Dispositifs numériques d'évaluation des compétences en langues vivantes étrangères : concevoir, tester des procédures de positionnement (semi)-automatisées

Elina Polchynski

Université Bordeaux Montaigne

ED Montaigne Humanités - EA CLIMAS

Dispositifs numériques d'évaluation des compétences en langues vivantes étrangères : concevoir, tester des procédures de positionnement (semi)-automatisées

Thèse de doctorat

Linguistique anglaise et didactique des langues

Elina POLCHYNSKI

Direction: M. le Professeur Jean-Rémi LAPAIRE

Soutenue à Pessac le 23.06.2016

Devant un jury composé de

M. Jean ALBRESPIT, Professeur à l'Université Bordeaux Montaigne
MME Laurence DURROUX, Professeure à l'Université Grenoble-Alpes
M. Jean-Rémi LAPAIRE, Professeur à l'Université Bordeaux Montaigne
MME Emmanuelle ROUSSEL, Professeure à l'Université de Caen

Remerciements

J´utilise cette occasion tout d'abord à remercier mon directeur de thèse **Jean-Rémi LAPAIRE** d'avoir accepté de diriger ce travail et de m'avoir accompagnée durant ces années. Il m'a notamment donné accès au projet qu'il pilotait et ouvert ses groupes d'étudiants pour tester POSILANG in situ. Grâce à sa motivation, son aide constant, ses nombreux conseils et critique constructive, j'ai eu la possibilité de mener à bien mon travail de recherche ainsi que de le présenter et défendre lors de la soutenance.

Je voudrais aussi remercier chaleureusement **Mesdames Durroux** et **Roussel**, ainsi que **Monsieur Albrespit** d'avoir accepté de siéger à mon jury, alors qu'ils sont très sollicités ailleurs. Je n'oublie pas non plus **Matt Armittage** qui m'a apporté une aide linguistique précieuse et généreuse pour l'anglais.

Enfin et surtout, mes pensées affectueuses et reconnaissantes vont à mes proches, en particulier, ma **mère Ludmilla** et ma **sœur Olga**. Même si elles sont malheureusement géographiquement éloignées, elles m'ont émotionnellement soutenue et encouragée tout au long de la durée de mes recherches. Grâce à leur patience, conseils et compréhension, il est devenu possible de terminer ce travail qui est énormément important dans ma vie et pour mon développement. Ma famille signifie tout pour moi et sans elle je n´aurais pas pu terminer mon projet. Je n'oublie pas non plus à remercier sincèrement mes amis et collègues qui m'ont soutenue et relue, notamment Marion**.** Enfin, je tiens particulièrement à remercier **Michel** pour son soutien constant, ses encouragements et aussi ses relectures.

Sommaire

INTRODUCTION : EVALUER DES COMPETENCES EN LANGUES A L'AIDE DE DISPOSITIFS STANDARDISES ET (SEMI)AUTOMATISES. POURQUOI? COMMENT?....12

1 PRATIQUES ET DISPOSITIFS: QUELQUES EXEMPLES....20

1.1 DEFINITION DES PRINCIPAUX TERMES DANS LE DOMAINE DE L´EVALUATION LANGAGIERE....20

1.1.1 Définition du terme « évaluation »....21

1.1.2 Définition du terme « test »....24

1.1.3 Définition du terme « mesure »....28

1.1.4 Lien entre les trois concepts : évaluation, test et mesure....31

1.2 LES PRATIQUES D'EVALUATION OBSERVEES DANS LES DIFFERENTS DOMAINES....33

1.2.1 Les pratiques d'évaluation dans l'enseignement supérieur....33

1.2.2 Les pratiques d'évaluation en didactique des langues vivantes étrangères....35

1.3 TYPOLOGIE DES TESTS DE LANGUES: VARIATIONS ET CONVERGENCES....36

1.3.1 Variations entre les tests des langues....36

1.3.2 Diversité de la typologie des tests en langues....37

1.3.2.1 Les tests de positionnement....37

1.3.2.2 Les tests d'acquisition et de progression....40

1.3.2.3 Les tests diagnostiques....42

1.3.2.4 Les tests de compétence....43

1.3.2.5 Les tests d'aptitude....45

1.3.3 D'autres paramètres de variation entre les tests....45

1.3.4 La grille d'évaluation « maison »....46

TABLEAU 1 – GRILLE « MAISON » (OU « PROJET REGION »)....47

1.3.5 La diversité de l'impact des tests langagiers....59

1.3.6 La grille utilisée dans les centres de langue en Italie....59

1.3.7 Le bilan de l'analyse comparative des tests de positionnement en italien....60

1.3.8 Convergences....62

1.4 LES QUALITES DES TESTS LANGAGIERS....63

1.4.1 L'utilité d'un test....63

1.4.2 Praticité....65

1.4.3 Fiabilité....66

1.4.4 Validité....70

1.4.4.1 Validité de contenu....74

1.4.4.2 Validité du test liée à l'usage réel de la langue....75

1.4.4.3 Validité liée aux construits d'un test....75

1.4.4.4 Validité apparente....76

1.4.4.5 L'aspect consécutif de la validité de construit....77

1.4.5 Impact *79*
1.4.6 Authenticité *81*
1.4.7 Interactivité *84*
1.5 QUESTIONNEMENTS ACTUELS 84
1.5.1 Les réflexions sur l'équité et l'éthique d'un test *84*
1.5.2 Evaluation traditionnelle ou alternative ? *90*
1.5.3 L'usage des technologies numériques dans l'évaluation des *92*
1.6 EVALUATION DE TESTS DE POSITIONNEMENT EXISTANTS 94
1.6.1 Oxford Quick Placement Test *95*
1.6.1.1 Description et évaluation des caractéristiques du test 95
1.6.1.2 Analyse de la typologie d'exercices et de compétences 97
1.6.2 Oxford Placement Test 2 ..... *100*
1.6.2.1 Description et évaluation des caractéristiques du test 100
1.6.2.2 Analyse de la typologie d'exercices et de compétences 103
1.6.3 Energy Placement Test *108*
1.6.3.1 Description et évaluation des caractéristiques du test 108
1.6.3.2 Analyse de la typologie des exercices et des compétences 111
1.6.4 Upstream Enterprise Placement Test *114*
1.6.4.1 Description et évaluation des caractéristiques du test 114
1.6.4.2 Analyse de la typologie des exercices et des compétences 115
1.6.5 CUTTING EDGE PLACEMENT TEST 120
1.6.5.1 Description et évaluation des caractéristiques du test 120
1.6.5.2 Analyse de la typologie des exercices et des compétences 124
1.6.6 Success Placement Test *125*
1.6.6.1 Description et évaluation des caractéristiques du test 125
1.6.6.2 Analyse de la typologie des exercices et des compétences 127
1.6.7 Le test du CAREL *133*
1.6.7.1. Description et évaluation des caractéristiques du test 133
1.6.7.2 Analyse de la typologie des exercices et des compétences 135
1.6.8 Bilan de l'évaluation des tests *136*
1.7 L'UTILISATION DES TESTS DE POSITIONNEMENT EN ANGLAIS DANS L ENSEIGNEMENT SUPERIEUR 137

2 ADOSSER UN TEST AU CADRE EUROPEEN COMMUN DE REFERENCE POUR LES LANGUES 144

2.1 ENJEUX 144
2.1.1 La perspective philosophique du Cadre Européen Commun de Référence *148*
2.1.1.1 Comparaison entre la perspective actionnelle et l'approche communicative 151
2.1.1.2 L'évaluation des compétences dans la perspective actionnelle 152
2.1.1.3 Les qualités inhérentes au Cadre européen commun de référence pour les langues 153
2.2 ORIGINE DES ECHELLES DE DESCRIPTEURS DU CECRL 155
2.2.1 Origine des niveaux communs de référence *155*

2.2.2 Origine des échelles des descripteurs *158*
2.3 PRESENTATION DES ECHELLES ILLUSTRATIVES DU CECRL 163
2.3.1 Le schéma descriptif *164*
2.3.2 Les niveaux de compétence *166*
2.3.2.1 Les types d'échelles de compétences 169
2.3.3 La critique du système basé sur les échelles *170*
2.4 ANALYSE DES DESCRIPTEURS 172
2.4.1 Procédé d'élaboration des descripteurs *173*
2.4.2 Critique des descripteurs contenus dans les CECRL *176*
2.4.2.1 Opacité des descripteurs 180
2.4.2.2 Les descripteurs des niveaux avancés 183
2.4.3 Dimensions qualitative et quantitative des descripteurs *184*
2.4.4 Formulation des descripteurs *186*
2.5 USAGE DU CECRL 188
2.5.1 Influence du CECRL sur la politique éducative *188*
2.5.2 Les enjeux sensibles de l'usage du CECRL *190*
2.5.2.1 La question de la validité 193
2.5.3 Le ratio de l'influence du CECRL sur l'enseignement et l'apprentissage des langues et sur l'évaluation des compétences *195*
2.5.3.1 Reconnaissance mutuelle de qualifications langagières 199
2.5.4 Devoirs futurs du CECRL *200*
2.5.5 Application du CECRL I: les Portfolios européens des Langues *202*
2.5.6 Application du CECRL II: la conception de la Grille d'analyse du contenu des items de compréhension écrite et orale *210*
2.5.6.1 Etape 1 : Analyse du CECRL 212
2.5.6.2 Etape 2 : Mise en application du cadre d'analyse 213
2.5.6.3 Etape 3: Mise en application élargie de la grille révisée 214
2.5.6.4 La conception de la grille numérisée *215*
2.5.6.4.1 Le module d'entrainement 220
2.5.6.5 Conclusion tirée après l'élaboration de la grille 227
2.6 DIALANG ET AU-DELA 229
2.6.1 Les origines et le développement de DIALANG *229*
2.6.1.1 Les problèmes rencontrés lors du processus de développement 230
2.6.1.2 Le pilotage 230
2.6.1.2.1 Résultat du pilotage en anglais 231
2.6.2 Type du test *233*
2.6.3 Conception de DIALANG *233*
2.6.4 Usage de DIALANG *242*
2.6.5 Usage de l'auto-évaluation en général *243*
2.6.6 Lien entre le CECRL et le système d'évaluation DIALANG *244*

2.6.6.1 L'usage du CECRL dans le cadre d'évaluation de DIALANG 245
2.6.6.2 Spécifications pour l'auto-évaluation 248
2.6.7 Format de DIALANG 248
2.6.8 Remarques conclusives 250

3 DE LA CONCEPTION A LA VALIDATION D'UN NOUVEAU TEST DE POSITIONNEMENT 260

3.1 OBJECTIFS 260
3.1.1 Objectif 1: choix du format 263
3.1.1.1 Tests informatisés 264
3.1.1.2 Tests adaptatifs 265
3.1.2 Objectif 2: évaluation des compétences dans le test 267
3.1.3 Objectif 3: conception du test de positionnement POSILANG avec des modules complémentaires « filières » adaptables 268
3.2 DEFINITION DE LA COMPETENCE LANGAGIERE 269
3.2.1 Définition de la connaissance et de la capacité 271
3.3 LES MODELES DE LA COMPETENCE LANGAGIERE EN LANGUE SECONDE 272
3.3.1 Des exemples de modèles de communication langagière 273
3.3.2 Le modèle de la compétence langagière communicative de Bachman 276
3.3.2.1 La compétence langagière dans le modèle de Bachman 277
3.4 MATERIEL 279
3.4.1 Documents officiels 279
3.4.2 Les sources non-officielles 281
3.5 METHODE 281
3.5.1 La modélisation par facettes de Bachman 284
3.5.1.1 Application du cadre méthodologique et de ses composantes à l'évaluation langagière 298
3.6 L'ELABORATION D'UN TEST 299
3.6.1 Le respect des principes valables pour un test 300
3.6.2 Les étapes de l'élaboration d'un test 301
3.6.3 Etape 1: L'élaboration des spécifications d'un test 304
3.6.3.1 Les spécifications du test POSILANG 305
3.6.4 Etape 2 : conception de la maquette 309
3.6.4.1 Les directives pour la conception des items à choix multiples 310
3.6.5 Etape 3 : validation des tâches 315
3.6.6 Validation rationnelle des tâches 319
3.6.6.1 Validation rationnelle des tâches de POSILANG 320
3.6.7 Passation des items en amont de leur opérationnalisation 327
3.6.7.1 Le pilotage des items 329
3.6.7.2 Principaux essais de passation des items 330
3.6.8 Analyse des résultats du pilotage 331
3.6.8.1 Analyse qualitative 331

3.6.8.2 Analyse quantitative des items 331
3.6.8.3 La valeur de facilité des items 332
3.6.8.4 L'index de discrimination des items 333
3.6.8.5 Analyse de distracteurs 335
3.7 ANALYSE DU TEST ENTIER 336
3.7.1 Le Tableau de fréquence 336
3.7.2 Les Mesures de tendance centrale : la Moyenne, le Mode et la Médiane 336
3.7.3 Fiabilité 336
3.8 ITEM RESPONSE THEORY 337
4 VALIDATION EMPIRIQUE DES TACHES DE POSILANG 358
4.1 ETABLISSEMENT DES NIVEAUX DE REFERENCE 359
4.1.1 Groupe L1 359
4.1.2 Groupe M2 (Sciences du Langage) 363
4.2 ANALYSE DU TEST POSILANG 372
4.2.1 Conception de l'échelle d'évaluation 372
4.2.2 Répartition des candidats selon les niveaux 373
4.2.3 Comparaison avec le niveau attribué avant la passation du test 373
4.2.4 Analyse des tâches 374
4.2.4.1 Valeur de facilité 375
4.2.4.2 Tâches non fiables 376
4.2.4.3 Index de discrimination 376
4.2.4.4 Coefficient de fiabilité 378
4.2.4.5 Conclusion 379
CONCLUSION 386
BIBLIOGRAPHIE 391
1 REFERENCES 391
1.1 Ouvrages cités 391
1.2 Articles universitaires cités 396
1.3 Ouvrages et articles consultés mais non cités 403
2 AUTRES DOCUMENTS 405
2.1 Tests de langue pour l'anglais 405
2.2 Manuels scolaires 406
2.3 Rapports et référentiels institutionnels 407
2.4 Articles de presse 409

Introduction : évaluer des compétences en langues à l'aide de dispositifs standardisés et (semi)automatisés. Pourquoi? Comment?

L'évaluation des compétences langagières au moyen de dispositifs standardisés et (semi)automatisés suscite actuellement un vif intérêt auprès de groupes professionnels et sociaux différents. On peut constater cet intérêt non seulement chez les linguistes ou les experts en évaluation, mais aussi chez les enseignants ordinaires, les étudiants et les parents d'élèves qui se sentent directement concernés par le changement des pratiques d'évaluation en salle de classe (Brown 2010 : 1x). Les « dispositifs automatisés » désignent ici les tests de langues vivantes numérisés et installés sur des plateformes pour une passation intégrale en ligne, doublée d'une correction entièrement prise en charge par la machine. Les dispositifs semi-automatisés peuvent, quant à eux, inclure un support papier en complément de l'évaluation automatique en ligne, ou encore contenir des phases expressives moins contrôlées, à l'écrit comme à l'oral, soumises à l'appréciation individuelle de correcteurs. Un test de langue, quel que soit son type et son objectif, est un dispositif de mesure ou, plus exactement, un instrument servant à mesurer la compétence langagière (Douglas 2010 : 2).

Mais pourquoi chercher à comprendre les principes de l'évaluation langagière ? Et quel intérêt y a-t-il à décrire les procédures mises en œuvre ? La première des raisons est le plus souvent un besoin professionnel d'appréhender les aspects théoriques et matériels des tests. Ce besoin concerne surtout les personnes directement impliquées dans l'achat ou la conception de tests langagiers pour leurs établissements. Celles-ci doivent connaître l'évolution des bonnes pratiques (McNamara 2000 : 4). Cependant, même les personnes qui ne sont pas directement concernées par l'évaluation s'intéressent souvent à la notion de compétence langagière et à l'évaluation qui peut en être faite. Enfin, les tests de langues jouent aujourd'hui un rôle clé dans la formation et la professionnalisation des candidats. Les tests fonctionnent souvent comme une véritable passerelle permettant de passer à une étape suivante dans un parcours éducatif ou professionnel (McNamara 2000 : 4).

L'évaluation des compétences langagières par des tests standardisés est aujourd'hui favorisée par la politique linguistique du Conseil de l'Europe (CECRL 2005 : 10). Cet organisme supervise les activités visant l'apprentissage des langues dans les systèmes éducatifs des états membres, depuis que la *Convention culturelle européenne* est entrée en vigueur en 1954[1]. Le Conseil de l'Europe a en effet inscrit le plurilinguisme dans sa politique de développement et de rapprochement culturels. Il reconnaît que l'apprentissage des langues étrangères et l'acquisition d'un niveau de compétence communicative dans plusieurs langues est un droit fondamental de tous les citoyens européens. Cette organisation supranationale se charge non seulement de garantir ce droit, mais considère le développement du plurilinguisme tout au long de la vie comme particulièrement important. La promotion du plurilinguisme correspond à l'un des principes définis dans le programme « Langues Vivantes » du Conseil de L'Europe. Conformément à ce principe, il convient de développer et d'améliorer les modalités et les instruments permettant l'évaluation des programmes d'apprentissage (CECRL 2001 :10).

Le Conseil de l'Europe a procédé en 2001 à la publication du *Cadre Européen Commun de Référence pour les Langues* (CECRL). Il existe désormais une base commune pour la conception de dispositifs d'évaluation et l'attribution de niveaux standardisés de compétence en langues vivantes. En effet, le CECRL décrit de manière détaillée quelles compétences les apprenants doivent acquérir pour être capables d'utiliser une langue efficacement (CECRL 2005 : 9). Le Cadre permet de rapporter les résultats des évaluations en langue à des critères standardisés, définissant différentes compétences. Par ailleurs, ce même Cadre rend possible la comparaison entre pays des niveaux de compétence assignés aux apprenants. Cette internationalisation des critères d'évaluation et des niveaux de référence est rendue nécessaire par la circulation croissante des étudiants et des professionnels dans l'espace Européen. Enfin, l'évaluation standardisée constitue un progrès manifeste dans l'objectivisation de la procédure de mesure. Auparavant, les compétences individuelles avaient tendance à être caractérisées de façon plus locale, aléatoire et subjective au moyen de qualificatifs assez impressionnistes comme *avancé, intermédiaire* ou *débutant* (Alderson et al. *Final Report of The Dutch CEF Construct Project* 2004 :

20). Ce n'est plus le cas désormais, puisque des niveaux et des descripteurs précis permettent de sortir de cet impressionnisme.

Cependant, l'aide apportée par le CECRL aux concepteurs de tests et aux examinateurs n'est pas la seule raison qui milite en faveur de l'évaluation des compétences langagières au moyen de dispositifs standardisés. L'évaluation standardisée est aujourd'hui, à juste titre, partie intégrante de l'enseignement et de l'apprentissage des langues dans les établissements scolaires et les organismes de formation à tous les niveaux. Elle fournit une base empirique pour prendre de nombreuses décisions éducatives, tant sur le plan pratique que théorique (Purpura 2004 : 50). L'usage de tests adossés au CECRL a de nombreux avantages par rapport à des formes d'évaluation moins standardisées. Sur le plan théorique, l'atout principal de ce genre de test est l'équité, désigné *fairness* en anglais, résultant de plusieurs facteurs. Ainsi, tous les items sont présentés dans le même format à tous les candidats, ce qui impose une même procédure de réponse et d'attribution des scores à l'ensemble des candidats. L'équité est renforcée par l'imposition des mêmes conditions de passation, notamment des mêmes contraintes temporelles (Douglas 2010 :6).

Par ailleurs, les résultats obtenus permettent de tirer des conclusions sur les compétences d'un candidat et sur son degré de maîtrise de la matière, dans le domaine évalué. Par cette fonction, les tests aident à vérifier l'évaluation préalable des compétences ou des acquis d'un candidat. Ils permettent ainsi de confirmer ou de modifier le jugement antérieur porté sur ce dernier (Douglas 2010 :1). Il est évident que les conclusions tirées ne peuvent être justes que si le test utilisé est de bonne qualité (Purpura 2004 : 50). La qualité des dispositifs se révèle notamment par leur standardisation, car celle-ci garantit une même manière d'évaluer les progrès des candidats lors de passations répétées et étalées dans le temps (Douglas 2010 : 1). L'écart entre les différents résultats obtenus rend possible la formulation de conclusions sur l'apprentissage au fil du temps et même sur l'efficacité de l'enseignement (Purpura 2004 : 49).

Sur un plan pratique, les scores obtenus aux tests permettent de prendre des décisions éclairées sur le placement des candidats dans une formation convenant à leur niveau de compétence en langue (Purpura 2004 : 50). Une

mauvaise décision ne peut avoir que des conséquences néfastes pour les candidats, de sorte que les développeurs de tests et les examinateurs ont la responsabilité éthique de veiller à ce que les décisions prises à partir des tests qu'ils proposent soient aussi dignes de confiance que possible (Douglas 2010 :10).

En dehors des avantages évoqués ci-dessus, les instruments d'évaluation standardisés en langues sont un atout pour l'institution scolaire. Les résultats des élèves peuvent être comparés en tenant compte des critères établis pour une classe particulière ou bien pour plusieurs classes dans une même école, ou encore entre différents établissements (Douglas 2010 :9). Cependant, il ne suffit pas de connaître les caractéristiques majeures de l'évaluation (semi)automatisée et standardisée des compétences pour concevoir des tests véritablement équitables qui permettent des évaluations fiables au cours du temps. Des principes définissant des bonnes pratiques ont été élaborés par deux organisations spécialisées dans l'évaluation en langues, l'EALTA et l'ILTA.

L'EALTA (*European Association for Language Testing and Assessment*) a pour objectif déclaré de promouvoir la compréhension des principes théoriques de l'élaboration des tests et de l'évaluation en langues. L'association entend améliorer et partager les bonnes pratiques concernant les tests et l'évaluation, à l'échelle européenne.[3] Cette organisation à but non lucratif a été créée pour fédérer les responsables de l'évaluation en Europe, ainsi que les groupes chargés d'élaborer et d'utiliser des tests. [4] L'objectif de cette association est de garantir que la qualité lors de l'évaluation des compétences est assurée par tous ses membres (Byrnes 2007 : 644). La nécessité de fonder cette organisation répond à la politique linguistique menée par le Conseil de l'Europe. Elle est étroitement liée à la conception et à la mise en œuvre du CECRL et du portfolio européen des langues [5]. Cette association énonce les principes suivants dans ses lignes directrices pour une bonne pratique dans l'élaboration et l'utilisation de dispositifs d'évaluation en langues: le respect des candidats, la responsabilité, l'équité, la fiabilité, la validité et la collaboration entre les partis concernés [6]. L'EALTA exige de ses membres qu'ils respectent l'ensemble de ces principes, qu'ils participent à la conception ou à la mise en œuvre de tests, et quelle que

soit leur fonction**.** Bien que tous ces principes soient d'une importance égale, la fiabilité et la validité sont particulièrement complexes et exigent une réflexion particulière (Douglas 2010 : 26-29). Ce sont donc ces principes qui seront explorés en priorité dans ce travail.

La deuxième association chargée de l'évaluation en langue est l'ILTA (*International Language Testing Association*). La mission de cette association, également destinée aux professionnels de l'évaluation langagière, est d'améliorer les pratiques partout dans le monde. C'est un objectif qui en englobe de nombreux autres.[7] Cette organisation a également élaboré un code d'éthique qui a été adopté lors de sa réunion annuelle en 2000. Ce code, relevant de la philosophie morale, sert à garantir un comportement éthique satisfaisant chez tous les professionnels de l'évaluation langagière. Le code contient neuf principes fondamentaux qui sont fondés sur des valeurs comme la bienveillance envers les candidats, la justice, le respect de l'autonomie et de la société civile.[8] La dimension éthique est explicitement soulignée dans un principe particulier. Celui-ci demande aux professionnels de l'évaluation langagière de respecter tous les principes éthiques pertinents, postulés dans les lignes directrices nationales et internationales. Chacun de ces principes est complété par des annotations qui clarifient leur nature respective. Par exemple, les obligations et les modalités de comportement attendues sont précisées sous cette forme, tout comme les sanctions résultant du non-respect de celles-ci. En outre, les difficultés qui peuvent émerger lors de l'application des principes sont évoquées par ce code. [9]

Comme on le voit, la conception de tests de positionnement en langues conformes aux bonnes pratiques internationales est un acte professionnel d'une grande complexité, comportant à la fois un volet technique et éthique. C'est dans ce contexte d'exigence et de professionnalisation croissante que les universités de Bordeaux ont émis le souhait de disposer d'outils d'évaluation à usage local mais de niveau international, permettant de positionner rapidement des milliers de primo-entrants avant de les répartir dans des groupes de niveaux. La réalisation d'un test de positionnement a donc été inscrite dans les priorités du projet interuniversitaire "Didactique des langues: ressources numériques et hybridations " (2011-2014), cofinancé par la région Aquitaine. La présente thèse est née de ce projet, au titre du Volet 1 « Conception d'un test de positionnement

automatisé en langues pour les établissements du PRES de Bordeaux ». Dans sa version originale, le projet devait préparer l'avènement de la *Maison des Langues et des Cultures* [10] Nous verrons que des facteurs politiques, liée à la fusion mouvementée des établissements du PRES, ont entravé le financement et la collaboration interuniversitaire : la Maison des Langues n'a pas été réalisée, l'espace numérique qu'elle devait abriter et qui devait accueillir le test n'a pas été monté, et enfin l'Université Bordeaux Montaigne s'est retrouvée seule pour reconfigurer le projet et mener à terme ce qui pouvait être sauvé. Cet état de fait a eu des conséquences non négligeables sur le déroulement de nos recherches, sans pour autant avoir raison de notre détermination à mener à bien la conception et l'évaluation d'une première version du test de positionnement POSILANG.

Avant de concevoir un test de positionnement, il est important d'en définir la forme et les fonctions, tout en fournissant un aperçu critique des dispositifs déjà réalisés. Notre premier chapitre s'ouvre donc sur une caractérisation des processus d'évaluation de compétences en langues vivantes étrangères chez l'adulte. Les dispositifs automatisés ou semi-automatisés, dans des langues à larges effectifs comme l'anglais ou l'espagnol[11] sont rapportés à la problématique générale de la définition et de la mesure des compétences. Plusieurs tests de positionnement disponibles sur le marché, gratuits ou payants, jugés représentatifs de l'offre actuelle, sont examinés. Il est important de souligner que certains tests demandent aux étudiants d'évaluer leurs compétences linguistiques eux-mêmes, c'est-à-dire, de s'auto-évaluer. En conséquence, il est nécessaire de les examiner et les comparer afin de déterminer ceux qui permettent l'auto-évaluation la plus fiable.[12]

Le deuxième chapitre se concentre sur les enjeux décrits dans le *Cadre Européen Commun de Référence pour les Langues*. Il s'intéresse tout particulièrement aux échelles d'activité et de compétence en communication langagière. Ces dernières comprennent les compétences linguistiques, sociolinguistiques et pragmatiques, mais seules les habiletés linguistiques sont pertinentes pour cette thèse. Ce sont donc elles qui sont examinées ici. Sont également analysés les descripteurs des différents domaines d'activité de

communication langagière. Pour clore ce second chapitre, le test DIALANG est étudié en tant qu'exemple de dispositif étroitement adossé au CECRL.

Le troisième chapitre présente les étapes de conception d'un nouveau test (POSILANG) réalisé dans le cadre du projet aquitain mentionné plus haut, avec l'appui de chercheurs grenoblois rattachés au projet IDEFI-ANR Innovalangues (2011-2015). Les étapes successives du travail effectué (conception de la maquette, pilotage des versions pilotes) sont décrites et analysées. Le quatrième et dernier chapitre rapporte les évaluations de passation des versions pilotes et commente les premiers résultats obtenus.

Nous voudrions que cette thèse serve non seulement la recherche dans le domaine de l'évaluation automatisée des compétences en langues mais aussi les besoins concrets des équipes de formation. Notre idée est de fournir les éléments de théorie et de méthode permettant à chaque équipe de construire ses propres dispositifs. Nous verrons en effet que les meilleurs tests sont ceux qui, tout en respectant des principes généraux, s'inscrivent dans un contexte local et répondent à des besoins spécifiques d'évaluation.

Notes:

[1] La convention culturelle européenne peut être trouvée sur le site suivant : http://www.coe.int/t/dg4/linguistic/Division_FR.asp#TopOfPage.

[2] http://www.coe.int/t/dg4/linguistic/manuel1_fr.asp.

[3] http://www.ealta.eu.org/index.htm

[4] Il y a trois groupes chargés de l'élaboration, de la mise en œuvre et de l'administration des tests qui sont visées par les directives de l'EALTA. Il s'agit, premièrement, des personnes responsables de la formation des enseignants en évaluation langagière. Sont également concernés les enseignants eux-mêmes, chargés de l'évaluation des compétences en classe. Le troisième groupe visé par les lignes directrices est constitué des responsables du développement des tests dans les institutions nationales ou internationales (http://www.ealta.eu.org/guidelines.htm).

[5] Le Portfolio européen des langues (PEL) a été conçu par la Division des politiques linguistiques au sein du Conseil de l'Europe pour atteindre deux objectifs majeurs. Le premier but est de permettre aux utilisateurs de noter les résultats de leur apprentissage linguistique ainsi que leur expérience d'apprentissage et d'utilisation des langues. Le deuxième objectif est de développer l'autonomie de l'apprenant, son plurilinguisme et sa compétence interculturelle. Cet instrument est étroitement lié au CECRL par le dispositif d'auto-évaluation qui est la composante centrale du PEL (http://www.coe.int/t/dg4/education/elp/ELP-REG/CEFR_FR.asp#TopOfPage).

[6] http://www.ealta.eu.org/guidelines.htm

[7] www.iltaonline.com

[8] http://www.iltaonline.com/images/pdfs/ILTA_Code.pdf

[9] http://www.iltaonline.com/images/pdfs/ILTA_Code.pdf

[10] J-R. Lapaire : bilan étape : 2012.

[11] J-R. Lapaire : bilan étape : 2012.

[12] J-R. Lapaire : bilan étape : 2012.

1 Pratiques et dispositifs: quelques exemples

La conception d'un test de langue ne s'effectue jamais ex nihilo et tout créateur de dispositif est forcément amené à consulter l'existant, qu'il veuille s'en inspirer ou s'en démarquer. Les formes auxquelles le concepteur est confronté sont l'expression visible et finale d'un processus invisible de réflexion, idéalement adossé à des cadres théoriques déjà construits, analysés et confrontés par des chercheurs appartenant à des disciplines aussi diverses que la linguistique, la didactique des langues, la psychométrie, la sociologie, l'éthique ou d'autres encore.

A l'évidence, tous les tests ne se « valent » pas, à compter qu'il soit possible de s'entendre sur la notion controversée de validité, comme nous allons le voir. Certains dispositifs ne prétendent être rien d'autre que d'astucieuses fabrications de circonstance, répondant dans l'urgence à une injonction administrative ou pédagogique locale. D'autres, au contraire, sont portés par de prestigieuses institutions publiques ou privées. Ils sont destinés à une large diffusion, se présentent comme le fruit d'un patient travail d'élaboration, et s'inscrivent dans un ensemble de bonnes pratiques professionnelles (qu'il conviendra d'identifier). Tous se proposent d'évaluer, à une fin ou une autre, des compétences. Mais en quoi consiste cette évaluation ? De quels biais souffre-t-elle ? Quel est son impact sur les pratiques pédagogiques mais aussi l'organisation sociale ? Ce ne sont là que quelques questions, tant les ramifications d'un test sont nombreuses. Nous proposons donc d'aborder le processus de développement, ses fondements théoriques et ses implications pratiques, au travers d'un échantillon représentatif de tests déjà en circulation. Mais il nous faut au préalable saisir les caractéristiques, les fonctions et les enjeux sociétaux de cette forme d'évaluation.

1.1 Définition des principaux termes dans le domaine de l'évaluation langagière

Avant d'entamer la description de tests particuliers, il convient d'introduire les concepts fondamentaux qui structurent le domaine de la mesure. La compréhension en est essentielle pour le développement et l'usage des tests

langagiers. Ces concepts clés sont couverts par les termes « mesure », « test » et « évaluation, » qui sont parfois à tort considérés et utilisés comme synonymes. Il est important de définir chacun de ces termes en faisant ressortir les traits qui les différencient (Bachman 1990 : 18). Pour cette raison, les trois termes seront présentés à la fois de manière autonome et contrastive dans les pages qui suivent.

1.1.1 Définition du terme « évaluation »

Le terme « évaluation » est souvent utilisé, à tort, comme synonyme de « test. » Cet usage, malheureusement répandu, est tout à fait contestable, dans la mesure où la passation d'un test n'est pas la seule forme d'évaluation (Conseil de l'Europe 2005 : 135). L'évaluation recouvre en réalité de très nombreuses formes de contrôle des connaissances ou des compétences, ce qui en fait un terme beaucoup plus large que « test », généralement lié à un événement ou à une série d'événements ponctuels. Le contrôle continu est également une forme d'évaluation, sans pour autant être réductible à un test (Conseil de l'Europe 2005 : 135). Ce dernier exemple vient rappeler que l'évaluation est un processus plus englobant, qui s'inscrit généralement dans une temporalité élargie. Le caractère « continu » s'explique par le fait que le processus d'évaluation est engagé en permanence (Brown 2010 :3). Cependant, l'évaluation est une démarche qui dépasse celle de contrôle, notamment lorsqu'elle prend un tour informel et même inconscient. Le CECRL évoque les observations informelles par les enseignants comme un exemple des formes d'évaluation allant au-delà des procédures de contrôle, qui sont forcément formelles dans tous les cas (Conseil de l'Europe : 135). Or, certains linguistes vont même plus loin que le Conseil de l'Europe, en attribuant au processus d'évaluation un caractère souvent inconscient, arguant que l'évaluation n'est pas toujours organisée de façon calculée, et peut également prendre un caractère fortuit (Brown 2010 : 3). Néanmoins, cette définition de l'évaluation n'est pas acceptée unanimement. Selon certains chercheurs, l'évaluation demande la collecte systématique de données dans le but de prendre une décision (Bachman 1990). Or, la collecte systématique d'informations s'oppose au caractère fortuit et aléatoire de cette procédure, soulignée par d'autres linguistes (Brown 2010: 3). Le besoin de

collecter des données de façon consciente et systématique est pourtant justifiable par la nécessité de prendre une décision ultérieurement. Pour s'assurer que la décision soit bonne, il faut veiller à ce que les informations soient pertinentes et fiables (Bachman 1990 : 22).

Le terme « évaluation » a donc un double sens. Il renvoie aussi bien à une estimation subjective et approximative qu'à une mesure objective et précise (Huver & Springer 2011 : 5). Cette ambivalence est logée dans l'étymologie, car « évaluer » provient du mot « avaluer » en ancien français, qui possède déjà les deux sens : le premier est « déterminer approximativement par une appréciation la valeur de quelque chose » et le deuxième « fixer le prix, la valeur de quelque chose » (Huver & Springer 2011 : 5). On retrouve bien cette double orientation tant dans les discours quotidiens que dans les discours et les pratiques d'enseignement et de recherche. Les deux sens contenus dans le concept d'évaluation convergent néanmoins autour de la notion centrale de valeur (Huver & Springer 2011 : 5).

Avant de décrire les manifestations et les fonctions de l'évaluation, il est important de se rappeler qu'il n'en existe pas de théorie générale. Cela peut paraître surprenant mais c'est ainsi et cela a pour conséquence l'hétérogénéité extrême des référents théoriques des actes individuels d'évaluation. Toutefois, la variation ne doit pas cacher des éléments de convergence, voire d'unité, entre des procédures censées partager une idée méliorative. Mais là encore, l'absence d'une théorie générale de l'évaluation, jette le trouble sur la légitimité même de cet objectif mélioratif (Chardenet 1999 :11).

Il est indispensable de prendre en compte les deux types de relations qui sous-tendent toute évaluation. Celles-ci permettent de mieux comprendre le mécanisme proprement dit d'évaluation et de mieux interpréter les actes qui en relèvent. La première de ces relations est celle qui unit le sens et le discours, tandis que la deuxième unit l'évaluateur et le phénomène évaluable, désigné *évaluataire* (Chardenet 1999 :11). Le discours d'évaluation joue un rôle primordial parce qu'il remplit trois fonctions. Premièrement, le discours établit les conditions de la mesure, comme la fixation des critères, et le repérage des indicateurs. Deuxièmement, il revient au discours de choisir la valeur de la

notation, c'est-à-dire, de choisir si celle-ci sera arithmétique ou non. La troisième fonction du discours est de porter un jugement sur des phénomènes évaluables, sous forme d'argumentation ou bien sous celle d'annotation (Chardenet 1999 :11).

Le système éducatif est en demande d'évaluation exacte, transparente et à forte valeur symbolique. Ces trois exigences, celles de l'exactitude, de la transparence et de la valeur symbolique, sont liées car la distribution des notes doit être exacte et transparente afin de posséder une valeur symbolique importante (Huver & Springer 2011 : 5).

L'importance de l'évaluation ne se limite cependant pas à la sphère de l'éducation et de l'enseignement car les pratiques évaluatives occupent une place de plus en plus importante dans d'autres domaines de la vie, notamment dans les secteurs économique, politique et médiatique. Une véritable culture de l'évaluation s'est répandue à tous les niveaux de l'organisation sociale, à tel point qu'on parle désormais de « l'expansion de la logique évaluative tout au long de la vie » (Martuccelli 2010 : 123). Cette expression souligne non seulement l'importance accordée à l'évaluation dans notre société, mais également le fait que tout individu est désormais concerné par les pratiques évaluatives dans des contextes sociaux très différents (Martuccelli 2010 : 120). Dans les sociétés contemporaines, les individus sont désormais sujets à d'incessantes évaluations. Celles-ci ne se déroulent pas nécessairement de façon formelle, dans un cadre institutionnel. Qu'elles soient ouvertes ou implicites, ces évaluations n'en demeurent pas moins de véritables « épreuves factuelles » qui servent à la sélection des personnes (Martuccelli 2010 : 120). Les évaluations sont des défis sociaux auxquels les citoyens sont confrontés et qui ont des conséquences importantes sur leur avenir (Martuccelli 2010 : 122). Visant le même objectif que les épreuves formalisées, à savoir, la sélection des personnes, les épreuves qui ne sont ni formalisées ni institutionnalisées, sont permanentes. Elles s'inscrivent dans un processus de sélection qui se déroule en continu (Martuccelli 2010 : 122).

Au sein des pratiques évaluatives, force est de constater l'existence de deux courants diamétralement opposés. L'un vise à obtenir des mesures exactes

et transparentes, censées rendre compte de la réalité sur un mode aussi objectif que possible. L'autre, valorise l'auto-évaluation, ainsi que le jugement intuitif et donc subjectif (Huver & Springer 2011 : 5). Cependant, ces deux tendances n'ont pas un poids égal dans nos sociétés car c'est le courant objectiviste qui domine clairement à l'heure actuelle. Ce dernier cherche à éliminer au maximum la subjectivité du jugement en recourant à des techniques spéciales. Les moyens utilisés doivent permettre une mesure aussi objective que possible de la performance. Pour cela, il est nécessaire d'expliciter les critères dans le champ des connaissances et des compétences évaluées (Chardenet 1999 :8). Cette «entreprise méthodique d'évacuation de la subjectivité du jugement par des techniques appropriées » (Chardenet 1999 : 7) n'en demeure pas moins paradoxale, au moment même où se répandent les procédures d'auto-évaluation prises en charge par le sujet et où est reconnue l'omniprésence de la subjectivité (Huver & Springer 2011 : 5). Un second paradoxe tient au fait que la pratique évaluative reste peu comprise du public, malgré son rôle majeur dans tous les domaines de la société de nos jours (McNamara 2000 : 3). Souvent, on constate même un manque de volonté de comprendre les procédures d'évaluation par le grand public qui préfère s'en remettre au jugement final des experts, sans s'intéresser au détail des procédures (McNamarra 2000 : 3).

1.1.2 Définition du terme « test »

Le test est un outil parmi d'autres d'évaluation, qui possède ses caractéristiques propres. Il s'agit d'une forme de contrôle qui a lieu à des moments identifiables dans le cursus et qui repose sur des procédures instituées. Lors des moments assignés de passation, les élèves cherchent à atteindre la meilleure performance possible, en sachant que leurs réponses seront mesurées et évaluées (Brown 2010 :3).

Défini en termes scientifiques, un test est une méthode de mesure de l'habileté, de la connaissance ou de la performance d'un individu dans un domaine particulier (Brown 2010 : 3). Pour mieux comprendre cette définition, il convient de regarder de près les termes utilisés. Premièrement, l'idée de méthode implique que le dispositif d'évaluation constitué par le test est explicite

et structuré. Deuxièmement, un test doit permettre de mesurer la performance d'un candidat selon des règles et des procédures explicites de quantification (Bachman 1990: 20). L'usage de procédures explicites pour mesurer la performance ne constitue pourtant pas un trait propre au test et concerne également d'autres outils de mesure (Bachman 1990 : 20). Le troisième point évoqué dans la définition citée est qu'un test de langue mesure toujours une performance au travers d'un échantillon de l'usage individuel de la langue (Brown 2010 : 3). Le but est de faire produire des instances particulières d'usage de la langue qui ne soient pas le fruit du hasard, mais spécifiques, révélant par là même un comportement particulier du candidat (Bachman 1990 : 21).

Recueillir des échantillons d'usage de la langue permet de se focaliser plus précisément sur telle ou telle compétence et de tirer des conclusions pertinentes. Seuls des échantillons spécifiques, reflétant l'usage précis de la langue par les sujets, permettent d'obtenir des scores qui soient fiables et utiles (Bachman 1990 : 21). Cette particularité des tests explique non seulement pourquoi on a forcément besoin de ce genre d'échantillon, mais fournit aussi la justification pour l'usage des tests en général. La nécessité d'obtenir les échantillons spécifiques de la performance des candidats conditionne la conception et la mise en œuvre d'un dispositif d'évaluation (Bachman 1990 : 21). Le trait distinctif des tests souligné par Bachman (1990) a été reconnu déjà à la fin des années 1960, comme en témoigne la définition généralement acceptée de Carroll (1968): « a psychological or educational test is a procedure designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual » (Carroll 1968: 46).[2]

Quelles conclusions peut-on tirer de cette définition générale d'un test sur les tests de langue ? Dans la plupart de ces derniers, les résultats de la performance d'un candidat servent à établir sa compétence dans une activité de communication langagière, à savoir, la compréhension de l'oral et de l'écrit, ainsi que la production orale et écrite. En revanche, certains tests utilisent les résultats de la performance d'un candidat pour tirer des conclusions sur sa connaissance dans un certain domaine linguistique, que celui-ci soit lexical, grammatical ou rhétorique (Brown 2010: 4). Cette dernière caractéristique implique qu'un test mesure toujours la performance dans un domaine particulier, dont la nature et

l'étendue varient d'un test à l'autre, et qui doit être bien défini dans tous les cas (Brown 2010: 4).[3] Ainsi, un test de compétence générale vise à évaluer le niveau de compétence dans tous les domaines langagiers. D'autres tests, plus ciblés, comme ceux de grammaire ou de vocabulaire, ont des critères d'évaluation plus spécifiques qui varient selon leur but respectif (Brown 2010 : 4). Il en existe de nombreux sur le marché, par exemple, *tolearnenglish*, e-anglais ou bien *ifg langues,* dont l'objectif consiste à évaluer une seule compétence, en grammaire et en compréhension écrite respectivement. [4]

La définition du test proposée par Purpura (2004) dépasse celle de Brown car elle va au-delà du processus de mesure de performance. Cette définition inclut aussi les étapes qui doivent succéder à celle de la mesure en faisant référence à deux autres termes clés, à savoir, les *inférences fondées sur les scores* et les *décisions justifiées* sur les individus. Le premier de ces termes est développé de la manière suivante dans son ouvrage :

> The responses to the test items can then be used as a basis for assigning scores and for making inferences about the student's underlying […] ability. […] we must infer the underlying ability from responses to questions or from samples of actual performance. Since responses to test items are ultimately converted into scores, we say we can make score-based inferences about an examinee's grammatical ability. (Purpura 2004: 147).

Les inférences fondées sur les scores servent à prendre des décisions qui, à ce titre, sont toujours justifiées, malgré leur diversité possible :

> Score-based inferences from grammar tests can be used to make a variety of decisions. […]. These inferences can then serve to provide feedback for learning and instruction, assign grades, promote students to the next level or even award a certificate. They can also be used to help teachers or administrators make decisions about instruction or the curriculum (Purpura 2004: 147).

La prise d'une décision constitue le trait commun entre le test et d'autres procédures d'évaluation. En effet, la décision est considérée comme faisant partie intégrante du processus d'évaluation aussi par d'autres chercheurs, qui considèrent la décision à la fois comme le point de départ et le point d'arrivée de l'évaluation: « Nous entendons par évaluation la collecte systématique de données dans le but de prendre une décision. C'est en effet la notion de décision qui est, à notre avis, le point de départ de toute démarche évaluative » (Doucet 2001 : 2). La plus grande prudence s'impose néanmoins lors de prises de décision importantes pour les candidats, lorsque le jugement émis se fonde sur une interprétation d'indices fragmentaires de performance recueillis durant la

passation du test. Il faut savoir que la compétence en langue seule ne suffit pas à la prise de décision, car les résultats obtenus aux tests ne sont jamais infaillibles (Douglas 2010: 20). Pour cette raison, il est nécessaire de faire preuve de distance et de prudence vis-à-vis des résultats révélés par le test, tout en essayant de minimiser le potentiel d'erreur et de tirer les conclusions sur la compétence des candidats le plus précisément possible (Douglas 2010 : 29). Malgré ces précautions, il est indispensable de tenir compte de toute information disponible sur un candidat, y compris, de ses résultats antérieurs, des recommandations par d'autres enseignants, du niveau de motivation et de l'expérience professionnelle du candidat en question (Douglas 2010 : 20).

Selon une définition proposée par Davies (2007), trois étapes sont indispensables lors de la conception d'un test et de l'évaluation des compétences des candidats. Il s'agit de la description du standard ou du niveau, de la formulation explicite de la mesure et du rapport des résultats de l'évaluation. La détermination du standard ou du niveau est la première étape car c'est la condition préalable pour pouvoir mesurer la performance des candidats (Davies 2007 : 437-438). En tant que deuxième étape, la mesure qui indiquera la validation ou non du niveau postulé doit être formulée explicitement. Le bilan des résultats de l'évaluation, opérant une synthèse des scores, des notes ou des profils, intervient en tant que troisième étape (Davies 2007 : 437-438). Ce rapport varie d'un test à l'autre, mais est nécessaire dans tous les cas, notamment afin d'informer des candidats de leurs résultats.

Tous les tests langagiers visent à déterminer la connaissance ou la compétence d'un candidat à partir de sa performance. Mais cela nécessite aussi la prise en compte de l'identité des candidats et de leur parcours antérieur. Cette intégration d'éléments personnels et contextuels doit intervenir en amont de la procédure, dans le but d'adapter le test aux capacités et au milieu des participants. La connaissance des pré-acquis des candidats est indispensable afin de permettre aux administrateurs et aux apprenants d'interpréter les scores obtenus au test de façon juste (Brown 2010 : 4).

1.1.3 Définition du terme « mesure »

Le deuxième terme qu'il faut bien distinguer de celui d'évaluation est celui de mesure. L'évaluation ne nécessite pas forcément le recours à une mesure stricto sensu pour avoir lieu. Comme on l'a vu dans la partie précédente, les composantes générales de l'évaluation sont la collecte et la prise en compte de données ainsi que la prise de décision (Doucet 2001 : 2). Ces qualités délimitent l'évaluation de la procédure de mesure, car cette dernière a pour fonction de fournir des informations pertinentes sur un phénomène observé, sans inclure la prise de décision à la base des données recueillies (Bachman 1990 : 23). Pour pouvoir remplir cette fonction, la procédure de mesure recourt à l'observation du phénomène qui suscite l'intérêt. Cependant, comme il s'agit d'une observation naturaliste, il est probable que les échantillons du comportement observé n'incluent pas les compétences ou les attributs spécifiques constituant la cible d'intérêt des chercheurs (Bachman 1990 : 23).

La mesure peut être définie comme le processus qui permet de quantifier la performance observée des candidats. La nécessité de quantifier la performance en recourant à des échelles résulte du fait qu'il est impossible de comparer directement des compétences et d'autres attributs internes des individus. Pour ce faire, il faut établir des échelles de mesure qui contiennent des nombres codant les différentes valeurs de la performance (Bachman 1990 : 26). La mesure peut aussi être effectuée sur des échelles dites qualitatives. Ces dernières sont de deux types : nominale et ordinale (Doucet 2001 : 3).

L'échelle nominale comporte les nombres qui servent à coder les différentes catégories d'un même attribut. L'échelle nominale est conçue en quantifiant l'attribut, par exemple, « la langue maternelle », ce qui implique qu'on assigne des nombres différents aux diverses catégories de cet attribut, comme « anglais », « français », « espagnol » et « italien ». Le lien entre le numéro assigné et la catégorie est arbitraire, puisque le nombre assigné à chaque catégorie est indifférent, la seule exigence étant qu'il soit unique (Bachman 1990 : 27). Il faut souligner qu'il n'y a pas de relation hiérarchique entre les catégories distinctes de cette échelle (Doucet 2001 : 3). Dans la mesure où elles

ont pour fonction de quantifier des catégories, ces échelles sont parfois désignées comme *catégorielles* (Bachman 1990 : 27).

La deuxième échelle d'évaluation qui est communément appliquée en langues est ordinale. Cette échelle comporte l'attribution de nombres aux catégories d'un attribut qui sont non seulement distinctes mais ordonnées, c'est-à-dire, situées en relation hiérarchique les unes par rapport aux autres (Bachman 1990 : 28). Il faut souligner que les catégories de l'échelle ordinale sont séparées par un intervalle qui est variable (Doucet 2001 : 3). Les définitions des différents niveaux d'une même compétence en constituent un exemple. Les échelles et les sous-échelles des compétences contenues dans le CECRL constituent les échelles ordinales car les définitions des niveaux communs de référence exprimées par les descripteurs respectifs permettent l'évaluation de la performance individuelle à l'aide de catégories distinctes et hiérarchiquement ordonnées.

Les deux échelles quantitatives sont l'échelle d'intervalles et l'échelle proportionnelle (Doucet 2001 : 3). L'échelle d'intervalles est utilisée pour établir des catégories distinctes et ordonnées. L'intervalle qui sépare une catégorie d'une autre est ici constant (Doucet 2001 : 3). Pour cette raison, cette échelle fournit des informations complémentaires par rapport à l'échelle ordinale. Pour donner un exemple, l'échelle d'intervalles ordonne les candidats non seulement sur une échelle en fonction de leur niveau de compétence respectif, mais montre également que la distance entre les niveaux de compétence de chaque candidat est la même (Bachman 1990 : 27-28). Pour pouvoir se servir de cette échelle, il faut évidemment déterminer au préalable que la distance entre toutes les catégories est vraiment identique. Or, en matière de niveaux de compétences, il est impossible d'évaluer que la distance entre ceux-ci soit identique, de sorte que l'échelle d'intervalles ne peut pas être utilisée dans le domaine de l'évaluation des compétences.

Ce même argument constitue la raison pour laquelle il est impossible d'appliquer le deuxième type d'échelle quantitative, l'échelle proportionnelle, pour l'évaluation langagière. Ce type d'échelle se distingue par l'existence d'un zéro absolu permettant de comparer les résultats de la performance en termes de

proportion (Doucet 2001 : 4). Cette manière de comparer ne peut quasiment pas être appliquée dans le domaine de l'évaluation langagière car il est impossible de dire qu'un score de 60 indique un niveau de compétence deux fois plus élevé que celui symbolisé par un résultat égalant 30 points. De même, un résultat chiffré de zéro n'implique pas l'absence absolue de compétence en langue dont la définition n'est, d'ailleurs, pas possible actuellement (Doucet 2001 : 4). En raison des difficultés qu'il y a à tirer des conclusions sur les compétences langagières des candidats à partir de mesures effectuées sur des échelles quantitatives, on recourt généralement à des échelles qualitatives dans le but d'évaluer ce genre de compétences (Doucet 2001 : 4).

L'usage présenté des quatre échelles n'est pas interchangeable, car il varie en fonction des caractéristiques de l'attribut mesuré ainsi que des procédures de mesure utilisées (Bachman 1990 : 27). Chacune de ces échelles de mesure est dotée de qualités différentes, qu'il est essentiel de comprendre autant pour le développement que pour l'usage d'un test. L'importance de cette compréhension est due au fait que les caractéristiques des échelles de mesure déterminent la manière d'interpréter et d'utiliser les scores (Bachman 1990 : 27).

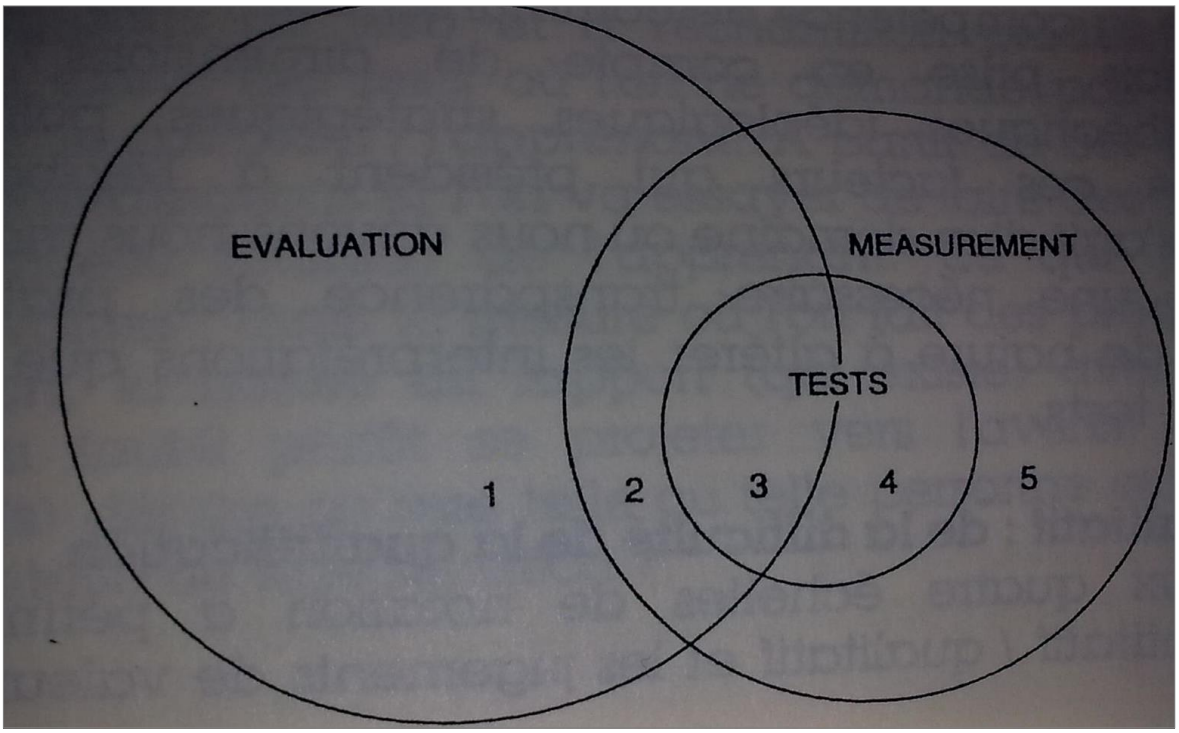
Rien n'oblige à recourir à une mesure quantitative, à attribuer une valeur chiffrée à la performance. On peut très bien mesurer la performance observée moyennant des rapports écrits et oraux non-quantifiables (Brown 2010 :5). Dans ce cas, on recourt à une mesure qualitative (Brown 2010 :5). Toutefois, la quantification est volontiers considérée comme porteuse de plusieurs avantages par rapport à une mesure qualitative. La description de la performance des candidats est plus concise, précise et homogène, ce qui permet de comparer les individus plus rapidement et facilement, surtout dans un contexte de massification aussi prononcé que celui de l'enseignement secondaire ou supérieur. En outre, la quantification permet souvent de décrire la performance avec une plus grande objectivité. Cependant, cette dernière caractéristique, considérée comme un point fort de la description quantitative, n'est pas dépourvue d'inconvénients: une objectivité chiffrée peut cacher une subjectivité ou une approximation dans les items servant à l'établir. Elle peut aussi gommer des nuances dans la performance. Au final, le résultat quantitatif est prononcé avec une certitude excessive par rapport à son objectivité réelle (Brown 2010 :5).

Cela ne signifie pas qu'il faille renoncer à quantifier mais cela invite à prendre un certain nombre de précautions, en amont (fabrication des outils d'évaluation quantitative) et en aval (interprétation puis utilisation des résultats pour orienter). Une description qualitative, en revanche, offre la possibilité de prendre en considération les nuances de la performance et d'individualiser le retour fait au candidat (Brown 2010 :5). Elle est cependant plus complexe et chronophage à mettre en place.

L'emploi d'une échelle quantitative conduit à une procédure de notation le plus souvent sous forme de score. Selon Doucet (2001 :4), il n'y a pas de nécessité mathématique à ce que la compréhension de la moitié des informations corresponde à la note 10/20 dans le système éducatif français. Il affirme même que la notation a un caractère relativement arbitraire. On peut en conclure que malgré l'opinion répandue qui est en faveur d'une mesure quantitative, celle-ci n'est pas forcément la meilleure dans tous les cas. Comme pour le choix de l'échelle, la décision pour l'un des deux types de mesure doit être prise en fonction des caractéristiques de l'attribut mesuré ainsi que des procédures de mesure utilisées (Bachman 1990 : 23).

1.1.4 Lien entre les trois concepts : évaluation, test et mesure

Chacun des trois domaines décrits peut être représenté schématiquement à l'aide de cercles qui s'imbriquent l'un dans l'autre. Ce schéma, rassemblant l'évaluation, le test et la mesure, a été proposé par Bachman (1990 : 23).



Les trois concepts sont représentés par trois cercles et leurs relations par des zones d'intersection. Le graphique montre que l'intersection de trois concepts, symbolisée par la zone 3 est possible mais ne constitue qu'un cas de figure parmi les cinq envisageables. L'intersection des trois cercles correspond au cas d'un test d'évaluation fondé sur une mesure. En ce qui concerne les autres zones, chacune d'elles renvoie à des constellations diverses. La zone 1 signale qu'il est possible d'évaluer sans mesurer, ce qui peut être le cas non seulement dans les situations quotidiennes, mais aussi lors de l'évaluation de la performance des étudiants. Dans ce cas, la performance est évaluée à l'aide de grilles listant des critères et d'autres outils permettant une observation ciblée et structurée de la performance. Quant à la zone portant le chiffre 2, elle indique qu'on peut évaluer à partir d'une mesure sans forcément recourir à un test. Ce second cas de figure se produit, par exemple, lorsqu'un professeur a recours à un classement dans le but de noter les performances des étudiants (Doucet 2001 : 4). La zone 4, à l'intersection du test et de la mesure, rappelle que l'utilisation d'un test ne relève pas toujours d'une démarche évaluative. Les tests peuvent être utilisés à des fins autres que l'évaluation des compétences, notamment à des fins de recherche (Doucet 2001 :4). La dernière zone du graphique, portant le numéro 5, représente la possibilité de mesurer sans que

cela ait lieu dans le cadre d'un test ou d'autres procédures d'évaluation. Comme on vient de le voir, le recours à une mesure hors procédure d'évaluation est pratiqué lorsqu'on s'intéresse aux informations fournies par la procédure de mesure sans pour autant avoir besoin de prendre de décision à partir des données recueillies (Bachman 1990 : 23). L'intérêt du schéma ci-dessus est qu'il permet aux acteurs de l'évaluation de prendre en considération tous les cas de figure possibles, puisque différentes intersections sont envisageables entre test, mesure et évaluation. Comme nous avons pu le constater, le cas de figure classique, symbolisé par la zone 3 dans le graphique, n'est pas la seule option possible.

1.2 Les pratiques d'évaluation observées dans les différents domaines

1.2.1 Les pratiques d'évaluation dans l'enseignement supérieur

A notre époque, l'enseignement et l'apprentissage des langues sont concernés par de multiples réformes politiques sur les plans international et national, parmi lesquelles figurent l'influence croissante du *Cadre européen commun de référence pour les langues*, le Processus de Bologne ainsi que l'implémentation continue des portfolios européens de langues dans les cursus d'enseignement (Dervin & Suomela-Salmi 2007 : 10). Ces réformes ont engendré des modifications dans les pratiques didactiques, notamment, l'explosion de l'usage des nouvelles technologies d'information et de communication (NTIC) dans la formation en langues (Dervin & Suomela-Salmi 2007 : 10). Malgré la nécessité d'accompagner ces réformes d'une réflexion approfondie sur les questions d'évaluation, ceci n'a pas été suffisamment le cas (Dervin & Suomela-Salmi 2007 : 10). L'évaluation n'a pas profité de la nouvelle ère induite par les réformes pour se renouveler de façon fondamentale (Dervin & Suomela-Salmi 2007 : 10). Ceci est regrettable puisque les pratiques d'évaluation ont une grande influence sur tous les aspects liés à l'apprentissage d'une langue (Boud & Falchikov 2006: xviii). Pour certains chercheurs, les modifications dans le domaine des pratiques évaluatives auraient même un impact plus élevé sur l'apprentissage que les changements effectués dans les programmes d'enseignement (Boud & Falchikov 2006 : xviii). Il en résulte la nécessité de réfléchir aux innovations en

matière d'évaluation, non seulement pour faire évoluer les procédures mais pour améliorer l'apprentissage lui-même (Dervin & Suomela-Salmi 2007 : 10).

Puisque le contexte et les acteurs de l'enseignement-apprentissage dans le Supérieur diffèrent largement de ceux du secondaire, les pratiques évaluatives doivent également se différencier. Un trait distinctif essentiel de l'enseignement des langues dans le Supérieur est l'extrême hétérogénéité des catégories d'apprenants (Dervin & Suomela-Salmi 2007 : 10). Parmi les étudiants de langues, on trouve aussi bien des spécialistes de langues vivantes, des spécialistes d'autres disciplines que des apprenants souhaitant développer de simples compétences communicatives dans une langue étrangère. Tous ces groupes d'apprenants ont des objectifs d'apprentissage distincts. Les enjeux de formation en langues vivantes étrangères ne sont pas les mêmes, si bien qu'il faut envisager de fortes variations dans les méthodes d'enseignement-apprentissage et dans les pratiques évaluatives. Un autre facteur de diversité est la différence en matière de formation parmi les enseignants du supérieur (Dervin & Suomela-Salmi 2007 : 10). Tous n'ont pas nécessairement été formés aux spécificités de la pédagogie universitaire.

La variété des pratiques évaluatives dans les établissements d'enseignement supérieur se manifeste par l'usage de méthodes d'évaluation sommatives, formatives ainsi que de méthodes hybrides combinant ces deux méthodes, souvent dans le même département et par le même enseignant. Cette variété des méthodes d'évaluation ne constitue pas un problème en soi, mais peut le devenir si les évaluateurs ressentent un manque de contrôle dans ce domaine. Pareil sentiment peut s'installer pour des raisons objectives, comme l'ignorance d'une méthode ou le manque d'expérience d'un enseignant, mais peut également émerger suite à des convictions pédagogiques individuelles (Dervin & Suomela-Salmi 2007 : 10).

De tout ceci il ressort que l'évaluation en langues est aujourd'hui un domaine complexe englobant une multitude d'approches. Celles-ci ne sont pas interchangeables puisqu'elles sont tributaires des objectifs d'évaluation visés (Dervin & Suomela-Salmi 2007 : 22). Le point commun de toutes les méthodes d'évaluation contemporaines est que celles-ci ne servent plus uniquement à mesurer les connaissances, mais entendent accompagner le processus d'apprentissage lui-même. A ce titre, les méthodes d'évaluation sont devenues

de véritables « outils pédagogiques » qui ne sont pas nécessairement élaborés par l'enseignant (Dervin & Suomela-Salmi 2007 : 22). Cette dernière qualité peut constituer un problème car elle peut aussi expliquer le manque de contrôle du processus d'évaluation, qui se complexifie et qui peine à trouver une cohérence en poursuivant plusieurs objectifs à la fois. En effet, pour être pertinent et adapté, un acte d'évaluation doit simultanément respecter les objectifs d'apprentissage, tout en étant adapté aux qualités valables pour tous les dispositifs d'évaluation (Bachman & Palmer 1996 : 17).[1]

1.2.2 Les pratiques d'évaluation en didactique des langues vivantes étrangères

La question de l'évaluation attire tout particulièrement l'attention des chercheurs en didactique des langues, comme en témoigne la parution de divers textes dédiés à l'évaluation des compétences langagières, dont le plus connu est le *Cadre européen commun de références pour les langues*, abrégé CECRL (Huver & Springer 2011 : 7). Dès sa parution en 2001, le CECRL s'impose comme le document de référence incontournable pour définir et harmoniser l'évaluation des compétences en langues, dans l'espace européen. On constate que le CECRL, ainsi que d'autres textes de référence, s'inscrivent dans un paradigme objectiviste et quantitatif car leur but est de concevoir des dispositifs permettant une évaluation fiable et « juste » se prêtant à l'utilisation, indépendamment du contexte (Huver & Springer 2011 : 8). Bien qu'une grande latitude soit laissée aux concepteurs dans la réalisation de dispositifs d'évaluation des compétences, le cadrage procède par objectivation des compétences soumises à évaluation et listage de descripteurs précis.

En didactique des langues, la question de l'évaluation est traitée sous un angle méthodologique nécessairement inscrit dans un courant particulier (Huver & Springer 2011 : 8). Ce courant, qui dessine un cadre pour les dispositifs d'évaluation, relève plutôt de l'approche communicative dans certaines publications (Tagliante 1991 : 3) ou de l'approche communicative-actionnelle dans d'autres (Goullier 2005 : 5, Tagliante 2005 : 36). On note cependant que le récent numéro de la *Revue française de linguistique appliquée* (RFLA 2010) se distingue en « proposant des approches diversifiées susceptibles de fournir au

lecteur une vue plus globale de la multiplicité des problèmes que pose l'évaluation […] » (RFLA 2010 :5). Cette manière particulière de conjuguer diverses approches, effectuée par la RFLA, paraît raisonnable, vu qu'une analyse strictement méthodologique des outils de l'évaluation et de leurs usages a ses limites. Celles-ci sont dues au fait que pour comprendre les dispositifs et les pratiques d'évaluation, il faut non seulement analyser la méthodologie dont ces derniers relèvent, mais également prendre en compte leurs dimensions individuelles, sociales et politiques. La raison à cela est que les outils et les pratiques d'évaluation sont, au-delà d'une approche méthodologique particulière, toujours inscrites dans ces trois paradigmes opposés. Il en résulte que les pratiques d'évaluation basées sur un même outil peuvent néanmoins varier fortement en fonction de ces dimensions (Huver & Springer 2011 : 9).

1.3 Typologie des tests de langues: variations et convergences

1.3.1 Variations entre les tests des langues

Malgré l'existence de caractéristiques communes à tous les tests, les tests de langues sont loin d'être homogènes. Ils varient en fonction d'une multiplicité de paramètres qui sont l'objectif poursuivi, la méthode utilisée, les tâches contenues, les domaines linguistiques faisant l'objet d'une évaluation, les contraintes de temps imposées et les procédures d'attribution des scores (Douglas 2010 : 6). Les paramètres cités ne font pas l'objet d'un consensus entre les linguistes, car la liste des paramètres utilisés pour évaluer les tests varie d'une grille d'évaluation à l'autre. Néanmoins, il est couramment admis de classer les outils d'évaluation en différentes catégories, en fonction de leur but principal (Brown 2000 : 9). Un tel procédé peut être expliqué en recourant à deux arguments. En premier lieu, le but principal d'un test est la raison même de sa mise en place et de son usage (McNamara 2000 : 5). En deuxième lieu, la spécification du but principal constitue la première étape dans le processus contenant les étapes du choix, de la conception, de la révision et de l'adaptation de toute procédure d'évaluation (Brown 2010 : 9). Bien que cette position théorique soit dominante, elle n'est pas unanimement acceptée par les chercheurs. Selon certains linguistes, la catégorisation des tests doit s'appuyer

sur les types d'information fournis par les dispositifs d'évaluation (Hughes 2003 : 11). Il convient de réfléchir, sur la base des informations recueillies, à l'adéquation du test à un usage particulier, requis dans un contexte d'évaluation spécifique (Hughes 2003 : 11). La catégorisation présentée dans cette thèse s'appuie néanmoins sur le principe majoritairement admis qui accorde la prévalence à l'objectif principal visé.

Bien que les tests soient prioritairement classés en fonction de leur objectif principal, il est incontestable que toutes les caractéristiques évoquées sont pertinentes lors de l'élaboration et de l'évaluation d'un test. Il faut donc à la fois identifier l'objectif principal d'un dispositif et prendre en compte ses autres caractéristiques, décrites par plusieurs paramètres.[5] La connaissance de ces autres caractéristiques est tout aussi essentielle, car chacune de celles-ci contribue à la constitution de la typologie d'un test langagier.

1.3.2 Diversité de la typologie des tests en langues

Les tests de langue se rangent le plus souvent dans l'une des grandes catégories suivantes : positionnement, progrès, acquisition, compétence, diagnostic et aptitude (Alderson, Clapham & Wall 1995 : 9). Cette catégorisation est contestée par certains chercheurs qui ne distinguent que quatre types de tests, à savoir, les tests de positionnement, d'acquisition, de compétence, ainsi que les tests diagnostiques (Hughes 2003 : 12). Il faut préciser que dans cette seconde nomenclature, les tests d'acquisition englobent ceux de progrès (Hughes 2003 : 14).

1.3.2.1 Les tests de positionnement

Les tests de positionnement ont pour fonction essentielle de classer les candidats. Pour cette raison, ils sont également connus sous l'appellation « test de classement » (Laurier 1998 : 250). Cette fonction exige qu'on évalue la compétence langagière des candidats afin de les placer dans le niveau adapté d'un cursus de langue bien défini (Brown 2010 :10). Le classement des candidats en fonction de leurs compétences individuelles est primordial dans la mise en œuvre et le déroulement des programmes d'enseignement d'une langue. Néanmoins, la majorité des établissements sont confrontés à la difficulté de

placer leurs étudiants dans un groupe qui soit adapté à leur véritable niveau de compétences (Laurier 1998 : 250).

Le programme d'enseignement d'une institution particulière peut servir de base à la confection des tâches, mais ce n'est pas une obligation. Les tâches peuvent également impliquer un matériau indépendant du curriculum suivi (Alderson, Clapham & Wall 1995 :9). Mais en réalité, beaucoup de tests de ce type contiennent des tâches liées au curriculum enseigné. Cette catégorie est donc considérée comme liée au contexte (Purpura 2004).

Les tests de positionnement doivent contenir des tâches de niveaux différents pour que le classement établi soit aussi fiable pour les débutants que pour les candidats plus avancés (Laurier 1998 : 250). Le classement est effectué en plaçant les apprenants sur un « continuum » qui reflète la « progression » créée par le programme d'enseignement (Laurier 1998: 250). En conséquence, les niveaux couverts par un test doivent couvrir tout le continuum du programme d'enseignement proposé par une institution (Laurier 1998 : 250). Cette procédure permet de répartir les apprenants entre les divers groupes des niveaux (Laurier 1998 : 250). Cette manière de procéder, estime-t-on, permet d'effectuer un placement plus précis des candidats dans un groupe d'apprentissage (Brown 2010 :10). En cas d'adossement des tâches à un programme, les tests de positionnement incluent habituellement des échantillons des matériaux langagiers inclus dans les cours. De cette façon, on peut déterminer que les contenus sont ou non ajustés aux connaissances et compétences individuelles des candidats (Brown 2010 :10).

Les tests de positionnement peuvent adopter une variété de formats et évaluer des compétences différentes. Le choix dépend en premier lieu des caractéristiques et des besoins du programme d'étude dans lequel les candidats seront placés à l'issue du test (Brown 2010 : 11). En second lieu, ce choix dépend des domaines de compétences visés, liés à l'offre de formation des établissements d'enseignement (Laurier 1998 : 250). Dans la pratique toutefois, peu d'institutions sont en mesure d'évaluer le niveau des apprenants dans leurs différents domaines de compétences. C'est en effet le niveau de compétences général qui le plus souvent sert à repartir les candidats dans les différents cours (Laurier 1998: 250). Cette manière de procéder montre que la « fonction de

positionnement poursuivie par ces établissements est réductrice » (Laurier 1998 : 250).

Les tests de positionnement déjà disponibles sont génériques et donc imparfaitement adaptés aux besoins spécifiques d'une institution (Hughes 2003 : 16). Le test de positionnement idéal serait celui qui s'adapterait à une situation particulière. Son élaboration serait fondée sur l'identification des caractéristiques clés des différents niveaux d'enseignement couverts par un établissement particulier (Hughes 2003 : 17). Pour cette raison, il est d'habitude nécessaire de produire ce test localement afin d'obtenir une plus grande qualité (Hughes 2003 : 17). Or ce travail de confection demande non seulement un savoir-faire, mais du temps et de l'effort, si on veut parvenir à un placement précis des candidats (Hughes 2003 : 17). Une autre option, que nous avons envisagée lors de la confection de POSILANG, consisterait à prévoir des modules additionnels spécifiquement liés aux besoins d'une filière, en complément du test général.

Certains programmes de formation, proposés par une minorité d'institutions, disposant d'une infrastructure suffisante, offrent aux candidats une « évaluation par profil » (Laurier 1998 : 250). Un tel programme rend possible l'évaluation séparée des différentes compétences des candidats (Alderson, Clapham & Wall 1995:10). Celle-ci implique une évaluation qualitative des résultats par domaine de compétences, et non pas seulement l'établissement d'un niveau de compétence général. Dans ce type de programme, chaque étudiant est placé dans des cours axés sur le développement d'une compétence langagière particulière (Alderson, Clapham & Wall 1995:10). Les cours peuvent se situer à différents niveaux de compétence, de sorte qu'un candidat peut se retrouver dans un cours de prononciation de niveau débutant, par exemple, et dans un cours de compréhension écrite de niveau intermédiaire (Laurier 1998 : 250).

Il est également possible d'utiliser les tests de positionnement à d'autres fins que le placement des candidats dans un groupe d'apprentissage. Beaucoup de ces dispositifs ont également une fonction diagnostique. Ils fournissent une information sur les points forts et les déficiences d'un candidat. Bien que le rôle diagnostique des tests de positionnement soit secondaire, il est néanmoins très utile, en raison des indications fournies aux enseignants sur ce qui doit être appris ou accentué dans le cours à venir (Brown 2010: 10). Le diagnostic des

atouts et des points faibles permet également de décider si les candidats ont besoin de cours de soutien. Un tel usage du test de positionnement se rencontre dans beaucoup d'universités (Alderson, Clapham & Wall 1995:10). Au-delà des usages évoqués, les tests de positionnement sont également très utiles à des fins d'apprentissage en autonomie car l'obtention du niveau de compétences indique le meilleur point d'entrée (Laurier 1998 : 250).

1.3.2.2 Les tests d'acquisition et de progression

Les tests d'acquisition, ou *achievement tests* en anglais, sont fondés sur le programme suivi ou le manuel utilisé en cours. Par conséquent, les tâches de cette catégorie de tests sont limitées aux contenus des enseignements dispensés. Ces tests ont lieu au bout d'une période de temps définie, après une ou plusieurs séquences correspondant aux objectifs d'acquisition visés par le test (Alderson, Clapham & Wall 1995:10). Le but principal de cette catégorie de dispositifs consiste à déterminer si les objectifs visés par le programme ont été atteints et ainsi, si les connaissances et les compétences requises ont bien été acquises par les candidats au bout d'une période de temps donnée (Alderson, Clapham & Wall 1995:10). Leur fonction en tant que base d'enseignement et d'apprentissage constitue indéniablement le grand avantage de ce type de tests. Cependant, leur adossement à un programme constitue aussi une limitation qui peut être perçue comme négative. Cette restriction est constituée par le fait que ces dispositifs ne sont pas en mesure de prédire la performance future des candidats, à moins que le programme ait été conçu à cette fin (McNamara 2000 : 88). Il faut noter que le programme qui sert de base à cette catégorie de tests peut concerner non seulement une discipline, mais plusieurs matières, voire le système éducatif tout entier. C'est le cas des examens scolaires externes en Grande-Bretagne, par exemple (McNamara 2000 : 87).

Comme les tests de positionnement, les tests d'acquisition peuvent également avoir une fonction diagnostique en pointant les domaines langagiers qui requièrent un travail futur du candidat. Cependant, ce rôle diagnostique reste secondaire et n'est pas toujours exploité. Dans la mesure où ce type de test est administré à la fin d'une unité d'enseignement, il est souvent sommatif. Néanmoins, il a également une fonction formative parce que, pour être efficace, le test d'acquisition doit évaluer la performance des candidats après plusieurs

parties d'une unité d'enseignement (Brown 2010: 9). Cette catégorie de tests se caractérise par une très grande variété de durées, de types d'items et de formats utilisés (Brown 2010: 10).

Lorsqu'elle est utilisée correctement, cette catégorie de tests a un grand avantage, qui est de fonctionner en tant que soutien des activités d'enseignement et d'apprentissage. Cependant, le risque existe que la passation de ce test à la fin d'une unité d'enseignement focalise l'attention sur le contenu du test lui-même, aux dépens du reste des enseignements-apprentissages. L'utilisation d'un test standardisé de type QCM augmente ce risque (McNamara 2010 : 6). Même si le test reflète fidèlement le programme suivi, il existe un deuxième risque dont il faut tenir compte, à savoir, l'écart entre le test et l'usage réel de la langue. Ce risque apparaît en cas des tests focalisés sur des aspects particuliers de domaines linguistiques séparés, notamment la grammaire et le lexique. On peut tenter de corriger cette distorsion en ajustant le contenu du programme et du test lui-même à l'usage de la langue dans le monde social réel. L'effet souhaité, à savoir que le test reflète l'usage de la langue dans la situation cible, sera atteint uniquement si le test est véritablement adossé à un programme qui lui-même respecte les usages réels de la langue, ce qui est loin d'être toujours le cas (McNamara 2010 : 6).

Malgré ces risques, les tests d'acquisition présentent des avantages. Ils sont notamment ouverts à l'innovation. C'est la raison pour laquelle ils ont trouvé leur place dans le mouvement connu sous le nom *d'évaluation alternative* qui constitue une innovation dans le domaine de l'évaluation langagière (McNamara 2000 : 7). L'approche alternative met l'accent sur la nécessité d'intégrer l'évaluation aux objectifs de la formation et de responsabiliser les apprenants dans leur propre apprentissage. L'auto-évaluation fait partie de ce processus de responsabilisation puisqu'elle permet aux apprenants d'évaluer leur propre performance dans des contextes variés (McNamara 2000 : 7).

Les tests de progression sont similaires aux tests d'acquisition dans la mesure où ils sont adossés au programme suivi ou au manuel du cours. Toutefois, à la différence des tests d'acquisition, les tests de progression sont administrés tout au long du cursus d'apprentissage, et non seulement à la fin

d'une période, dans le but de voir les progrès réalisés par les candidats pendant une période intermédiaire (Alderson, Clapham & Wall 1995:10).

1.3.2.3 Les tests diagnostiques

Contrairement aux tests d'acquisition, les tests diagnostiques ont pour fonction de fournir une information sur les acquis, les manques ou les points faibles de l'apprenant avant le début d'un cours ou d'un module d'enseignement. Ce type de test est censé renseigner sur les aspects langagiers mal maîtrisés par l'apprenant, nécessitant la mise en œuvre d'un apprentissage (Hughes 2003 : 15). Pour cette raison, les tests diagnostiques offrent des informations généralement plus détaillées que les tests d'acquisition (Brown 2010: 10). Afin de fournir ce genre d'informations, les tests diagnostiques doivent contenir plusieurs items qui testent à la fois la connaissance et la capacité d'usage d'une même structure linguistique (Hughes 2003 : 15). Chaque structure linguistique doit être testée en différents contextes jugés importants pour évaluer la maîtrise de son usage (Hughes 2003 : 15). La prise en compte de ces impératifs dans les phases de conception et d'administration alourdit le dispositif, ce qui le rend peu pratique (Hughes 2003 : 15).

Les caractéristiques que nous avons énumérées expliquent pourquoi la conception d'un test diagnostique reste délicate (Alderson, Clapham & Wall 1995:10). Il existe en fait très peu de tests purement diagnostiques. Lorsqu'ils existent, estiment certains chercheurs, ils ne fournissent pas une information suffisamment détaillée et fiable (Hughes 2003 : 16). La fonction diagnostique est souvent accomplie par d'autres dispositifs, notamment ceux d'acquisition et de compétence. Mais il ne s'agit alors que d'une fonction secondaire, qui n'est pas systématique (Alderson, Clapham & Wall 1995:10). La difficulté qu'il y a à élaborer des tests purement diagnostiques concerne en particulier les tests spécifiques. Contrairement aux tests généraux, les tests spécifiques ont pour but de repérer les acquis et les manques d'un candidat dans un domaine langagier bien délimité, comme, par exemple, la grammaire (Alderson, Clapham & Wall 1995:10). Les déficiences repérées dans le domaine grammatical indiquent à l'enseignant quelles structures et fonctions langagières devront être étudiées dans les cours à venir.

Certains linguistes regrettent le manque de tests purement diagnostiques, car ceux-ci sont d'une grande utilité pour l'apprentissage autonome ou individualisé (Hughes 2003 : 16). Une fois que les lacunes des candidats ont été repérées, il devient plus aisé de les orienter vers des ressources et des activités adaptées à leurs besoins (Hughes 2003 : 16). L'utilité reconnue des tests diagnostiques suscite l'espoir qu'ils pourront être développés grâce aux nouveaux programmes numériques et administrés dans un format automatisé facilitant leur diffusion (Hughes 2003 : 16). DIALANG est à la fois un bon exemple de test diagnostique et la preuve qu'il est possible d'évoluer vers des dispositifs informatisés (Hughes 2003 : 16).

1.3.2.4 Les tests de compétence

Les tests de compétence ont souvent pour but d'évaluer la compétence langagière générale des candidats. Pour cette raison, on fait référence à ce type de tests par le terme anglais *proficiency*, qui implique la compétence et l'aisance générales dans tous les domaines de la langue (Brown 2010 : 4). Cependant, tous les dispositifs de ce type n'ont pas pour fonction de déterminer si un niveau donné de compétence langagière générale a été acquis par les candidats. Certains tests de cette catégorie ont pour objectif de déterminer si les candidats ont les compétences suffisantes dans un domaine particulier, comme la médecine ou le tourisme (Alderson, Clapham & Wall 1995:10). Ce dernier type de dispositif demande l'analyse de la langue requise pour un objectif spécifique (Alderson, Clapham & Wall 1995:10). Même lorsqu'on évalue un domaine de langue spécifique, le test n'est jamais limité à une seule compétence, par exemple, la compréhension de l'oral (Brown 2010 : 11). Il y a toujours au moins deux compétences langagières qui sont visées par l'évaluation. Pour donner un exemple, les nouveaux tests TOEIC, publiés en 2011, ont pour objectif d'évaluer deux compétences : la compréhension orale et la compréhension écrite (*Tests complets pour le nouveau TOEIC* 2011 : VI-VII).

Contrairement aux tests d'acquisition, aucune référence n'est faite à un processus d'enseignement antérieur dans ce type de test, car l'enjeu n'est pas d'évaluer l'acquisition de compétences au travers d'un programme donné. N'étant pas tributaires de contenus d'enseignement particuliers, les tâches du

test de compétence dépendent uniquement des définitions théoriques générales données de telle ou telle compétence (Purpura 2004).

Ce que les tests de compétence cherchent à évaluer est donc l'usage réel de la langue, en mode hors programme. Pour pouvoir évaluer cet usage, il faut que le test lui-même se rapproche autant que possible de la réalité sociale de la langue, comme l'illustrent les tests conçus durant les deux dernières décennies (McNamara 2000 : 7). Par exemple, un test de compétence ayant pour fonction d'évaluer la capacité à utiliser la langue dans le secteur de l'aviation doit forcément intégrer des caractéristiques relevant de l'usage de la langue dans ce secteur d'activité précis. La première contrainte est d'évaluer la compétence communicative des candidats, vu que l'usage réel de la langue des professionnels d'aviation a un caractère éminemment communicatif. La deuxième contrainte, résultant également des caractéristiques de l'usage réel de la langue dans ce secteur d'activité, est de privilégier les actes de communication orale.

Même les tests de compétence traditionnels, réputés pour leurs tâches standardisées au format QCM, évoluent vers des activités de production en proposant des échantillons de performance écrite et orale. L'un de ces dispositifs le très connu *test d'anglais langue étrangère* (TOEFL), évalue la production écrite et orale des candidats (Brown 2010 : 11). Il est pourtant vrai que ce n'est pas la fonction traditionnelle de ces tests d'évaluer la compétence communicative. Traditionnellement, ces tests ne balaient pas l'entier du spectre de la performance communicative des candidats. Ils évaluent essentiellement les compétences réceptives ainsi que les habiletés grammaticales et lexicales des participants (Brown 2010 : 11). Pour ce faire, ils utilisent typiquement des instances de langue décontextualisées ou peu contextualisées.

Puisque ces tests évaluent la compétence des candidats par rapport à un niveau de langue prédéterminé constituant une norme, ils sont sommatifs à référence normative (Brown 2010 : 11). La référence normative implique que les réponses des candidats soient comparées à des réponses fixes et préparées, qui constituent la norme. On obtient un seul score, sous forme de chiffre numérique ou de pourcentage, qui sert à placer les candidats dans un ordre de classement.

Les tests de compétence ne prévoient pas de fournir de retour diagnostique sur un candidat car leur rôle ultime est d'admettre ou d'empêcher le passage de quelqu'un à un niveau d'instruction avancé ou à un poste professionnel particulier (Brown 2010 : 11). En dehors de l'évaluation de la compétence globale, le but de cette catégorie de tests est de certifier un niveau de compétence particulier ou de refuser d'attester un niveau, en cas de défaillance par rapport à la compétence cible.

1.3.2.5 Les tests d'aptitude

Cette catégorie de tests a pour but de mesurer la capacité générale d'apprendre une langue et de prédire les chances de succès, en mode a priori, autrement dit en amont de la formation. Ces tests sont donc censés mesurer des dispositions et un potentiel, ce qui est toujours délicat (Brown 2010 : 12). Ils sont utilisés rarement aujourd'hui, sauf lorsqu'on s'interroge sur l'incapacité d'un sujet à apprendre une langue étrangère. Ce type de situation demeure cependant exceptionnel. Par ailleurs, en l'absence de cause manifeste objectivable, il est hasardeux de se prononcer sur le succès ou l'échec attendus avant même que le sujet ait débuté son apprentissage (Brown 2010 : 12). Ces tests sont donc utilisés de nos jours pour renseigner les candidats sur leurs styles d'apprentissage préférés, sur leurs points forts ou faibles potentiels, ainsi que sur les stratégies qui leur conviennent (Brown 2010 : 12).

1.3.3 D'autres paramètres de variation entre les tests

Comme nous l'avons vu, l'analyse d'un test ne doit pas se limiter à ses seuls objectifs. Il est tout aussi nécessaire de prendre en considération et d'évaluer ses autres caractéristiques. Certes, les tests restent classés selon leur objectif principal, mais leur typologie révèle qu'ils sont porteurs d'une multiplicité des caractéristiques tout aussi pertinentes. Voilà pourquoi une analyse détaillée est indispensable. Elle seule permet de comparer les dispositifs.

Pour être en mesure d'analyser et de comparer les différents tests en langues disponibles sur le marché de façon non seulement détaillée, mais aussi homogène, il faut recourir à une grille d'évaluation comparative. Il existe des modèles divers de grilles d'évaluation qui varient en fonction des paramètres

utilisés. Pour donner un exemple de grilles comparatives d'évaluation, deux modèles différents seront présentés dans ces pages.

1.3.4 La grille d'évaluation « maison »

Le premier modèle de grille que nous présentons ici est un modèle qui a été conçu dans le cadre du projet région « Didactique des langues. Ressources numériques et hybridations » financé par l'Université Bordeaux Montaigne et la Région Aquitaine (2011-2014). Cette grille maison a été élaborée dans le but d'évaluer plusieurs tests de positionnement en anglais disponibles sur le marché dont l'analyse sera décrite et discutée par la suite. A l'origine, il s'agissait de répondre à la première phase du projet, qui visait à faire un état des lieux de l'existant en matière de tests de positionnements. La grille « maison » (ou « projet région ») que nous avons construite a été approuvée par le responsable du projet, sans préjuger d'améliorations ultérieures possibles. Priorité fut donnée à la lisibilité, à la maniabilité et au pouvoir de synthèse. La présentation claire des résultats d'évaluation nous a paru aussi importante que la procédure d'analyse elle-même, ce document devant pouvoir être inséré dans des dossiers institutionnels d'aide à la décision. La diversité des paramètres retenus garantit le recueil de résultats variés, tandis qu'une bonne lisibilité permet de comparer facilement les différents dispositifs. La grille « projet région» est insérée ci-après.

Nom du test	
Source	
Prix	
Auteur et année d'apparition	
Fonction du test	
Public visé	
Contraintes (contrôle externe, temps, lieu, date fixe)	
Langues testées	
Niveaux couverts	
Objectifs	
Conception du test	

Compétences testées	
Contenu des tâches	
Format des tâches	
Mode d'évaluation	
Transparence des critères d'évaluation	
Forme du bilan (texte, tableau, grille)	
Elaboration (contenu) du bilan (conseils, points forts et faibles)	
Ma propre Evaluation	

Tableau 1 – Grille « maison » (ou « projet région »)

Les premiers paramètres inclus dans la grille visent à saisir les données administratives d'un test, à savoir, son nom, la source, le prix, son auteur et son année officielle de lancement. Les catégories intervenant par la suite sont de nature méthodologique et linguistique. Leur évaluation est importante lors de l'analyse d'un test afin de comprendre la typologie de ce dernier. Il s'agit de catégories comme la fonction du test, le public visé, les contraintes imposées, notamment le contrôle externe, le temps attribué, le lieu de passation, l'attribution éventuelle d'une date fixe. Suivent les langues testées, les niveaux couverts, la conception et les objectifs du test, les compétences évaluées, le contenu des tâches ainsi que les méthodes appliquées. Les objectifs incluent l'objectif principal ainsi que les objectifs secondaires. Figurent ensuite le mode et les critères d'évaluation, ainsi que le contenu et la forme du bilan. La dernière catégorie évoquée dans la grille concerne ma propre évaluation sommative du test.

Les paramètres choisis incluent un repérage des formes, des catégories et des fonctions langagières des tests, sans qu'il s'agisse à ce stade de procéder à une évaluation qualitative de leur pertinence. Les principes d'évaluation qualitative des tests, au nombre de six, sont détaillés plus loin dans notre étude.

1. Les données administratives

Les données administratives permettent seulement d'identifier le test, non de l'évaluer. Ni le concepteur (qui reste généralement en retrait et a rarement la visibilité d'un auteur), ni l'année de parution, ni l'éditeur ne sont des indices fiables de qualité. Le prix éventuel n'indique rien d'autre que l'inscription du dispositif dans une logique de développement et de commercialisation privée. Il existe d'ailleurs des tests gratuits qui sont considérés comme de très haute qualité par la communauté d'experts. Un bon exemple est DIALANG qui est un système d'évaluation totalement gratuit fournissant aux utilisateurs des informations diagnostiques sur leur compétence en langues. La haute qualité de ce dispositif d'évaluation résulte de son adossement direct au CECRL (Conseil de l'Europe 2001 : 161). Ce test est libre, quoique d'accès instable et aléatoire (changements de serveurs, incompatibilité avec certains systèmes de navigation, manque de maintenance). De même, l'année de parution ne conditionne pas nécessairement la qualité ou la pertinence d'un dispositif. Si on considère à nouveau DIALANG, on note qu'il a été lancé en 2001. Ce test, pourtant ancien, est encore considéré comme l'un des tests de positionnement les plus accomplis et le mieux adossé au CERCL (Conseil de l'Europe 2001 : 161). A maint égard, DIALANG reste une référence incontournable quinze ans après son lancement, alors que la technologie informatique et la réflexion ont progressé.

2. La fonction du test

La fonction d'un test est son but principal, la raison même de sa conception et de sa mise en œuvre. Chacun des types de tests présentés dans la sous-partie précédente a une fonction différente. Comme on a pu le voir, la fonction des tests de positionnement est de placer l'apprenant dans un cours de langue correspondant à son niveau de compétence. Ce n'est donc pas un hasard si ce paramètre apparait en premier dans notre grille. Pour mémoire, la fonction figure en premier rang dans le modèle des spécifications reconnues d'Alderson, Clapham & Wall (1995 : 11)[6].

3. Les objectifs

Dans notre grille, les objectifs de chaque test apparaissent dans une rubrique séparée de la fonction. Les objectifs comprennent à la fois les objectifs primordiaux et les objectifs secondaires d'un dispositif d'évaluation. Les tests de positionnement, par exemple, peuvent avoir d'autres buts en dehors du placement des candidats dans un cours de langue approprié. L'objectif secondaire souvent poursuivi par ce type de tests est d'identifier les points forts et les déficits dans la compétence des participants (Brown 2010: 10).

4. Public visé

Au même titre que la fonction, le public visé par un test varie d'un dispositif à l'autre. Les catégories concernées par ce paramètre sont les caractéristiques personnelles des candidats. L'enjeu est ici de déterminer à qui précisément s'adresse le test, car ce facteur a nécessairement un impact sur les contenus. Il s'agit principalement de l'âge, de la langue maternelle ainsi que du niveau de maitrise estimé de la langue cible. Les tests de positionnement, présentés dans le sous-chapitre qui suit, s'adressent à un public adulte large et ne sont donc pas prévus pour être administrés dans un contexte scolaire. Concernant la langue maternelle, les tests d'anglais analysés au moyen de la grille s'adressent à des locuteurs non natifs. L'autre caractéristique individuelle ciblée par ce paramètre est le niveau de compétence du candidat en langue évaluée. Pour que la passation du test soit réalisable et pour qu'elle ait un sens, le niveau du candidat doit être estimé en amont et jugé adapté au répertoire de niveaux que couvre le test. Il faut en effet savoir que certains dispositifs n'offrent qu'une évaluation restreinte de niveaux, comme le *Cutting Edge Placement Test* (examiné en détail plus loin).

5. Les Contraintes

Les contraintes sont un autre paramètre qui doit forcément être appliqué lors de l'analyse d'un test. Il faut distinguer deux types généraux de contraintes : d'une part celles qui concernent les concepteurs et les administrateurs du test, d'autre part celles qui doivent être prises en compte par les candidats.

Nous nous intéresserons d'abord aux contraintes pouvant survenir lors du développement et de l'administration d'un test. Celles-ci se définissent comme les limitations des ressources nécessaires à la bonne administration de l'instrument d'évaluation (Fulcher & Davidson 2007 : 128). Il convient de partir de la compréhension des ressources requises pour le développement et l'administration d'un dispositif. Ces ressources se répartissent en quatre catégories: les ressources humaines, physiques, financières ainsi que la sécurité du test. Concernant les ressources humaines, il s'agit du personnel et des compétences disponibles. Les ressources physiques comprennent l'équipement local ainsi que les technologies d'information et de communication qui sont à la disposition des responsables pour le développement et l'opérationnalisation du test (Fulcher & Davidson 2007 : 128). Les ressources financières sont nécessaires pour assurer la disponibilité et le bon fonctionnement des autres catégories de ressources, d'ordre humain et physique.

La sécurité d'un test est également intégrée aux ressources. Pour garantir un bon déroulement des opérations, il faut veiller à ce que les contenus détaillés de toutes les versions du test soient tenus secrets jusqu'au moment de l'épreuve (McNamara 2000 :24). Mais le bon déroulement n'est pas seul en cause. La sécurité est indispensable pour la validité des scores et la prise de décision qui en découle (Fulcher & Davidson 2007 : 128). L'importance de la sécurité explique pourquoi cette contrainte a trouvé sa place dans l'ouvrage *The Standards for Educational and Psychological Testing:* « Reasonnable efforts should be made to assure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means » (AERA 1999: 64).

Selon McNamara (2000), la fonction d'un test peut également faire partie des contraintes. Il n'est pas rare, de nos jours, que les Etats et leurs administrations exigent des rapports de réussite à des tests de langues (par exemple pour intégrer un établissement d'enseignement supérieur lorsqu'on est étranger, pour demander un titre de séjour ou monter un dossier de naturalisation). Cette demande plus politique doit être intégrée par les responsables de la conception de tests afin que la construction et la fonction de ce qu'ils proposent soient compatibles avec les attentes ou prescriptions institutionnelles de ce type (McNamara 2000 :24).

Tous les systèmes d'évaluation sont concernés par la limitation de certaines ressources. Puisque leur influence sur le développement et l'administration d'un dispositif est inévitable, il est essentiel de les repérer dès le début de la conception du test (Fulcher & Davidson 2007 : 128).

Un second type de contrainte concerne les candidats lors de l'administration d'un test. La contrainte peut prendre la forme d'un contrôle externe lors de la passation, ou encore être de nature spatio-temporelle. Le contrôle externe impose la mise en place de procédures lors de l'administration d'un test pour garantir la sécurité des participants. Les contraintes spatiales se traduisent le plus souvent par l'imposition d'un lieu fixe d'administration exigeant le déplacement des candidats. Les contraintes temporelles incluent la durée ainsi que l'imposition d'une date ou d'une période de passation. Les autres caractéristiques physiques comme le niveau sonore, la température, les conditions d'illumination, le placement dans la salle, peuvent également être considérées comme autant de contraintes liés aux candidats parce qu'ils sont confrontés à celles-ci et doivent y faire face (Bachman & Palmer 1996 : 48). Pour l'analyse des tests à l'aide de la grille « projet région », seules les contraintes du deuxième type ont été prises en compte. En effet, les contraintes concernant les responsables du développement et de l'administration d'un test ne sont pas repérables de l'extérieur. Elles n'ont donc pas pu être intégrées à notre évaluation.

6. Les langues testées

C'est un paramètre qui varie largement en fonction des tests. Leur nombre peut varier entre une seule et 14, le nombre maximum atteint à l'heure actuelle, étant celui évalué par le système DIALANG. A l'exception de ce dispositif, les tests évalués dans le cadre du projet de recherche évoqué, portent sur une seule langue qui est l'anglais.

7. Les niveaux couverts

Pour désigner les niveaux de compétence des tests analysés, nous avons eu recours aux niveaux communs de référence déterminés par le CECRL. Il s'agit des niveaux suivants : Introductif (A1), Intermédiaire (A2), Seuil (B1), Avancé

(B2), Autonome (C1), Maîtrise (C2). Les deux premiers constituent les niveaux qui doivent être atteints par un utilisateur élémentaire, les deux niveaux intermédiaires correspondent à un utilisateur indépendant. Enfin, les deux niveaux supérieurs renvoient à un utilisateur expérimenté.

Il est important de noter que tous les tests ne sont pas automatiquement adossés aux niveaux communs de référence du *Cadre Européen Commun de Référence pour les Langues*. La référence varie considérablement d'un dispositif à l'autre. Si certains tests sont explicitement construits autour des niveaux communs de référence, d'autres en dévient de façon considérable. Les tests comme *DIALANG*, mais aussi *Oxford Quick Placement Test* et *Oxford Placement Test 2* font partie de la première catégorie. En revanche, d'autres dispositifs ne correspondent pas aux niveaux de référence distingués par le *Cadre Européen*. Ce peut être parce qu'ils couvrent un spectre de niveaux plus limité que ce référentiel, ou alors parce qu'ils ne distinguent pas les mêmes niveaux de compétence. Les dispositifs comme *Cutting Edge Placement Test* et *Success Placement Test* illustrent ce second cas de figure : ils n'incluent pas les niveaux C1 et C2 d'un utilisateur expérimenté, ils ajoutent un niveau de compétence supplémentaire (*pré-intermédiaire*) aux deux niveaux intermédiaires distingués par le CECRL. Ce niveau est situé entre les niveaux A2 et B1 et correspond donc au niveau A2+ du CECRL.

8. Conception du test

Ce paramètre renvoie à la structure d'un test qui peut être définie comme la manière dont ses différentes parties sont combinées et présentées aux candidats (Bachman & Palmer 1996 : 51). Cette définition suscite plusieurs interrogations. La première concerne le nombre de parties ou de tâches dont un test se compose (Bachman & Palmer 1996 : 51). Il est important de noter qu'un instrument d'évaluation peut se composer non seulement de parties simples, mais de tests complets englobant plusieurs parties à leur tour. C'est le cas du *Cutting Edge Placement Test,* par exemple, qui comporte cinq tests différents (A-E). La deuxième interrogation porte sur le nombre et le degré de différenciation entre les parties et les tâches. La troisième porte sur l'ordonnancement des parties ou des tâches, qui a un impact manifeste sur la structure du test. Il est

important de déterminer si les parties ou les tâches d'un test exigent une suite de réponses fixes ou variables, c'est-à-dire si on est autorisé ou non à placer les parties dans un ordre libre (Bachman & Palmer 1996 : 51). Le dernier enjeu qui a une influence sur la structure d'un test est la conception de chacune de ses parties. Pour l'analyser, il convient de prendre en considération le nombre des tâches contenues dans chaque partie d'un test (Bachman & Palmer 1996 : 51). Lors du recensement et de l'évaluation des tests disponibles sur le marché, deux questionnements dans l'ensemble qui vient d'être décrit ont été considérés. Il s'agit du nombre de parties ainsi que de la quantité des tâches contenues à l'intérieur de chacune d'elles.

Les tâches, quant à elles, se distinguent selon plusieurs critères qui sont leur nombre, leur ordonnancement séquentiel ainsi que le format et le contenu de chaque tâche donnée. Dans la mesure où seuls le nombre et l'ordonnancement des tâches structurent de façon visible et globale la conception du test, priorité a été donnée à ces deux catégories dans la rubrique « conception du test ». Cela ne signifie pas pour autant que les deux autres critères, le contenu et le format des tâches, aient été évincés. Nous les avons seulement redéployées vers des rubriques ultérieures : respectivement « contenu des tâches » et « méthodes du test ». Même si ces critères sont séparés dans la grille d'évaluation, pour des raisons de commodité, elles restent de toute évidence liées, car le contenu et le format des tâches ont une influence directe sur leur nombre. Ainsi, les items à *réponse fixe* sont facilement multipliables et ont donc toutes les chances d'apparaître en nombre plus élevé que d'autres items.

9. Compétences testées

Les tests varient selon les compétences évaluées, leur nombre et l'ordre dans lequel celles-ci sont testées. Il y a un désaccord entre les chercheurs sur la question de savoir si on peut mesurer les habiletés langagières séparément l'une de l'autre. Certains experts expriment leur scepticisme par rapport à ce point de vue, en raison de l'usage communicatif de la langue en contexte réel qui demande la combinaison de plusieurs compétences (Douglas 2010 : 19). Selon ce point de vue, l'objectif doit être de mesurer la compétence langagière générale au travers d'habiletés particulières : la compréhension de l'écrit, la production de

l'écrit, la compréhension de l'oral, la production de l'oral, la compétence lexicale, la compétence grammaticale. La procédure qui consiste à mesurer ces habiletés de manière séparée, en attribuant aux candidats un score pour leur performance dans chaque domaine de compétence, est contestée (Douglas 2010 : 19). Malgré ces réserves, malgré la nécessité de conjuguer plusieurs habiletés en situation de communication réelle, la plupart des tests procèdent à des évaluations séparées. C'est le cas pour tous les dispositifs analysés dans le cadre du projet « Didactique des Langues : ressources numériques et hybridations ». Les dispositifs se distinguent pourtant par le fait que certains tests évaluent les compétences lexicales et grammaticales, tandis que d'autres n'évaluent que les compétences communicatives. Ces particularités seront décrites et analysées en temps utile. Il sera par ailleurs tenu compte de l'ordre dans lequel l'évaluation des différentes compétences a lieu, comme l'exige la procédure car, comme on a pu le voir, la disposition des tâches avait un impact sur la structure du test tout entier (Bachman & Palmer 1996 : 51).

10. Contenu des tâches

Les caractéristiques des tâches contenues dans un test constituent un facteur aussi important que les compétences langagières des candidats. Ces deux facteurs exercent la plus grande influence sur les scores obtenus (Purpura 2004 : 100). Les tâches possèdent un ensemble de caractéristiques uniques. Bien que le contenu et le format des tâches fassent partie de ces caractéristiques, ces deux composantes ont été séparées dans la grille présentée pour des raisons de précision. Le format des tâches est saisi par le paramètre « méthodes du test », décrit par la suite. Les caractéristiques des tâches peuvent être spécifiées de différente façon, selon différents degrés de précision.

Il existe des cadres conçus pour décrire les caractéristiques des tâches d'un test de manière très détaillée, par exemple celui de Bachman & Palmer (1996). Dans ce cadre très complexe, les différentes parties d'une tâche sont considérées et analysées séparément l'une de l'autre, en ayant recours à un grand nombre de critères précis. Ces parties sont les *instructions fournies*, l'*input,* les *réponses attendues* ainsi que le *lien entre l'input* et *les réponses attendues.* Les caractéristiques de l'input et des réponses attendues sont évaluées selon le

même spectre de critères, à savoir, le format, les caractéristiques langagières et les caractéristiques thématiques. Lors de l'analyse des tests en question, exposés dans la suite de ce chapitre, seules les caractéristiques langagières des réponses attendues ont été considérées. Ce choix résulte de la décision d'effectuer une analyse brève et concise.

11. Format des tâches

Le format des tâches concerne la manière dont les candidats sont incités à traiter les matériaux soumis (McNamara 2000 : 26). La méthode peut être définie comme la manière dont les tâches sont conçues (McNamara 2000 : 5). Le format est l'une des deux composantes de la méthode d'un test. Outre le format des tâches, la méthode englobe la manière d'attribuer les scores et d'évaluer les réponses des candidats.[7] En termes de format, deux larges catégories de tests peuvent être distinguées. La première catégorie se distingue par un format qui présente un nombre fixe de réponses possibles pour les candidats. Il existe plusieurs types de formats à réponse fixe. Le *questionnaire à choix multiples* est considéré comme le plus important de ces formats à cause de sa fréquence (McNamara 2000 : 5). Ces tests sont typiquement utilisés pour évaluer la réception à l'oral et à l'écrit ainsi que la grammaire et le vocabulaire (McNamara 2000 : 5).

La seconde catégorie de dispositifs sont les tests de performance. Ceux-ci ne recourent pas aux réponses fixes, mais évaluent les compétences langagières à travers un acte de communication. Les tests de performance sont habituellement utilisés pour évaluer les compétences en production orale et écrite des candidats. Pour ce faire, un échantillon de la production orale ou écrite est élicité et, ensuite, soumis au jugement d'un ou de plusieurs évaluateurs qui utilisent une procédure d'évaluation par critères (McNamara 2000 : 6). Pour pouvoir tirer les conclusions sur les compétences langagières dans la situation de communication cible à partir des performances mises en évidence dans un test, les échantillons de langue sont obtenus en simulant des tâches réelles dans des contextes réalistes (McNamara 2000 : 6).

Comme nous l'avons déjà signalé, les méthodes ont été analysées séparément du contenu des tâches, ce qui nous a permis de rester concis dans

l'analyse des dispositifs. Pareille séparation est un choix pragmatique, non une obligation. En effet, il existe deux façons de concevoir le lien entre le contenu d'un test et la méthode. La première considère la méthode comme une partie intégrante du contenu tandis que la deuxième, plus répandue, traite la méthode de manière indépendante du contenu (McNamara 2000 : 26). L'approche intégrative est partagée par plusieurs spécialistes de l'évaluation en langues. Le format de la réponse attendue fait ainsi partie des critères d'évaluation des tâches élaboré par Bachman & Palmer (1996 :50). De même, Purpura (2004) affirme que le format utilisé fait partie des caractéristiques des tâches (Purpura 2004 : 100). Ces positions corroborent le résultat des recherches effectuées dans les années 1980 qui ont révélé un effet de méthode considérable (Bachman & Palmer 1982a : 462, Shohamy 1983). La méthode appliquée pour obtenir une performance a un impact important sur les scores obtenus par les candidats (Purpura 2004 :101).

12. Le mode d'évaluation

Le mode d'évaluation concerne la manière d'attribuer des points et d'interpréter les scores. Ce paramètre varie en fonction des tests. Le mode d'évaluation le plus fréquent consiste à indiquer le nombre de bonnes réponses sous forme de chiffre ou de pourcentage. Cependant, il y a des tests dont le mode d'évaluation ne prévoit pas l'affichage d'un score, comme par exemple DIALANG. Dans ce système d'évaluation, le niveau de compétence est déterminé uniquement à partir du nombre de réponses fausses.[8] Comme il a été évoqué dans le sous-chapitre précédent, le mode d'évaluation est considéré comme faisant partie intégrante de la méthode du test, avec son format : « the term test method covers these aspects of test design together with the issue of how candidate responses will be rated or scored » (McNamara 2000 :26).

Quel que soit le mode d'évaluation adopté dans un test, les scores attribués ne résultent jamais de la seule compétence langagière des candidats, mais de tout un ensemble de facteurs dont les principaux sont les attributs personnels et les schémas affectifs des candidats, leurs connaissances thématiques, l'utilisation des stratégies ainsi que les caractéristiques du test lui-même (Purpura 2004 : 100). Il faut bien avoir conscience que les caractéristiques

du test constituent un facteur aussi important que les compétences langagières des candidats, ces deux facteurs exerçant un impact majeur sur les scores obtenus (Purpura 2004 : 100). Enfin, le mode d'évaluation a également une influence très importante sur le test et détermine le degré de transparence des critères d'évaluation.

13. La transparence des critères d'évaluation

Ce paramètre a été inclus dans la grille car il existe un lien fort entre la transparence d'évaluation et les qualités d'un test, notamment la perception de la validité et de l'équité du test par les candidats.

La transparence de l'évaluation se manifeste par l'affichage du score atteint par chaque candidat ainsi que par la séparation (*cut off*) entre niveaux de compétence différents. En outre, il faut veiller à ce que le score soit bien présenté et compréhensible pour les candidats. Or, le résultat ne devient compréhensible que lorsque le candidat peut faire le lien entre le score obtenu et le nombre de bonnes ou mauvaises réponses.

Cependant, il ne suffit pas d'afficher le score atteint, le niveau et le relevé des réponses justes ou fausses pour rendre une évaluation transparente. Il faut s'entendre sur les qualités de la performance considérées comme décisives et déterminer les critères pour les juger (McNamara 2000 : 36). Le repérage des critères pertinents implique la prise en compte des composantes de la compétence ainsi que la pondération de chacun des éléments (McNamara 2000 : 36). Pour que la transparence de l'évaluation soit réelle et partagée, il faut non seulement que les concepteurs soient au clair sur les critères adoptés mais qu'ils en informent les candidats.

Une autre forme de transparence est la transparence dans l'attribution des scores aux différentes parties d'un test. Ces scores se combinent et déterminent le score final (Doucet 2001 : 15). En deuxième lieu, la pondération exerce une forte influence sur l'interprétation des résultats et peut altérer ces derniers (Doucet 2001 : 15). Cela explique pourquoi ces questions doivent être résolues clairement lors de la conception d'un test. Il faut se demander si l'incidence relative des parties séparées reflète fidèlement l'importance relative des

différentes composantes de la compétence communicative ou s'il y a d'autres règles auxquelles elle répond. Ainsi, l'élaboration de la pondération peut se fonder sur des considérations théoriques, idéologiques, stratégiques, politiques ou une combinaison de ces facteurs (Doucet 2001 : 15). En raison de la signification majeure de la pondération lors de la mesure de la performance, une transparence des pratiques est nécessaire.

14. Forme du bilan

Le bilan est une partie du test qui n'est pas moins importante que les instructions et les items. Pour cette raison, il mérite d'être pris en considération et évalué lors de l'analyse des tests. Le bilan se compose de deux composantes, de forme et de contenu. Chaque paramètre est suffisamment important pour être saisi séparément. En ce qui concerne la forme, le bilan peut prendre l'apparence d'un tableau, d'une grille ou simplement d'un texte courant. On ne peut pas dire sous quelle forme il est préférable de rédiger le bilan. Cela dépend de l'étendue des informations à transmettre ainsi que de la préférence de ses concepteurs.

15. Contenu du bilan

Le bilan peut être plus ou moins détaillé. Il doit inclure, au minimum, le score atteint et l'évaluation du niveau de compétence d'un candidat. Cette rubrique peut aussi contenir une liste de points forts ou faibles repérés dans la performance individuelle, offrir la possibilité de voir ses réponses encore une fois ou juste les fautes. Il faut souligner qu'un bilan donnant un feedback détaillé sur la performance est d'une haute valeur pour le candidat, car elle lui permet de comprendre sa performance très vite après l'évaluation et ainsi favorise son apprentissage.

16. Ma propre évaluation du test

Cette rubrique de synthèse, qui comporte une part de subjectivité assumée, vise à évaluer globalement le test, à établir sa valeur en tant qu'instrument d'évaluation des compétences, à s'interroger sur les bénéfices qu'il est susceptible d'apporter aux candidats.

1.3.5 La diversité de l'impact des tests langagiers

Toutes les variables ne sont pas saisies par les paramètres inclus dans la grille qui vient d'être présentée. L'un des éléments de variation est l'impact produit par le résultat obtenu au test : impact sur la société, sur les institutions, sur les personnes responsables des évaluations, ainsi que sur les candidats eux-mêmes (Fulcher & Davidson 2007 :372). En ce qui concerne les candidats individuels, le résultat à un test peut avoir un impact sur la possibilité des individus concernés de poursuivre leurs études ou leur carrière. Ce phénomène est lié à celui de la décision, abordée lors de la définition du test. L'impact est corrélatif au but principal d'un instrument d'évaluation et en conséquence, à sa typologie. On distingue un test *high stake*, ayant un impact élevé sur l'avenir éducatif ou professionnel des candidats, d'un dispositif *low stake*, n'ayant pas un tel effet (McNamarra 2000 :48). Tandis que les tests diagnostiques ont peu d'impact sur l'avenir des candidats, les tests de positionnement se distinguent par un impact moyen car ces derniers ont pour objectif de placer les candidats dans un programme d'études. Entre ces deux catégories évoquées se trouvent les tests d'acquisition et de progrès. Les dispositifs qui ont l'impact le plus élevé sur l'avenir des candidats sont les tests de compétence car leur passation mène soit à la délivrance d'une récompense, sous forme d'une certification, soit au refus de délivrer ce document à un candidat.

1.3.6 La grille utilisée dans les centres de langue en Italie

Le second modèle de grille est la grille du projet Innova-Langues[9] qui a servi à une analyse comparée des tests de positionnement en italien. Ces tests sont employés par les centres linguistiques des universités en Italie. Les tests de positionnement pris en compte dans le cadre de l'analyse comparative sont ceux utilisés par les universités italiennes qui ont répondu à l'appel, à savoir, *l'Università Ca' Foscari Venezia, Università degli Studi di Parma* et *l'Université de Sassari*. Les paramètres ont été sélectionnés pour décrire les aspects didactiques, méthodologiques et organisationnels d'un test. Il y a onze paramètres inclus qui apparaissent dans la grille dans l'ordre suivant.

Accès et inscription	
Auto-évaluation initiale	
Le temps à disposition pour compléter le test	
Type de test	
Approche didactique	
Typologie des exercices	
Compétences et habiletés visées	
Nombre total d'items dans le test	
Consignes fonctionnelles (le degré de clarté, l'emploi de métalangage, le temps verbal utilisé)	
Feed-back final	
Points de force et de faiblesse	

L'analyse comparative des tests de positionnement en italien a permis d'établir le bilan qui sera présenté par la suite.

1.3.7 Le bilan de l'analyse comparative des tests de positionnement en Italien

En ce qui concerne l'accès et l'inscription, les tests en ligne nécessitent une inscription préalable. Les tests papier demandent d'insérer les coordonnées du candidat. Concernant l'auto-évaluation initiale, 80% des tests n'incluent pas cette phase. Le reste des dispositifs prévoit une auto-évaluation en ligne. Le temps à disposition pour compléter le test est de 120 minutes en moyenne. Dans 75 % de cas, le temps est explicité au début du test, tandis qu'un test sur 4 ne le mentionne pas. Concernant leur typologie, 60% des tests sont progressifs, 20%

adaptatifs et 20% ne sont pas adossés au *Cadre européen commun de référence pour les langues*. La totalité des tests sont *low-stake*, c'est-à dire qu'ils n'ont pas un impact important sur le devenir des candidats. L'approche didactique choisie est grammaticale dans 65 % des cas et communicative dans 35%.

En ce qui concerne la typologie des exercices, le format le plus courant est la « réponse à sélectionner» (*selected response*) (70%). Ce format inclut le QCM, le Vrai ou Faux (V/F) et l'appariement. 20% des exercices sont de type « production courte » (*limited production*). Le procédé consiste à compléter un texte lacunaire. Enfin, 10% des exercices sont du type « production longue » (*extended production*).

En ce qui concerne les compétences et les habiletés visées, la compétence écrite est évaluée par 80% de tests, en utilisant le format QCM ou V/F. La production écrite est testée par la totalité des dispositifs au travers des tâches de type « limited production » ou « extended production ». En revanche, la production orale n'est pas évaluée par 90% des tests.

En ce qui concerne le nombre d'items, leur distribution est hétérogène et fluctue en fonction des niveaux. La plupart des items se situent entre les niveaux A1 et B1. Le nombre d'items proposé aux trois niveaux supérieurs, B2 à C2, est nettement plus réduit. Il faut noter que le score est indiqué à la fin des exercices proposés, mais non les paramètres d'évaluation.

Pour ce qui est des consignes fonctionnelles, à savoir le degré de clarté, l'emploi de métalangue et les formes verbales utilisées, elles sont brèves et suffisamment claires. Le mode dominant est l'impératif singulier. Quant au feedback final, 75% des tests ne l'autorisent pas. Parmi les 25 % qui le donnent, seulement 10% prévoient un feedback détaillé et d'autres dispositifs indiquent uniquement le niveau de compétence évalué.

Globalement, il est possible d'identifier les points de force et de faiblesse de ces tests italiens. Les points forts, qui concernent uniquement les tests en ligne, sont une grande variété d'exercices (facilités par le support interactif) et la possibilité d'un parcours personnalisé quand le test est adaptatif. Par ailleurs, lorsqu'une approche communicative est adoptée, un large éventail de

compétences est évalué. Les points faibles sont le manque d'adossement de certains items au CECRL, le nombre limité de compétences évaluées lorsque l'approche est grammaticale, l'impasse faite sur la production orale par la quasi-totalité des tests ainsi que le manque de feedback pour les étudiants.

1.3.8 Convergences

Malgré les variations relevées, on constate qu'il existe un certain nombre de convergences entre les tests, qui transcendent leurs différences. La convergence principale concerne la nature de ce qu'on peut mesurer au moyen de tests langagiers ainsi que la finalité même de la procédure de mesure. En effet, le but ultime, commun à tous les instruments d'évaluation, est de tirer des conclusions sur la compétence langagière d'un candidat à partir de sa performance au test, souvent désignée par le mot anglais *outcome* (Douglas 2010 : 19). Ce lien établi entre une performance ponctuelle à un test et une compétence langagière générale ne va pas forcément de soi mais il est accepté par la plupart des linguistes faisant de la recherche dans le champ de l'évaluation des compétences langagières (Bachmann & Palmer 1996 ; Fulcher & Davidson 2007). La prise de décision, nous l'avons vu, constitue un autre élément partagé par tous les types des dispositifs d'évaluation (Douglas 2010 :19- 20). La décision à prendre varie forcément en fonction du type de test.

Un autre point de convergence entre les diverses catégories de tests concerne les qualités principales de l'évaluation langagière: la praticité, la fiabilité, la validité, l'authenticité et leur impact sur l'enseignement. Non seulement celles-ci doivent être respectées par tous les tests formels, mais elles doivent être applicables à toute forme d'évaluation. Ces qualités constituent autant de principes qui déterminent la convenance, l'efficacité et au final l'utilité d'un test (Brown 2010 : 25). Ces qualités, ainsi que leur mise en pratique dans les instruments d'évaluation, seront expliquées par la suite.

1.4 Les qualités des tests langagiers

Dans les sous-chapitres suivants, les six qualités valables pour tout type de test seront définies et décrites, à savoir, la praticité, la fiabilité, la validité, l'authenticité, l'interactivité et enfin l'impact des tests sur l'enseignement et l'apprentissage (Bachman & Palmer 1996 : 17).[10] Ces qualités doivent être respectées tant au moment de l'évaluation que lors de la conception des tests. Il est important de noter que ces principes s'appliquent non seulement aux tests formels mais à tout type d'évaluation (Brown 2010 : 25).

La prise en considération et le respect de ces six qualités sont extrêmement importants parce qu'ils permettent d'évaluer l'utilité d'un test (Bachman & Palmer 1996 : 17). L'examen de ces six qualités servira à la fois à évaluer les dispositifs existants et à concevoir notre propre test « POSILANG ». Avant d'aborder les six qualités évoquées, il convient de décrire le modèle d'utilité ainsi que les principes nécessaires pour opérationnaliser ce modèle lors du développement et de l'usage de tests.

1.4.1 L'utilité d'un test

L'utilité est considérée comme la qualité la plus importante d'un test car l'usage auquel un dispositif d'évaluation est destiné doit être la préoccupation première lors de sa conception et de son développement (Bachman & Palmer 1996 : 17). L'usage du test qui est fait par une institution pour évaluer des compétences langagières engage, à un niveau ou à un autre, l'avenir du candidat. Le caractère primordial de l'utilité tient également au fait que cette qualité constitue une métrique qui doit s'appliquer lors de l'évaluation de tous les aspects liés au développement et à l'usage des tests (Bachman & Palmer 1996 : 17). En tant que qualité majeure d'un dispositif d'évaluation des compétences langagières, l'utilité doit servir de base aux contrôles de qualité tout au long du développement d'un test (Bachman & Palmer 1996:17). Un test de bonne qualité doit être efficace, approprié et utile (Brown 2010 : 25).

Selon la position de recherche dominante, les six qualités d'un test doivent être envisagées dans une relation de complémentarité (Bachman & Palmer 1996 : 18 ; Hughes 1989). La relation de complémentarité signifie que l'utilité du

dispositif doit être considérée à partir de la résultante de ces qualités, non de leur superposition : Utilité= Fiabilité + Validité de construit+ Authenticité+ Interactivité+ Impact + Praticité (Bachman & Palmer 1996 : 18). Une telle relation implique que, malgré la tension pouvant exister entre les qualités individuelles, un équilibre optimal doit être trouvé entre celles-ci. Pour parvenir à cet équilibre, il faut prendre en considération l'effet combiné des différentes qualités sur l'utilité d'un test particulier au lieu d'évaluer l'impact individuel de ces qualités sur l'utilité. La recherche d'un équilibre entre les six qualités a pour conséquence qu'on cherche à déterminer leur niveau minimum acceptable. Cela aide à éviter les deux extrêmes possibles qui consistent soit à favoriser quelques-unes de ces six qualités aux dépens des autres, soit à essayer d'atteindre un niveau maximal pour chacune d'entre elles (Bachman & Palmer 1996 : 134).

Trois principes sont à la base de ce modèle d'utilité, qui permettent de l'opérationnaliser lors de la conception et de la passation des tests. Le premier postule qu'il faut maximiser l'utilité globale d'un test et non pas les qualités individuelles. Il en résulte que les qualités individuelles doivent être évaluées au travers de leur effet combiné sur l'utilité globale et non pas indépendamment les unes des autres. Par ailleurs, l'utilité d'un dispositif particulier et l'équilibre entre ses différentes qualités ne peuvent pas être étendues à tous les tests en général. La situation d'évaluation et la démarche d'évaluation doivent être considérées dans leur spécificité (Bachman & Palmer 1996 : 18). Outre son caractère variable, l'appréciation de l'utilité globale d'un dispositif est subjective. Car c'est bien le concepteur qui décide quelles sont les qualités à optimiser dans le test qu'il conçoit. La prise de décision ne dépend donc pas uniquement d'éléments objectivables, comme le type de test et la situation d'évaluation spécifique, mais résulte également d'intuitions et de jugements plus personnels émanant directement du concepteur (Bachman & Palmer 1996:19). Il n'en demeure pas moins qu'un test sera d'autant plus utile qu'il sera développé en tenant compte de son objectif spécifique, du groupe des candidats particuliers à qui on le destine et de la situation spécifique de l'usage de la langue. Dans Bachman & Palmer (1996), un questionnaire permettant d'interroger chaque critère d'utilité d'un test est fourni. Ce questionnaire entend faciliter l'évaluation logique ou

conceptuelle de l'utilité pour aider les concepteurs à créer leurs dispositifs (Bachman & Palmer 1996 : 135).

1.4.2 Praticité

Ce principe concerne les questions logistiques et administratives qui entrent en jeu lors de la conception du dispositif, ou encore lors de l'administration des épreuves et de l'attribution des scores. Il y a plusieurs facteurs qui déterminent le degré de praticité d'un test. Parmi les principaux figurent : le respect des limites budgétaires imposées, l'administration et l'évaluation des passations, la mise à disposition des candidats du temps nécessaire à l'effectuation des tâches, l'énoncé de directives claires, ainsi que l'usage adéquat des ressources humaines et matérielles disponibles (Brown 2010 :26). Les ressources décisives pour évaluer la praticité se subdivisent en trois types : les ressources humaines, les ressources matérielles et le temps nécessaire au développement, jusqu'à la première administration opérationnelle (Bachman & Palmer 1996 : 37). Les coûts associés à chaque type de ressources doivent être calculés au plus près. Il faut souligner que la gestion des ressources disponibles dépasse la seule disponibilité de temps et d'effort nécessaires pour la conception du test ainsi que pour l'attribution des scores. La disponibilité seule ne garantit en rien un usage intelligent des ressources (Brown 2010 : 26). Il faut souligner que les types et l'étendue des ressources requises varient en fonction de la situation d'évaluation (Bachman & Palmer 1996 : 135). En résumé, la praticité est à définir comme la relation entre les ressources requises pour la conception, le développement et l'usage du test, d'une part, et les ressources disponibles pour ces activités, d'autre part (Bachman & Palmer 1996 : 34). On peut représenter cette relation au moyen de la formule suivante :

$$\text{praticité} : \frac{\text{ressources disponibles}}{\text{ressources requises}}$$

Selon cette formule, le développement et l'usage d'un test sont pratiques si le rapport est supérieur ou égal à 1. Si tel est bien le cas, les ressources requises n'excèdent pas les disponibilités et le dispositif est viable. Dans le cas contraire, lorsque les ressources requises sont supérieures aux ressources disponibles, la

praticité du test doit être remise en cause. Deux options existent alors pour corriger la situation : l'une est de diminuer la part des ressources requises en modifiant les spécifications ; l'autre consiste à augmenter les moyens disponibles ou à les allouer différemment afin d'augmenter leur efficacité (Bachman & Palmer 1996 : 34).

Il résulte de la définition fournie qu'il existe un niveau limite de praticité. Ce niveau constitue dans les faits un niveau minimal acceptable. L'existence d'un niveau limite montre que, contrairement aux cinq autres composantes de l'utilité, la praticité n'est pas une qualité continue. L'impossibilité d'attribuer un degré plus ou moins élevé de praticité constitue un trait distinctif de cette qualité (Bachman & Palmer 1996 : 135). Un deuxième trait distinctif est que la praticité ne se réduit pas à un score, à la différence des cinq autres qualités. La praticité répond essentiellement aux questions : le test peut-il être développé et utilisé? Si oui, sous quelles conditions et de quelle manière (Bachman & Palmer 1996 : 135)?

1.4.3 Fiabilité

La caractéristique principale d'un test fiable est la cohérence de la mesure qu'il permet d'effectuer (Bachman & Palmer 1996 : 19). Cette qualité est extrêmement importante pour l'évaluation à large échelle parce qu'il est connu que toute mesure inclut une marge d'erreur. Par ailleurs, il existe un lien évident entre la fiabilité et la validité d'un test, dans la mesure où la fiabilité est indispensable pour qu'un test soit déclaré valide (Hughes 2003 : 50). La fiabilité est fonction de la cohérence des scores obtenus aux tests et aux tâches des tests différents (Bachman & Palmer 1996 : 20). C'est au prix de cette cohérence que le test sera en mesure de fournir une information valable sur les compétences des candidats (Bachman & Palmer 1996 : 20). Une première façon de concevoir la cohérence est d'imaginer qu'un même test puisse être administré à un même candidat ou à un même groupe plusieurs fois dans le temps et dans l'espace, avec des résultats comparables (à compter que leur niveau n'ait pas eu l'occasion d'évoluer au travers d'un apprentissage et hors effet de tâche). Cette exigence (qui reste largement théorique) implique que le résultat atteint une fois est potentiellement reproductible, toutes choses étant par ailleurs égales :

« Whenever a test is administered, the test user would like some assurance that the results could be replicated if the same individuals were tested again under similar circumstances. » (Crocker & Algina 1986: 105). Une deuxième façon de concevoir la cohérence est l'interchangeabilité : deux formes de test réputées équivalentes doivent fournir des résultats comparables (Bachman & Palmer 1996 : 20).

En dehors de la cohérence des scores, la fiabilité se manifeste par plusieurs autres critères. Les principaux sont la cohérence des conditions d'administration, l'absence d'ambiguïté des tâches, ainsi que la formulation de directives claires et uniformes pour l'attribution et l'évaluation des scores. Au-delà de l'uniformité, les rubriques portant sur l'attribution des scores et l'évaluation des résultats doivent se prêter à une application cohérente lors des différentes administrations du test (Brown 2010 : 27). L'absence d'ambigüité dans les items est primordiale car il y a un lien très net entre le nombre d'items ambigus dans un test et son degré de fiabilité (Alderson, Clapham & Wall 1995 : 87)

Bien que la fiabilité soit une qualité primordiale, il faut reconnaître l'impossibilité d'éliminer complètement les incohérences dans un test. On désigne par incohérence des « changements asystématiques », c'est-à dire, des variations de scores qui ne reflètent pas des variations de niveau de compétence chez les candidats, mais qui dépendent d'autres facteurs, par exemple, des états psychologiques des candidats (Alderson Clapham & Wall 1995: 87). Malgré l'impossibilité d'éliminer complètement ces incohérences, il est souhaitable d'en contrôler les sources potentielles et d'en minimiser les effets (Bachman & Palmer 1996 : 20). Le but de l'évaluation est d'élaborer des tests qui mesurent les changements de niveau de compétences des candidats, non les « changements asystématiques» (Alderson, Clapham & Wall 1995 : 87). Le degré de fiabilité d'un test dépend de la proportion des « changements systématiques » dans le score (Alderson, Clapham & Wall 1995 : 87). Le test est d'autant plus fiable que la proportion de ces derniers est élevée.

Le degré de fiabilité peut être influencé dans un sens ou dans l'autre par les caractéristiques propres au test. Cinq facteurs majeurs ayant un effet significatif sur la cohérence des scores ont été identifiés. Les tâches, d'abord, qui

sont en partie contrôlables. En effet, lors de la conception d'un test, il est possible de réduire l'incohérence en minimisant les variations entre tâches, dès lors que ces variations ne sont pas liées aux usages réels et pluriels de la langue (Bachman & Palmer 1996 : 20). Le deuxième facteur de cohérence est le nombre d'items contenus dans un test. Une augmentation de ce nombre entraîne mécaniquement une augmentation de la cohérence des scores. Cependant, il est important de ne pas inclure un nombre trop élevé d'items car cela peut avoir des effets psychologiques néfastes sur certains candidats et induire, par contrecoup, une moindre fiabilité du test (Brown 2010: 29). Le troisième facteur concerne la variation de la difficulté des items. La difficulté identique des items réputés de même niveau augmente le niveau de fiabilité (Fulcher & Davidson 2010 : 106). En quatrième lieu, il est nécessaire de choisir des sujets hétérogènes lors de la phase de pilotage (Fulcher & Davidson 2010 : 106). Le cinquième facteur de cohérence est le format lui-même du test. Les tests « objectifs » ayant un ensemble de réponses fixes, préparées d'avance, sont réputés posséder une fiabilité plus grande que ceux qui sont « subjectifs » avec leurs réponses ouvertes, appelant un jugement de l'évaluateur (Brown 2010: 29). Cependant, les tests au format QCM, présumés plus « objectifs », requièrent une attention particulière lors de leur conception. Pour assurer une réelle fiabilité à cette « objectivité », il faut veiller à ce que tous les items aient vraiment le même niveau de difficulté, qu'ils soient correctement distribués et que toutes les options proposées à l'intérieur de chaque item aient été élaborées soigneusement (Brown 2010: 29).

Il existe quatre facteurs qui peuvent être à l'origine d'un manque de fiabilité d'un test. Ces facteurs sont liés au candidat, à l'évaluateur, aux procédures d'administration du test ou au test lui-même (Brown 2010 : 27). Ces facteurs ont en commun de modifier le score obtenu pour des raisons autres que la compétence, de sorte que le résultat observé ne correspond pas au véritable score de l'individu (Fulcher & Davidson 2010 : 105). Or, pour pouvoir tirer des conclusions valables sur les compétences d'un candidat, sur la seule base du score obtenu à un test, il faut que ledit score reflète fidèlement ses capacités et performances. Des méthodes numériques ont donc été mises au point pour

calculer l'écart type entre le score obtenu et le score véritable lors des passations différentes (Fulcher & Davidson 2010 : 105).

En ce qui concerne la non-fiabilité imputable au candidat, on constate que ce sont surtout des facteurs physiques ou psychologiques qui interviennent. Bien que ces facteurs puissent paraître au-delà du contrôle des administrateurs et des évaluateurs, il existe des stratégies capables d'en minimiser la cause et l'impact. En effet, ce type de non-fiabilité est souvent lié aux conditions dans lesquelles un test est administré. Or celles-ci ne sont pas irrémédiables. La non-fiabilité liée aux procédures d'administration concerne généralement les conditions matérielles dans lesquelles le test a lieu. Celles-ci incluent tous les aspects susceptibles d'avoir un impact sur la performance des candidats et donc des scores attribués. A l'évidence, de mauvaises conditions de passation ont un effet immédiat sur l'état psychologique des candidats. Ce constat est encore plus évident lorsque les dégradations concernent non seulement le lieu de passation mais les items inclus dans le corps du test (Brown 2010 : 29).

Lorsque le déficit de fiabilité est causé par les évaluateurs, il convient de déterminer si on est en présence d'une ou deux personnes, engagées dans l'attribution de scores incohérents. Dans le premier cas de figure, il s'agit très probablement d'un facteur interne, fréquent dans les tests ne prévoyant pas un ensemble normé de réponses correctes. On ne peut non plus exclure l'inattention, la fatigue, ou encore des biais par rapport à certains candidats dont le correcteur n'est pas forcément conscient. Une autre cause envisageable est le manque de clarté des critères d'attribution des scores (Brown 2010: 28). Lorsqu'il y a deux correcteurs et que des écarts sont constatés, il ne faut pas exclure que l'un d'eux au moins ne respecte pas les critères de notation, manque d'attention ou d'expérience, lorsqu'il ne souffre pas de préjugés à l'encontre de la population testée (Brown 2010 : 28).

Il est très précieux de connaître les sources possibles d'incohérence des scores. En premier lieu, pour tenter de réduire ou de contrôler les facteurs de perturbation de la fiabilité. Mais surtout pour intégrer l'existence d'une marge d'erreur à toute prise de décision fondée sur un score (Fulcher & Davidson 2010 : 114).

1.4.4 Validité

Avant d'être une théorie, la validité est une pratique : tel ou tel dispositif est jugé nationalement ou internationalement probant pour évaluer des connaissances ou des compétences, en vue de prononcer l'admission universitaire ou l'aptitude professionnelle d'un candidat (Chapelle 1999 : 255). Lorsque l'utilisation du dispositif se fait à très large échelle, auprès de publics nécessairement hétérogènes et dispersés, il est nécessaire d'avoir la garantie que le dispositif est bien adapté aux objectifs et aux candidats particuliers concernés (Fulcher & Davidson 2007 : 23). La situation est toutefois différente lorsque le dispositif est destiné à un usage plus restreint et ciblé, comme cela est le cas en situation classe. La méthodologie d'enseignement et les domaines à évaluer sont alors précisés par un manuel ou un programme donnés (Spolsky 1975 : 255).

Mais qu'entend-on exactement par « validité » ? De nos jours, la validité, aux côtés de la fiabilité, est considérée comme l'une des deux qualités principales de tout test réputé efficace (Alderson, Clapham & Wall 1995 : 7). Depuis les travaux de Cronbach & Mehl (1955), les recherches sur la validité sont devenues un axe majeur de l'évaluation langagière, psychologique et éducative (Fulcher & Davidson 2007 : 10). Ce paramètre est en effet aussi complexe qu'essentiel. Dans les années 1970, la validité est assimilée à l'authenticité (Chapelle 1999 : 256). Par exemple, un test de langue, pour être probant, doit permettre de vérifier une compétence effective, qui soit authentiquement exportable en situation réelle. En termes plus généraux, la validité peut être définie comme le degré auquel les conclusions tirées des résultats d'une évaluation sont pertinentes, significatives et utiles à la lumière du but de l'évaluation (Gronlund 1998 : 226). Cette définition souligne que la validité est une propriété étroitement tributaire de l'objectif du test. La notion d'objectif apparaît d'ailleurs dans d'autres définitions de la validité, plus axées sur le processus de mesure.

> 'Validity in testing and assessment has traditionally been understood to mean discovering whether a test 'measures accurately what it is intended to measure' (Hughes 1989 : 22), or uncovering the 'appropriateness of a given test or any of its component parts as a measure of what it is purposed to measure' (Henning 1987 : 170).

S'interroger sur la validité d'un instrument d'évaluation c'est donc déterminer si celui-ci mesure effectivement ce qu'il entend mesurer, autrement dit s'il est en phase avec les objectifs qu'il se donne (Fulcher & Davidson 2007 : 4).

Les principes contenus dans les définitions citées sont résumés par les deux arguments de Cronbach & Mehl (1955) qui n'ont cessé d'exercer une influence sur les chercheurs. Le premier de ces arguments stipule que la définition du construit est centrale dans la conception des tests et l'évaluation des compétences (ibid : 282). Le deuxième argument met l'interprétation des scores obtenus au cœur de toute recherche sur la validité (ibid : 300). La question majeure pour interpréter les scores est donc quelle preuve peut être fournie qui légitime leur interprétation (Fulcher & Davidson 2007 : 10) ? L'enjeu est majeur car il faut rassembler suffisamment de données susceptibles d'être acceptées comme preuves manifestes d'une compétence en langues, par le très large public des formateurs et des utilisateurs. Si les preuves fournies par le test sont convaincantes, alors celles-ci pourront être utilisées pour défendre l'usage du dispositif et l'interprétation qui est faite des scores (Fulcher & Davidson 2007 : 10). Il faut cependant avoir conscience que ce qui possède le statut de preuve à un instant *t* peut changer avec le temps, si bien que la lecture des scores doit être entendu comme un processus évolutif (Fulcher & Davidson 2007 : 10).

Les définitions de la validité que nous avons citées jusqu'ici énoncent clairement que pour être valable un test doit évaluer exactement les compétences qu'il se donne pour objectif. Cela suppose que soient systématiquement ignorées les autres variables, non pertinentes. Par ailleurs, pour être valable un test doit aussi livrer des informations permettant de tirer des conclusions précises à partir des scores obtenus (Alderson, Clapham & Wall 1995 : 7). Enfin, un test probant doit se fonder sur des données empiriques obtenues au travers de la performance du candidat. Les conclusions que l'on tire de ces données et des scores qui leur sont associés doivent être soutenues par des arguments théoriques (Brown 2010 : 30). En résumé, pour acquérir une validité, un test doit fournir des informations utiles et significatives sur les capacités réelles d'un candidat.

Il est important de souligner que la validité n'est pas une notion absolue, mais un principe qui doit être respecté à des degrés divers: « […] validity is a matter of degree, not all or none.» (Messick 1989 : 33). Traditionnellement, on distingue trois types de validité, liés au types de données récoltées au travers de test : une validité orientée vers l'usage réel de la langue, désigné *criterion* en

anglais, une validité de contenu et une validité de construit (Chapelle 1999 : 255). Cependant, cette tripartition a été réfutée par Samuel Messick dans son article de 1989 :

> Traditional ways of cutting and combining evidence of validity, as we have seen, have led to three major categories of evidence: content-related, criterion-related, and construct related. However, because content-and criterion-related evidence contribute to score meaning, they have come to be recognized as aspects of construct validity. In a sense, then, this leaves only one category, namely, construct-related evidence (Messick 1989: 20).

La dernière phrase montre que Messick ne reconnaît qu'un seul type de preuve pour établir la validité, le construit, qui englobe les autres. La théorie de Messick opère donc l'union de ce que ses prédécesseurs dissociaient. Elle est donc désignée par l'appellation "unified validity framework » (Messick 1989 : 13). Si nous citons avec quelque détail cet article c'est qu'il fut jugé subversif en son temps. Les positions défendues par Messick ont fondamentalement changé la manière dont est conçue la validité dans son domaine de recherche, et ce jusqu'à nos jours (Fulcher & Davidson 2007 : 12). Pour Messick, la validité n'est pas une propriété du test, mais le degré auquel on est habilité à construire une inférence à partir de la performance mesurée et à prendre une décision particulière sur la base du score obtenu (Fulcher & Davidson 2007 : 12) : "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick 1989 : 13).

Malgré son rôle capital, l'article de Messick (1989) n'est ni la première ni la seule publication à rejeter la classification traditionnelle de la validité. Dans *Standards for educational and psychological testing* (AERA et al. 1985), la définition des trois types de validité est déjà remplacée par le concept de validité unifié [11] (AERA et al. 1985). Comme chez Messick, la validité du construit est conçue comme centrale, tandis que les deux autres types sont présentés comme des méthodes servant à l'explorer (Chapelle 1999 : 256).

Le concept de validité de Messick va cependant au-delà de l'unification des trois types de validité traditionnels. Il englobe aussi les conséquences d'un test. Comme la validité, la notion de conséquence est complexe. Messick y inclut

plusieurs concepts préalablement forgés par d'autres linguistes : l'affect, décrit par Madsen (1983), l'impact décrit par Hughes (1989) ainsi que l'éthique, soulevé par Canale (1987) (Chapelle 1999 : 257).[12] Pour montrer que la validité est un concept unifié, bien que composé de facettes multiples, Messick a développé une matrice de la validité progressive. Celle-ci contient les deux facettes de l'évaluation : la justification et la fonction de l'évaluation. La justification de l'évaluation englobe elle-même deux aspects: les données sur lesquelles se fonde l'interprétation et l'usage des tests, ainsi que les conséquences qui en résultent. La fonction d'évaluation comprend l'interprétation et l'usage des tests (Messick 1989 : 20).

	Test interpretation	**Test use**
Evidential basis	Construct validity	Construct validity + Relevance/utility
Consequential basis	Value implications	Social consequences

Facets of validity (Messick 1989 : 20)

Selon Messick, les barrières entre les catégories de la matrice sont perméables, de sorte que chaque catégorie à droite ou en bas inclut la même information que la précédente, mais avec des ajouts respectifs (Messick 1989 : 20). La théorie de Messick est devenue le paradigme accepté de l'évaluation psychologique, éducative et langagière adopté par les travaux publiés dans les années 1990 (Chapelle 1999 : 257). L'idée d'inclure les conséquences de l'usage d'un test, aussi bien son impact que la responsabilité sociale, dans le concept de validité a engendré de vifs débats parmi les experts (Chapelle 1999 : 257).

Il existe trois approches de la validité reconnues par la communauté des experts (Chapelle 1998 : 34). Elles se distinguent essentiellement par le sens donné aux scores. La première approche est la théorie traditionnelle des traits. Elle postule que le score obtenu correspond bien aux attributs des candidats dans le domaine de compétences couvert par le test. Ces attributs sont estimés stables et réels (Chapelle 1998 : 34).

La deuxième approche, néo-behavioriste, voit dans le score un indicateur de contexte, englobant notamment le lieu de l'évaluation, le sujet et les

participants. Pour pouvoir tirer des conclusions valables sur la compétence d'un candidat, les aspects de la situation cible doivent être répliqués dans le test le plus précisément possible (Chapelle 1998 : 34).

La troisième approche, interactionnelle, traite la performance observée durant le test comme représentative des performances générales, en contexte identique: « performance is viewed as a sign of underlying traits, and is influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts » (Chapelle 1998 : 34). Le test a ici le statut d'un échantillon.

Bien que la tripartition du concept de validité en catégories soit dépassée depuis Messick (1989), nous estimons utile d'en conserver les articulations. Celles-ci restent à la fois pertinentes pour intégrer les travaux antérieurs à Messick et pour mettre en évidence les diverses facettes actuelles du concept unifié.

1.4.4.1 Validité de contenu

Un test satisfait au critère de validité de contenu s'il demande au candidat de véritablement effectuer la performance mesurée. Puisque ce type de validité tient au contenu du test, on fait référence à ce type de preuve par l'expression « validité liée au contenu » (Brown 2010 : 30). On peut identifier ce type de preuve de manière empirique, en observant et définissant la performance qui est mesurée. Pour donner un exemple, il est évident qu'un test visant à évaluer la compétence des candidats en interaction orale ne doit pas se composer d'items écrits au format QCM. Il existe deux façons, directe et indirecte, de tester la performance des candidats. Lorsque le test est direct, le candidat accomplit la tâche cible directement. Lorsque le test est indirect, ce n'est pas la tâche cible qui doit être effectuée, mais une tâche liée à celle-ci d'une façon ou d'une autre. Quelle que soit la stratégie, il est demandé au test de faire preuve de validité de contenu. La manière optimale de tester la compétence des candidats dépend de nombreux facteurs, en particulier le type de test choisi, le profil des candidats et les compétences qu'on cherche à évaluer. Dans de nombreux cas, il est possible d'évaluer les compétences indifféremment d'une manière ou l'autre, mais il arrive aussi que l'une des deux stratégies s'impose. Par exemple, si le test est destiné

à groupe classe, il est conseillé de tester la compétence de façon directe, après vérification préalable que le test en question couvre bien les objectifs du cours (Brown 2010 : 32). Beaucoup de tests de compétence standardisés manquent de validité de contenu, car ils sont réduits en contexte et donc ne demandent pas aux candidats de montrer toutes les habiletés constitutives de la performance communicative (Brown 2010 : 31).

1.4.4.2 Validité du test liée à l'usage réel de la langue

Cette forme de validité n'est reconnue au test que si les scores obtenus sont vraiment utilisés pour prédire la performance du candidat en situation d'usage réel de la langue (Fulcher & Davidson 2010 : 5). En pronostiquant la performance pour des situations à venir, les scores indiquent si les objectifs visés par le test ont été satisfaits ou non. Il est nécessaire que soit explicité ce que les concepteurs entendent par usage réel de la langue dans tel type de situation. La mesure doit permettre de constater qu'un niveau préalablement déterminé a bien été atteint (Brown 2010 : 32). Le test satisfait ce critère de validité si le dispositif proposé est conforme à ce type d'objectif et s'en rapproche avec un maximum de précision (Douglas 2010 : 63). Il faut noter que ce type de validité est le plus pertinent pour les tests conçus et administrés en classe. La situation est autre pour les tests standardisés à référence normative (Brown 2010 : 32).

Il convient de distinguer deux catégories au sein de ce type de validité : concurrentielle et prédictive. Un test est porteur de validité concurrentielle si les résultats obtenus sont soutenus par la performance extérieure à celle évaluée dans le dispositif. Un test est porteur de validité prédictive s'il permet d'évaluer et de pronostiquer le succès futur des candidats dans l'apprentissage de la langue. Dans la mesure où les tests de positionnement, au même titre que les tests d'admission à l'université ou de sélection pour un poste de travail, ont pour fonction de se prononcer sur les chances de succès des candidats dans l'avenir, il est très important qu'ils satisfassent au critère de validité prédictive (Brown 2010 : 33).

1.4.4.3 Validité liée aux construits d'un test

Un test repose sur un cadre théorique, explicite ou non, contenant plusieurs construits. De façon générale, le construit se définit comme une théorie, une

hypothèse ou un modèle qui tente d'expliquer les phénomènes observés (Brown 2010 : 33). Le construit d'un test désigne les compétences qui sont à la base d'une tâche. C'est ce construit qui permet l'interprétation des scores obtenus pour ladite tâche (Bachman & Palmer 1996 : 23). La validité des construits d'un test implique l'évaluation du degré auquel le test mesure effectivement les concepts sous-tendus. Ainsi est-on en mesure de déterminer si les scores obtenus sont interprétables comme des indicateurs avérés des compétences ciblées (Alderson, Clapham & Wall 1995 : 13). Pour valider les concepts construits, le cadre théorique sur lequel repose le test doit être explicité. Les relations entre les concepts construits, au sein du test, doivent être clairement identifiées, ainsi que la relation entre le cadre théorique général et les objectifs principaux du test (Alderson, Clapham & Wall 1995 : 13). Ce type de validité est absolument essentiel pour tous les tests standardisés à référence normative (Brown 2010 : 33).

Deux menaces principales pèsent sur la validité des construits d'un test. La première est la non prise en compte de dimensions importantes des construits. Ce manquement est désigné par l'expression « construct under-representation » (Messick 1996 : 244). L'authenticité d'un test souffrant de ce défaut peut être mise en péril. La deuxième menace pesant sur la validité des construits est une évaluation trop large, par inclusion de facettes non-pertinentes (Messick 1996 : 244). Ce risque est connu sous le nom de « construct-irrelevant variance » (Messick 1996 : 244). Au-delà de la menace que ce manque fait peser sur la validité, il met en péril le caractère direct de l'évaluation.[13] Il faut noter que les deux menaces sont présentes dans tout type d'évaluation. Cela explique la nécessité, lors de la validation d'un test, de collecter suffisamment de données et de preuves pour écarter ces risques (Messick 1996 : 244)

1.4.4.4 Validité apparente

La validité dite apparente (*facial validity*) a trait à la manière dont les candidats perçoivent un test comme équitable, pertinent ou utile pour leurs études futures (Gronlund 1998 : 210). L'estimation des qualités d'un test par les candidats est éminemment subjective puisqu'elle se fonde essentiellement sur des affects et des impressions personnelles. C'est l'apparence superficielle du test qui sert de fondement aux évaluations des candidats: « Face validity refers to the degree to

which a test looks right and appears to measure the knowledge and abilities it claims to measure, […].” (Mousavi 2009 : 247). La dimension subjective et superficielle de la validité apparente explique la critique dont elle a été l’objet (Bachman 1990 : 285). Malgré les très légitimes réserves qui entourent ce type de validité, force est d’admettre que l’apparence d’un test a un effet non-négligeable sur la perception qu’en développent les candidats. Il ne s’agit pas seulement de sentiments personnels sans conséquence mais d’impressions qui ont un impact réel sur la performance (Brown 2010 : 35).

1.4.4.5 L’aspect consécutif de la validité de construit

L’aspect consécutif (*consequential aspect*) de la validité de construit (*construct validity*), un terme proposé par Samuel Messick dans son article de 1996, englobe toutes les conséquences sociales calculées ou fortuites de l’usage et de l’interprétation d’un test (Messick 1996 : 251). Il s’agit d’évaluer la précision avec laquelle les critères spécifiés sont mesurés, de déceler la présence éventuelle de biais dans l’attribution et l’interprétation des scores, de se prononcer sur le degré d’équité dans l’usage du test, de déterminer son impact positif ou négatif sur les pratiques d’enseignement et d’apprentissage. L’effet d’un test sur la préparation des candidats constitue un exemple évident d’impact sur les pratiques pédagogiques. Messick attire l’attention sur le fait que cette forme de validité ne doit pas être considérée de façon isolée (Messick 1996 : 251). Etant donné que des valeurs sociales sont associées à la signification des scores, l’évaluation de l’impact social d’un test est un aspect faisant partie de la validité de construit (Messick 1996 : 251).[14]

Pour certains chercheurs, cette notion est de facto synonyme d’impact ou d’effet de retour, ce que traduit très bien le terme anglais *washback* utilisé en linguistique appliquée[15] (Bachman & Palmer 1996 : 39). Cette synonymie établie par un certain nombre de chercheurs entre aspect consécutif et impact n’est pas acceptée par tout le monde. Messick, par exemple, estime que la notion d’impact doit être clairement distinguée de celle de validité. Il justifie sa position par deux arguments. Premièrement, l’impact n’est qu’une forme de conséquence possible entrant dans l’estimation de la validité d’un test. Deuxièmement, les conséquences elles-mêmes ne sont qu’un aspect de la validité de construit dont un test doit faire preuve (Messick 1996 : 242). Fulcher & Davidson (2007)

également condamnent la confusion entre impact et aspect consécutif de la validité de construit. Selon eux, il existe des forces dans le système éducatif et dans la société qui peuvent prévenir l'apparition de l'impact souhaité ou affecter son caractère, malgré la validité d'un test (Fulcher & Davidson 2007 : 223). Ce point de vue repose sur l'argument que la validité est une propriété du test, liée à son usage, tandis que l'impact est un phénomène complexe, qui ne peut pas être relié directement à la validité du dispositif d'évaluation (Fulcher & Davidson 2007 : 223).[16] Cette position reflète celle d'Alderson & Wall (1993 : 116) qui furent les premiers à dénoncer le lien hâtivement établi entre le *washback* et la validité en avançant les mêmes arguments que ceux invoqués par Fulcher & Davidson (2007). L'impact d'un test est conçu comme un phénomène complexe qui, à ce titre, ne peut pas être directement lié à la validité : « washback is likely to be a complex phenomenon which cannot be related directly to a test's validity » (Alderson & Wall 1993 : 116).

Certains linguistes estiment qu'on peut établir la validité d'un test en se fondant sur l'influence que celui-ci a sur l'enseignement et l'apprentissage. Morrow (1986) écrit ainsi: « The first validity criterion that I would … put forward for [these examinations] would be a measure of how far the intended washback effect was actually being met in practice » (Morrow 1986 : 6). Mais ce n'est pas chose aisée et Morrow lui-même admet ne pas savoir comment mesurer cet effet de retour sur les pratiques d'enseignement : «I am not sure, at this stage, how it could be measured. » (Morrow 1986 : 6). Il propose donc une méthode d'observation directe : les concepteurs du test à valider viennent eux-mêmes en salle de classe pour en estimer l'impact (Morrow 1986 : 6). Un test qui n'aurait pas l'effet de retour escompté sur l'enseignement et l'apprentissage manquerait alors de validité (Morrow 1986 : 6). Morrow emploie à ce sujet l'expression « washback validity » (Morrow 1986 :5). Frederiksen & Collins (1989) introduisent un concept similaire à celui de « washback validity », dénommé « systemic validity » (Frederiksen & Collins 1989 : 27). Ce dernier est étroitement lié au concept de *washback* car un test ferait preuve de validité en induisant des changements dans les programmes et l'enseignement. Des capacités cognitives mesurables seraient développées sous son influence (Frederiksen & Collins 1989 : 27).

1.4.5 Impact

L'impact de l'administration des tests et de l'usage des scores est observable à deux niveaux : au niveau macro et au niveau micro. Le niveau macro traite des effets du test sur la société et sur les systèmes éducatifs, tandis que le niveau micro envisage l'effet de la passation et de l'évaluation sur les individus (Brown 2010 : 34). L'impact des tests sur la société (niveau macro) a souvent été critiqué en raison de l'abondance des cours préparatoires qui se créent autour de tel ou tel test. Ces cours sont souvent payants et les détracteurs de ces initiatives commerciales mettent en garde les institutions contre le risque d'inégalités engendrées par des écarts de condition socio-économique entre candidats (McNamara 2000 : 54). L'impact des tests sur les individus (niveau micro), peut être constaté sur le plan de la motivation, de l'attitude antérieure et postérieure des candidats, ainsi qu'au niveau des styles d'apprentissage qui en résultent (Gronlund & Waugh 2008).

Le concept d'impact a été interrogé pour la première fois de façon critique par Alderson & Wall (1993). Auparavant, il était entendu que les tests avaient un impact majoritairement négatif sur l'enseignement (Alderson & Wall 1993 : 115). Pour Vernon (1956), par exemple, les tests altèrent les programmes (« distort the curriculum ») (Vernon 1956 : 166). Sous la pression du test, les professeurs en viennent à négliger des activités et des domaines autres, qui ne contribuent pas directement à la réussite aux épreuves. Vernon déplore également l'entraînement excessif aux tests (Vernon 1956 : 166).

Alderson & Wall (1993) proposent un changement de perspective. L'influence exercée par l'évaluation sur les activités d'enseignement et d'apprentissage cesse d'être jugée systématiquement de façon négative. L'impact est défini de façon beaucoup plus neutre comme étant l'influence de l'évaluation sur l'enseignement et l'apprentissage (Bailey 1996 : 259). L'influence sur les pratiques pédagogiques se manifeste par une focalisation des professeurs et des candidats sur les enjeux de l'évaluation. L'effet de retour du test (*washback*) peut être de promouvoir ou d'inhiber l'apprentissage: « Washback refers to the extent to which the introduction and use of a test influences language teachers and learners to do things that they would not

otherwise do that promote or inhibit learning » (Messick 1996 : 241). Selon Fulcher & Davidson (2007), ce concept est limité aux activités d'enseignement et d'apprentissage, sans tenir suffisamment compte des supports pédagogiques, qui sont eux aussi affectés par les tests (Fulcher & Davidson 2007 : 225). Pour rendre le concept d'impact significatif, il convient d'identifier les changements dans l'enseignement et l'apprentissage qu'on peut attribuer à l'emploi d'un type défini de test. Lorsqu'un impact négatif est décelé, il faut en chercher les causes, pour s'assurer que cet impact ne trouve pas sa source dans une défaillance structurelle interne du dispositif:

> The primary measurement concern with respect to adverse consequences is that negative washback or, indeed, any negative impact on individuals or groups should not derive from any source of test invalidity such as construct under-representation or construct-irrelevant variance (Messick 1996:252).

La nécessité d'identifier les causes de l'impact négatif d'un test tient au fait que la responsabilité du concepteur est engagée. Certains chercheurs comme Messick (1996 :252), limitent toutefois cette responsabilité. Celle-ci serait mise en cause uniquement si l'effet négatif du test résulte d'un défaut de conception. D'autres estiment au contraire que la responsabilité du concepteur est toujours engagée, que chaque développeur a l'obligation de saisir l'impact de ce qu'il conçoit et diffuse sur l'environnement d'apprentissage et sur la société (Alderson 2004 : xi).

Force est de reconnaître que le lien entre un test et son impact, qu'il soit positif ou négatif, est complexe. Ainsi, la qualité de l'impact d'un test sur l'enseignement et l'apprentissage ne dépend pas exclusivement de la qualité de l'instrument d'évaluation lui-même. Les recherches empiriques menées depuis 1993 tendent à montrer que les tests de bonne et de mauvaise qualité peuvent avoir un impact positif en augmentant la motivation et l'activité d'apprendre des candidats (Fulcher & Davidson 2007 : 225, 229). Ces mêmes recherches ont établi que la préparation à des tests considérés comme de mauvaise qualité n'entraînait pas ipso facto une mauvaise qualité d'enseignement. L'une des raisons de cette complexité de relation entre le test et son impact sur l'enseignement est que l'introduction d'un test ne suffit pas à bouleverser l'enseignement et l'apprentissage, malgré l'importance des dispositifs d'évaluation dans la plupart des sociétés (Fulcher & Davidson 2007 : 226). La conception déterministe, répandue dans le domaine de l'évaluation langagière,

est qualifiée de naïve par Fulcher & Davidson qui lui reprochent de ne pas tenir compte d'autres facteurs également susceptibles d'influencer l'enseignement, notamment la compétence des enseignants, leur compréhension des principes déterminant le test et la disponibilité des ressources à l'école (Fulcher & Davidson 2007 : 226). Le fait que l'enseignement et l'apprentissage soient affectés par de nombreuses variables, souligne la nécessité d'étudier ce phénomène dans chaque contexte où on postule l'impact du test sur le processus d'enseignement et d'apprentissage (Fulcher & Davidson 2007 : 229).

1.4.6 Authenticité

Le critère principal qui fonde l'authenticité est la correspondance qu'on peut établir entre les tâches incluses dans le test et les tâches langagières effectuées au dehors, en contexte réel, lorsqu'on communique dans la situation cible. La définition généralement acceptée de ce principe est celle fournie par Bachman & Palmer (1996): « the degree of correspondence of the characteristics of a given language test task to the features of a target language task » (Bachman & Palmer 1996: 23). Cette définition implique la nécessité d'inclure des tâches dans un test qui ressemblent au maximum à celles à effectuer lors de l'usage réel de la langue. Pour assurer l'authenticité d'un test, non seulement les tâches doivent simuler celles rencontrées dans des contextes réalistes, mais également tenir compte du temps et des ressources dont disposent les candidats (Messick 1996 : 243).

La nécessité de développer des tests authentiques peut être expliquée par deux raisons. Premièrement, les tests authentiques permettent de relier l'usage contraint et l'usage réel de la langue. Ainsi devient-il possible de généraliser l'interprétation locale des scores obtenus (Bachman & Palmer 1996 : 23). Les tâches authentiques constituent en effet le moyen d'établir dans quelle mesure l'interprétation d'un résultat est extensible au-delà de la performance ponctuelle et encadrée de l'épreuve. En second lieu, l'authenticité du test affecte la perception qu'en ont les candidats et influence leur performance. L'une des façons qu'ont les candidats de réagir à un test est de percevoir la pertinence de ce qui leur est proposé par rapport à l'usage réel de la langue, tant du point de

vue des contenus que des tâches requises (Bachman & Palmer 1996 : 24). La perception de l'authenticité d'un test encourage une réaction affective positive de la part des candidats et les aide à atteindre les meilleurs scores dont ils sont capables (Bachman & Palmer 1996 : 24).

Les tests authentiques se distinguent par le fait qu'ils demandent aux candidats de comprendre non seulement une information linguistique explicite mais aussi un sens transmis de manière implicite (Bachman 1990 : 4). Pour concevoir des tests authentiques, il faut auparavant déterminer les besoins langagiers en contexte réel puis adapter les tests à ceux-ci le plus précisément possible.

Selon certains chercheurs, il y a un lien direct entre la validité et l'authenticité d'un test (Noël-Jothy & Sampsonis 2006 : 44). Dans cette perspective, un bon test évalue nécessairement les capacités langagières des apprenants en situation réelle. Dans le cas contraire, le critère de validité n'est pas respecté (Noël-Jothy & Sampsonis 2006 : 44). Ce point de vue est partagé par les partisans de l'approche communicative appliquée à l'évaluation en langues, mais il n'est pas accepté à l'unanimité par les chercheurs dans le domaine de l'évaluation (Fulcher & Davidson 2007 : 63). Les critiques de ce point de vue nient le lien entre authenticité et validité d'un test, affirmant que l'adaptation des méthodes du test aux tâches à accomplir en situation réelle ne rend pas ce test automatiquement probant (Fulcher & Davidson 2007 : 63). Un tel dispositif d'évaluation permet de modéliser le comportement des candidats de sorte qu'il soit possible d'observer l'usage des processus qui seraient employés par les candidats en situation d'usage réel de la langue. Ces tests sont par conséquent directs, mais cette caractéristique ne débouche pas automatiquement sur une validité (Fulcher & Davidson 2007 : 63).[17]

Certains chercheurs se sont interrogés sur le lien entre la validité et l'authenticité en recourant au concept de construit. En effet, les deux qualités ont en commun l'exigence de ne pas négliger ou sous-estimer des facettes importantes du construit d'un test. Pareille exigence est désignée par l'expression « minimal construct under-representation » (Messick 1996 : 243). Toutefois, cette exigence commune à la validité et à l'authenticité, n'implique pas

nécessairement qu'un test doit être authentique pour acquérir une plus grande validité (Messick 1996 : 243).

De même qu'il n'existe pas de lien direct entre authenticité et validité, un dispositif authentique ne s'accompagne pas systématiquement d'effets positifs. Cela est dû à une multiplicité de raisons, liées aux qualités d'évaluation et aux caractéristiques du système éducatif, notamment au terrain sur lequel l'instruction et l'évaluation se déroulent (Messick 1996 : 244). En ce qui concerne les qualités d'évaluation, l'authenticité idéale ne se trouve que très rarement dans les tests, en raison de la difficulté à évaluer toutes les dimensions importantes de ses construits. Un dispositif d'évaluation ne reflète donc jamais de manière fidèle les performances d'usage réel de la langue (Messick 1996 : 244). Ceci est inévitable pour au moins deux raisons. Premièrement, il est dans la nature même d'un test de créer une anxiété. Cette anxiété et les processus déclenchés pour y faire face n'opèrent pas de la même manière dans les situations d'usage réel de la langue. Deuxièmement, la performance au test est évaluée et interprétée d'une manière distincte de celle qui prévaut dans l'usage réel de la langue (Messick 1996 : 244). Il y a donc des limites évidentes au naturalisme dont pourrait se prévaloir l'authenticité d'un test. La raison à cela est que dans l'évaluation langagière, comme dans toute évaluation éducative ou psychologique, ce sont les processus appliqués à l'attribution des scores et à l'interprétation des résultats qui priment, non les habiletés opérationnelles dans la performance de la tâche. L'impact d'un test fait la part belle à l'obtention de scores élevés, aux dépens des habilités. Il faut donc tenter de diminuer la part des habilités non prises en compte lors de l'attribution et de l'interprétation des scores (Messick 1996 : 245).

En dernier lieu, il convient d'interroger la relation entre technologie et authenticité. On peut estimer a priori que l'usage de la technologie augmente l'authenticité d'un test en permettant un enrichissement d'input multimodal. L'input non verbal est censé avoir des conséquences bénéfiques sur la qualité d'authenticité d'un test. Mais on manque encore d'études empiriques démontrant la réalité des effets bénéfiques de la technologie sur l'authenticité d'un test (Douglas & Hegelheimer 2008 : 118).

1.4.7 Interactivité

L'interactivité prend en compte les caractéristiques individuelles des candidats qui sont requises pour accomplir les tâches du test. Dans le contexte de l'évaluation langagière, les caractéristiques individuelles pertinentes pour les tâches sont, bien sûr, la compétence langagière (qui se compose des connaissances de la langue et de la compétence stratégique) mais aussi les connaissances thématiques et les schémas affectifs (Bachman & Palmer 1996 : 25). A la différence de l'authenticité, l'interactivité concerne l'interaction entre le candidat et les tâches du test, comme il existe une interaction entre l'usager de la langue et des tâches à accomplir dans les situations cibles. L'interaction est présente dans tout test, mais à des degrés variables. Il faut noter toutefois que le haut degré d'interactivité d'un test ne signifie pas forcément que ce dernier permette une mesure valable des compétences langagières. Une interactivité élevée peut être due à la mise en œuvre de stratégies métacognitives et à la convocation de connaissances thématiques par les candidats, sans que soient pour autant impliquées un grand nombre de compétences langagières (Bachman & Palmer 1996 : 25).

1.5 Questionnements actuels

Avant de décrire et d'analyser un échantillon représentatif de tests de langues, il est utile d'évoquer les interrogations parcourant la recherche actuelle. Nous commencerons par les réflexions sur l'équité des tests. Nous abordons ensuite le choix entre évaluation traditionnelle et alternative. Nous terminons par l'intégration des technologies numériques à l'évaluation des compétences en langues vivantes étrangères.

1.5.1 Les réflexions sur l'équité et l'éthique d'un test

En dehors des six qualités précédemment décrites, un test doit faire preuve d'équité. Le concept d'équité n'est pas nouveau. Il est traditionnellement lié aux qualités techniques des tests, à leur validité et à leur fiabilité (Kunnan 2000 : 1). Kunnan souligne que la primauté du concept d'équité et sa place dans le cadre de la justice sociale n'ont pas été suffisamment reconnues et débattues dans le

passé (Kunnan 2000 : 1). Le lien entre équité et justice sociale n'a que récemment attiré l'attention des experts (Kunnan 2000 : 1). L'équité est en fait un concept complexe qui peut être ramené à trois composantes indépendantes: la validité, l'accès et la justice. L'équité apparaît ainsi comme une notion interdisciplinaire car fondée sur des considérations sociales, éthiques, légales et philosophiques (Kunnan 2000 : 5).

Le premier enjeu de l'équité est la validité, notamment la validité de construit du test. Les concepteurs de tests doivent ainsi s'assurer que les interprétations des scores ont la même validité de construit, quel que soit le groupe de candidats. Ces derniers varient notamment en fonction de leur origine, de leur sexe, de leur domaine de spécialisation ainsi que de leur langue et de leur culture d'origine (Kunnan 2000 : 3). Pour que la validité de construit soit égale pour tous, il faut que les scores mesurés reflètent uniquement les compétences visées par le construit du test, c'est-à-dire, que les scores obtenus ne soient pas affectés par des facteurs non pertinents, résultant des caractéristiques particulières des groupes évoqués (Kunnan 2000 : 3).

La deuxième composante constitutive de l'équité est l'accessibilité du test. L'accessibilité concerne aussi bien les domaines financier, géographique, éducatif que personnel. L'accessibilité financière est un enjeu clé. Le deuxième enjeu essentiel est l'accessibilité géographique du dispositif, qui est variable selon les régions (Kunnan 2000 : 4). L'accessibilité personnelle englobe les aménagements spéciaux permettant aux personnes handicapées de passer le test. Pour des raisons d'équité, il faut créer les conditions qui rendent ces personnes capables de passer le test sans compromettre le construit mesuré (Kunnan 2000 : 4). Enfin, l'équité d'un dispositif d'évaluation est influencée par l'accessibilité éducative. Ce type d'accès implique l'opportunité fournie aux candidats d'apprendre en amont de la passation du test. La possibilité d'étudier les contenus et la méthodologie du test, de se familiariser avec l'équipement utilisé et les conditions de passation d'un test, a une grande influence sur le succès, tant aux niveaux individuel que collectif. Le succès collectif renvoie aux situations où des candidats ayant suivi un entraînement en amont de la passation obtiennent de meilleurs résultats que des groupes qui en sont privés (Kunnan 2000 : 4).

Le troisième enjeu associé à l'équité est la justice : justice sociale mais aussi litiges juridiques. La notion de justice sociale englobe les conséquences sociales d'un test, qui peut aussi bien contribuer à plus de justice qu'à moins de justice sociale, selon qu'il est ou non accompagné d'effets pervers (Kunnan 2000 : 5). L'inégalité d'accès de différents groupes de candidats aux opportunités de formation ou d'emploi, selon qu'ils ont ou non réussi le test, peut être un signe que le test contribue à une forme d'injustice sociale. Dans ce cas, il est nécessaire de créer un mécanisme de contrôle pour démontrer que l'injustice sociale n'est pas provoquée par le dispositif d'évaluation lui-même (Kunnan 2000 : 5). Les litiges sont la conséquence du manque de normes d'évaluation clairement énoncées (Kunnan 2000 : 5). Aux Etats-Unis et en Grande-Bretagne, notamment, la validité d'un test, utilisé dans le contexte scolaire ou professionnel, ainsi que l'usage des scores, représentent souvent des sources de litiges menant à des actions en justice (Kunnan 2000 : 5).

De nos jours il ne fait aucun doute que l'équité soit décisive pour assurer un niveau élevé dans la construction et l'usage des tests, ainsi que dans l'évaluation de leurs résultats. Les efforts doivent se concentrer non seulement sur la satisfaction des différents critères de validité, mais aussi sur la distribution de l'autorité, ainsi que sur la collaboration entre les différents acteurs impliqués (Shohamy 2001 : 160-161). Non seulement les concepteurs mais les utilisateurs doivent contribuer au développement de dispositifs équitables et libres de biais pour tous (Kunnan 2000 : 2). Les candidats peuvent aussi choisir de ne passer que les tests qu'ils estiment équitables envers chacun, quelle que soit l'origine, le sexe ou le handicap (Kunnan 2000 : 2).

L'usage de pratiques d'évaluation qui sont éthiques et démocratiques est indispensable pour fournir à toute personne des opportunités d'accès à l'éducation et à l'emploi (Shohamy 2001 : 160-162). Pour assurer des pratiques d'évaluation éthiques, le concept de professionnalisme a été développé par Davies (1997b). Ce dernier repose sur la notion de « moralité professionnelle » : « What a professional morality does […] is to provide a contract for the profession and the individual with the public, thereby safeguarding all three » (Davies 1997b : 333). Pour remplir sa fonction protectrice, la moralité professionnelle doit conclure un contrat entre la profession, l'individu et le public. Ce contrat se

manifeste normalement sous forme d'un Code d'Ethique et d'un Code de Pratique, définis par une association professionnelle. Ces codes doivent être l'objet d'une adhésion par tous les membres. Il s'agit d'un acte symbolique qui marque l'appartenance à une profession particulière (Davies 1997b : 333).

L'une des associations dans le domaine de l'évaluation langagière s'appelle *l'Association Internationale pour l'Evaluation en Langues*, abrégée par l'acronyme ILTA [17]. Cette association a été créée pour officialiser et corroborer les décisions des chercheurs, afin de guider le comportement professionnel des personnes chargées des pratiques d'évaluation. L'association a travaillé sur l'élaboration du *Code d'Ethique* et du *Code de Pratique* depuis les années 1990. Le *Code d'Ethique*, adopté en 2000, a pour but principal de promouvoir le comportement professionnel de tous les acteurs chargés de la construction et de l'administration des tests ainsi que de l'évaluation des résultats. Pour assurer un comportement professionnel de penser et d'agir, ce code est fondé sur un ensemble des normes éthiques, qui incluent l'équité, mais aussi la bienveillance, le respect de l'autonomie et de la société civile [18]. La bienveillance se manifeste par l'inclusion de tâches permettant aux candidats de donner le meilleur d'eux-mêmes dans la démonstration de leurs compétences langagières (Douglas 2010 : 63). Ce principe est lié à la procédure « biased for best » qui implique un degré de participation stratégique de l'enseignant et de l'étudiant à la préparation et à la mise en œuvre du test, ainsi qu'au suivi en aval de la passation (Brown 2010 : 44). On note par ailleurs, que la responsabilité de l'enseignant est autant engagée que celle de l'étudiant dans le processus de réussite. Il appartient donc à l'enseignant de créer les conditions optimales pour la meilleure performance possible de l'étudiant (Brown 2010 : 44). Ce code identifie neuf principes fondamentaux. Chacun est complété par des annotations clarifiant le principe énoncé. Les principes prescrivent les comportements attendus des membres d'ILTA ainsi que les objectifs qu'ils doivent se fixer. Enfin, les difficultés et les exceptions accompagnant la mise en pratique des principes sont évoquées.

Après l'entrée en vigueur du Code d'Ethique, des principes pour la bonne pratique de l'évaluation ont été formulés et adoptés en 2007 sous forme d'un *Code de Pratique*.[19] Ce référentiel, conçu par l'association, révèle également que l'équité est cruciale pour la pratique d'une évaluation de bonne qualité. Plusieurs

directives se focalisent sur ce concept, en revendiquant l'équité lors de l'administration du test ainsi que lors de la prise de décisions : « Persons who utilize test results for decision making must use results from a test that is sufficiently reliable and valid to allow fair decisions to be made ».[20] Pour permettre une prise de décision équitable, les responsables de l'évaluation doivent s'assurer que les conclusions tirées des résultats au test sont valables et fiables.[21] Le principe cité montre que les notions de validité et d'équité sont intimement liées dans le modèle de l'évaluation langagière défendu par ILTA [22]. De ce fait, ce modèle se rapproche de la position des chercheurs dans la discipline, sans toutefois l'épouser intégralement, étant donné que dans le modèle théorique présenté plus haut, le concept d'équité englobe celui de validité (Kunnan 2000 : 3). Dans le cadre théorique publié en 2010, désigné « Test fairness framework », Kunnan conçoit l'équité comme un concept complexe contenant cinq qualités principales, en l'occurrence, la validité, l'absence de biais, l'accès, l'administration du test et les conséquences qui en résultent (Kunnan 2010 : 3).

L'absence de biais est également un enjeu actuel qui dépasse le domaine de la recherche, puisque le traitement égalitaire des candidats est revendiqué par *le Code de Pratique :* « Care must be taken to ensure that all test takers are treated in the same way in the administration of the test ».[23] Or il peut y avoir un biais dans un test, malgré le traitement égalitaire des candidats. Si cela se produit, les différences de scores entre les candidats ne sont pas liées aux différents niveaux des capacités évaluées, mais à des facteurs indépendants des habiletés testées (Kunnan 2010 : 5).

Il existe plusieurs types de biais, dont le premier est un biais de contenu ou de langage utilisé vis-à-vis de candidats appartenant à des groupes différents. Ce type de biais peut avoir de nombreuses raisons : des connaissances thématiques ou culturelles spécifiques, l'utilisation d'une terminologie technique ou de variations dialectales (Kunnan 2000 : 3). Le deuxième type de biais, qui est lié au format, peut provenir de l'usage des questions à choix multiple, mais également des items à réponse construite (Kunnan 2000 :3). Les tests informatisés peuvent aussi être à l'origine de ce type du biais (Kunnan 2000 :3). Ceci paraît logique car tous les groupes de candidats ne sont pas familiarisés

avec les technologies de l'information et de la communication dans la même mesure.

Le troisième type de biais est l'impact des différences ethniques, sociales, culturelles et linguistiques des candidats sur les scores obtenus au test (Kunnan 2010 : 5). Ces différences individuelles, qui sont des facteurs extérieurs au construit du test, peuvent influencer la performance des candidats. C'est le cas des matériaux ou supports blessants, montrant certains groupes de façon défavorable (Kunnan 2000 : 3). Les types de biais évoqués peuvent apparaître dans tous les tests, y compris anonymes. Le biais peut être dû à l'influence de facteurs personnels et contextuels sur les évaluateurs (Fulcher & Davidson 2007 : 27). Ces facteurs entrent alors dans la valeur attribuée aux scores, ce qui est un facteur de perturbation. Afin de garantir que les différences de scores résultent uniquement de la variation des capacités évaluées, visées par le construit du test, et ce faisant, afin de garantir la validité du construit, il est nécessaire de sonder la performance des différents groupes de candidats (Kunnan 2000 : 3). Pour des raisons d'équité, il est particulièrement important de contrôler ces variables dans les tests à grande échelle, afin d'en minimiser les effets négatifs autant que possible (Fulcher & Davidson 2007 : 27).

Le *Code de Pratique* insiste sur le fait que la prise de décision équitable est essentielle car l'équité joue non seulement un rôle au niveau de l'évaluation des compétences mais conditionne le succès du test. Ceci est conforme aux résultats de la recherche menée sur ce sujet. Les problèmes d'équité peuvent se manifester non seulement au niveau du test, mais aussi dans le contexte plus large de leur usage (Shohamy 2000 : 15). L'usage d'un test n'est pas sans conséquences sociales, comme nous l'avons vu plus haut (Kunnan 2000 : 4). Or, la justice repose sur l'équité des décisions prises à partir des scores mesurés (Bachman & Purpura 2008 : 456). Pour garantir l'équité des décisions, il faut tenir compte des usages spécifiques prévus pour un test ainsi que des conséquences de ces décisions pour les différents groupes de candidats (Bachman & Purpura 2008 : 456). Ce besoin s'explique par les conséquences négatives qui résultent souvent de l'usage d'un test qui ne correspond pas à l'emploi prévu pour ce dispositif d'évaluation (Bachman & Purpura 2008 : 456).

Concernant l'usage auquel un test est soumis, on note l'espoir d'un bénéfice futur (Bachman & Purpura 2008 : 461). Les bénéfices espérés peuvent être considérés dans la perspective des candidats, ou alors les intérêts politiques, sociaux et administratifs des établissements éducatifs peuvent prédominer. Dans le premier cas de figure, l'usage du test est destiné à maximiser les chances d'accès équitable des individus et des groupes aux opportunités fondées sur le mérite individuel (Bachman & Purpura 2008 : 461). Dans le deuxième cas de figure, l'usage du test permet de choisir le nombre de candidats pourvus des compétences souhaitées en fonction d'intentions politiques, sociales et éducatives particulières (Bachman & Purpura 2008 : 461). Ces deux cas de figure ne s'excluent pas réciproquement, car les intentions politiques, sociales et/ou éducatives peuvent coïncider avec des conséquences positives pour les individus (Bachman & Purpura 2008 : 461). Pour illustrer ce fait à l'aide d'un test de positionnement, on peut dire que la prise de décisions équitables entraîne le placement des candidats dans les groupes de niveaux les mieux adaptés à leurs compétences et à leurs besoins individuels. L'équité est bénéfique pour les candidats eux-mêmes ainsi que pour les établissements d'enseignement qui gèrent leur parcours. On voit qu'un positionnement équitable bénéficie à chacun et qu'il en résulte des conséquences sociales justes. L'exemple donné confirme le lien entre l'équité des décisions prises et la justice en termes des conséquences sociales, établi par la recherche (Bachman & Purpura 2008 : 456).

En raison de la grande importance accordée à l'heure actuelle aux enjeux d'équité lors des différentes étapes de développement, d'administration et d'interprétation d'un test, on assiste à une discussion permanente sur le rôle du test comme régulateur d'accès (« as gate-keeper or door-opener », Bachman & Purpura 2008 : 456).

1.5.2 Evaluation traditionnelle ou alternative ?

L'un des questionnements actuels concerne la meilleure méthode à adopter pour évaluer les compétences langagières des candidats. Faut-il privilégier une évaluation traditionnelle ou opter pour une évaluation alternative, plus propice à l'intégration des compétences communicatives ? L'évaluation alternative est

jugée plus authentique, plus apte à susciter de véritables actes de communication, durant le processus même de mesure. En effet, l'évaluation alternative valorise l'usage réel de la langue, tel qu'il existe en dehors des situations de test ou d'examen. Il s'agit d'une exigence d'authenticité formulée par les experts (Bachman & Palmer 1996 : 9). Sont notamment privilégiées les tâches communicatives contextualisées qui appellent des réponses ouvertes et créatives. Grâce à leur format, ces tâches sont censées fournir l'occasion d'une performance individuelle. L'évaluation en est forcément plus subjective et le processus de retour plus interactif (Brown 2010 : 18). Par contraste, l'évaluation traditionnelle utilise des formats standardisés et cadrés temporellement, comportant des tâches décontextualisées, souvent au format QCM. Les réponses attendues sont identiques, discrètes et également décontextualisées (Brown 2010: 18).

La préférence actuelle pour l'évaluation alternative ne signifie pas pour autant que tous les concepts constitutifs de ce type d'évaluation soient favorables à l'élaboration des tests de langues (Brown 2010 : 18). Les deux approches ont en réalité des fonctions différentes, pour lesquelles elles doivent être valorisées (Brown & Hudson 1998 : 672). Au lieu de considérer certains tests comme des exemples d'évaluation alternative, donc des cas particuliers, Brown & Hudson (1998) proposent de voir tous les dispositifs comme autant de possibilités offertes, quelle que soit la tradition d'évaluation qu'ils représentent (1998 :672). Ces possibilités doivent être considérées comme des outils avec leurs avantages et inconvénients respectifs pour atteindre des objectifs d'évaluation particuliers (Brown & Hudson 1998 : 672). Cette position est corroborée par le fait que beaucoup de tests combinent en réalité les deux types d'évaluation. Cependant, de nombreux pédagogues et réformateurs de l'éducation prônent une diminution de l'usage des tests standardisés pour privilégier des dispositifs d'évaluation contextualisés, communicatifs et fondés sur la performance (Brown 2010 : 21). Pour cette raison, il faut réfléchir lors de la conception d'un test, à l'intégration d'éléments d'évaluation alternative (Brown 2010 : 18).

1.5.3 L'usage des technologies numériques dans l'évaluation des compétences langagières

L'usage des nouvelles technologies connaît une forte croissance dans le domaine éducatif. L'évaluation des compétences langagières ne constitue pas une exception à cette évolution (Brown 2010 : 19). L'introduction progressive des technologies numériques dans le domaine de l'administration et de l'évaluation des tests de langue est non seulement engagée mais inévitable (Douglas 2010 : 116). Cette mise en pratique progressive des technologies n'est pas seulement due à la préférence actuelle accordée aux nouvelles technologies, mais elle est aussi liée aux résultats des recherches menées dans ce domaine, par exemple, par Cenoz & Gorter (2011). Les recherches montrent que l'usage combiné des différents médias lors de l'apprentissage des langues est favorable au développement du multilinguisme (Cenoz & Gorter 2011 : 340). La raison à cela est que les technologies multimédia permettent le développement d'une « lecture multicanalaire » (*multimodal literacy*), fondée sur l'usage combiné des différents codes de communication, qui sont sonore, visuel et gestuel (Cenoz & Gorter 2011 : 340). Quant au canal visuel, il comporte non seulement l'écriture, mais également d'autres stimuli visuels, en l'occurrence, les images, les diagrammes et les vidéos (Cenoz & Gorter 2011 : 340). Ces codes de communication incluent l'usage de la langue, mais n'y sont pas limités, car l'emploi unique de la langue ne permet pas une communication complète (Cenoz & Gorter 2011 : 340).

Le développement de la « lecture multicanalaire » (*multimodal literacy*), favorisée par les technologies multimédia, offre des avantages considérables pour le développement de la conscience linguistique et, finalement, du multilinguisme (Cenoz & Gorter 2011 :340). La raison à cela est l'affinité entre les notions de multilinguisme et de « lecture multicanalaire», au-delà du fait que l'usage des multiples modes de communication favorise lui-même le plurilinguisme. La multimodalité de la communication s'accompagne d'un brouillage des frontières auparavant nettes entre les modes de communication écrit et oral (Cenoz & Gorter 2011 : 340). Ce brouillage, ou, au moins, ce gommage des frontières entre les différentes langues est nécessaire pour que le multilinguisme puisse se manifester dans la communication (Cenoz & Gorter

2011 : 340). Le multilinguisme se manifeste dans la communication par l'usage juxtaposé des différentes langues maitrisées par un individu, désigné par les termes anglais *codeswitching* et *codemixing* en linguistique (Cenoz & Gorter 2011 : 340).

En raison de l'usage de plus en plus répandu des dispositifs numériques d'évaluation des compétences, il faut s'interroger sur l'impact des différents aspects liés à l'évaluation (semi)automatique des compétences langagières. Ainsi, faut-il penser aux effets de l'utilisation des technologies sur les attitudes et les émotions des candidats, au lien entre les types de technologies utilisés et la performance langagière, ainsi qu'à leur impact sur le contenu et le format des tâches concevables (Douglas 2010 : 116). De plus, il est nécessaire de se rendre compte du fait que l'application des technologies affecte la définition des compétences qu'on entend mesurer (Douglas & Hegelhemer 2008 : 117). L'usage des technologies à des fins d'évaluation des compétences langagières implique une montée en complexité du construit du test (Douglas & Hegelhemer 2008 : 117). La montée résulte de la nécessité de définir les compétences langagières en fonction des supports utilisés, en l'occurrence, du support papier, électronique ou sonore (Douglas & Hegelheimer 2008: 117).

Certains tests assistés par ordinateur sont locaux, petits et accessibles sur un site web particulier, tandis que d'autres sont au contraire larges, standardisés et ouverts à un large spectre de candidats (Brown 2010 : 20). La variété dans la dimension ainsi que l'opportunité de passer un test pour s'entraîner sont des atouts majeurs. Un autre point fort consiste en une évaluation autodirigée, donc autonome, de plusieurs domaines langagiers, ce qui peut servir de pratique pour un test standardisé avec impact sur l'avenir d'un candidat (Brown 2010 : 20). L'usage éclairé des technologies est d'une grande aide aux développeurs de tests. L'intégration de dispositifs numériques doit permettre une démarche efficace pour construire et organiser des tests ainsi que pour attribuer des scores et évaluer des résultats (Douglas 2010 : 139). Il faut cependant se garder d'utiliser la technologie pour elle-même sans égard pour les compétences à mesurer. Ainsi est-il essentiel de choisir les tâches pour leur pertinence dans le processus d'évaluation, non pour leur facilité d'administration ou d'évaluation avec les technologies numériques (Douglas 2010 : 139).

Les technologies numériques ne sont pas sans inconvénient. Certains sont inévitables, comme le manque de sécurité et la possibilité de fraude, en cas d'évaluation autodirigée et non-supervisée. Les autres peuvent être maitrisés, comme le développement du format QCM et la réduction du format à réponse construite. Il faut savoir que les technologies existantes ont des limites dans l'évaluation des tâches construites et que l'attribution automatique de scores se distingue de celle effectuée manuellement par les êtres humains (Douglas 2010 :117). Un autre point faible des tests automatisés est la tendance à considérer les tâches comme de simples composantes d'un dispositif au lieu de considérer ces dernières comme révélatrices de l'usage réel de la langue en contexte (Douglas & Hegelheimer 2008 : 124). Ce dernier inconvénient peut être évité par l'usage créatif des technologies numériques dans l'administration des tests, qui permet d'augmenter l'authenticité d'un dispositif, l'échange interactif durant sa durée ainsi que de promouvoir l'autonomie des candidats. En effet, les technologies numériques peuvent augmenter la chance à la fois d'enseigner et d'évaluer les compétences langagières de façon plus communicative (Douglas & Hegelheimer 2008 : 115).

Indépendamment de la question de savoir si on doit ou non se servir des technologies numériques, il faut toujours commencer la conception d'un test de langues par l'analyse de la situation cible pour laquelle le test est conçu, puis déterminer les tâches appropriées. C'est seulement après ces deux étapes préalables qu'il convient à réfléchir à la manière dont la technologie peut nous aider à fournir une évaluation des compétences équitable, fiable, authentique et pratique (Douglas 2010 : 139). Il est important de garder à l'esprit les six qualités décrites ci-dessus, auxquelles chaque test des langues doit satisfaire.

1.6 Evaluation de tests de positionnement existants

Avant d'entamer la réalisation d'un test de positionnement automatisé en langues (POSILANG), adossé au CECRL, il était indispensable d'évaluer plusieurs dispositifs d'évaluation déjà disponibles en anglais. Cette phase d'analyse de l'existant est essentielle, car elle permet de se familiariser avec le répertoire des tâches ou exercices, ainsi que d'identifier les compétences évaluées. Il est

important de préciser que l'étude très détaillée que nous avons opérée n'a jamais eu pour finalité un emprunt ou une imitation de contenu et de forme. L'objectif fondamental est d'observer les formats, de mesurer les écarts et surtout de comprendre les principes généraux qui parcourent l'ensemble des tests de positionnement en anglais langue étrangère. Les tests que nous avons passés au crible sont les suivants : *Oxford Quick Placement Test, Oxford Placement Test 2, Cutting Edge Placement Test, Upstream Placement Test, Success Placement Test, Energy Placement Test, CAREL* ainsi que *DIALANG.*

Le test que nous préparons pour les universités de Bordeaux devant être un test de positionnement il est logique que nous nous soyons essentiellement concentrés sur ce type de test. Signalons néanmoins que *DIALANG* a un statut particulier que nous expliciterons dans le deuxième chapitre de cette thèse, consacré au *Cadre européen commun de référence pour les langues* (CECRL).

L'évaluation des tests de positionnement va se dérouler de la manière suivante. Dans un premier temps, les caractéristiques pertinentes du test seront décrites à l'aide des paramètres de la grille « maison » présentée plus haut. Dans un deuxième temps, la typologie des exercices inclus dans ce test sera exposée et analysée. Dans un troisième et dernier temps, les types de compétences évalués par ces exercices seront révélés.

1.6.1 *Oxford Quick Placement Test*

1.6.1.1 Description et évaluation des caractéristiques du test

L'Oxford *Quick Placement Test* a été édité par *Oxford University Press* et *University of Cambridge Local Examinations Syndicate* en 2000. C'est un test de positionnement d'accès payant. Il s'adresse à toute personne indépendamment de son niveau de langue. Conformément à sa typologie, le but de ce test est d'évaluer le niveau de compétence en langue des candidats et de les placer dans un cours de langue adapté. Une contrainte temporelle particulièrement serrée est imposée aux candidats qui doivent passer le test en 30 minutes seulement (d'où le qualificatif « quick »). Le test comporte deux parties, qui se différencient par le nombre d'items et par leur conception. La première partie contient 40 items, dont 15 sont intégrés dans un texte cohérent portant sur un sujet précis. En ce qui

concerne les cinq premiers items, chacun se réfère à une phrase isolée. Cependant chacune de ces phrases permet de s'imaginer un contexte communicatif propice à son énonciation. Les candidats sont d'ailleurs invités à se figurer une situation précise au travers de la question « Where can you see these notices ? » figurant dans les consignes. La deuxième partie du test comprend 20 items dont dix sont intégrés à un texte cohérent et dix autres alignent des phrases isolées.

Il y a trois compétences évaluées par ce test, la compréhension écrite ainsi que les deux compétences communicatives langagières, lexicale et grammaticale. Tandis que la compréhension des syntagmes et des phrases complètes fait partie de la compétence de réception à l'écrit, la compréhension des mots isolés relève de la compétence lexicale (CECRL 2001 : 57, 87-88). En ce qui concerne la compétence grammaticale, celle-ci est sollicitée en choisissant le lexème qui convient le mieux parmi les représentants de plusieurs catégories lexicales ou grammaticales. Par exemple, choisir si une préposition ou une particule adverbiale convient dans une phrase, ou encore déterminer quelle forme verbale convient parmi plusieurs citées.

Par rapport au contenu des tâches, les trois ou les quatre options présentées se réfèrent soit à une phrase isolée soit à un texte cohérent servant de base. La méthode utilisée pour évaluer ces trois compétences consiste à choisir la bonne réponse parmi les options données sous forme de QCM. [24] A l'exception des cinq premiers items, la réponse correcte permet de compléter une phrase. Il y a donc une combinaison de deux formats : exercice à trous et QCM.

En ce qui concerne le mode d'évaluation de ce test, le résultat est transmis de façon chiffrée par simple indication du nombre de points atteints. Sur la base de ce score, un niveau de compétence est défini sur l'échelle du CECRL. Le candidat peut non seulement découvrir son score et son niveau CECRL, mais aussi apprendre quel est le score limite d'un niveau de compétence. Cela lui permet de voir où sa compétence individuelle est placée à l'intérieur d'un niveau particulier. En raison de la corrélation entre le nombre de points obtenu et le niveau attribué, ainsi que l'affichage du *cut off* entre les 6 niveaux différents, la transparence des niveaux d'évaluation est préservée. Le bilan est communiqué

au candidat sous forme d'un tableau qui fournit non seulement des informations sur le score atteint et le niveau de référence CECRL de l'utilisateur, mais aussi sur le niveau scolaire auquel correspond normalement ce niveau de compétence. Sont également indiqués les tests de certification en anglais général et en anglais des affaires accessibles sur la base du score obtenu. Cependant, le bilan n'est pas aussi précis et détaillé qu'il ne le paraît, car il ne contient pas de conseils sur la façon d'améliorer ses performances et n'explicite ni les atouts ni les points faibles du candidat. L'évaluation de ce dispositif permet de repérer plusieurs points forts mais aussi certains défauts. En ce qui concerne les atouts, la possibilité de passer très rapidement les épreuves, la transparence du mode d'évaluation, la mesure des compétences lexicales et grammaticales en contexte phrastique et non de façon isolée.

La principale réserve inspirée par ce test est sa focalisation exclusive sur la compréhension écrite. La compréhension orale n'est pas l'objet d'une évaluation, pas plus que ne le sont la production écrite ou l'expression orale. Or l'évaluation des compétences lexicale et grammaticale à l'écrit ne saurait se substituer intégralement à l'évaluation des trois autres compétences. La substitution est d'autant plus problématique que la compétence grammaticale n'est pas en corrélation simple et directe avec un niveau de référence du CECRL. En effet, peu d'indications dans le Cadre permettent de relier de façon univoque la maîtrise d'aspects grammaticaux particuliers à des niveaux de compétence (Westhoff 2007 : 678). Cependant, cette difficulté n'est pas admise par tous les chercheurs. Glabionat et al. (2002), par exemple, considère qu'il est possible de relier la maîtrise de points grammaticaux précis aux niveaux du CECRL. La question reste donc ouverte.

1.6.1.2 Analyse de la typologie d'exercices et de compétences

Après avoir décrit les caractéristiques et évalué les points forts ou faibles de l'*Oxford Quick Placement Test*, il convient d'expliciter les types d'exercices et de compétences contenus. La première partie du dispositif se compose d'exercices de lexique et de grammaire. Les items 1-5 contiennent des exercices de lexique servant à évaluer la compétence de réception à l'écrit. Le sens de syntagmes et

de phrases complètes et isolées doit être saisi, comme « Please leave your room key at the reception ». [25]

La deuxième compétence évaluée par ces items est la compétence pragmatique. Celle-ci désigne l'utilisation fonctionnelle des ressources de la langue, c'est-à dire, leur usage dans le but de réaliser les fonctions langagières (CECRL 2001 : 18). Un contexte communicatif est créé par des énoncés que les candidats ont besoin de comprendre avant de prendre une option d'identification. En donnant la réponse correcte, le candidat démontre à la fois sa compréhension du sens de l'énoncé et sa compréhension du contexte d'usage, autrement dit il manifeste sa compétence pragmatique. Toutefois, la compétence pragmatique testée par ces items reste limitée. Bien que les phrases soient rattachées à un contexte communicatif, leur fonction pragmatique n'est pas véritablement explicitée. Il n'est nulle part question de la finalité de ces énoncés, de leur sens d'intention, des présupposés qu'ils véhiculent.

Les items 6-10 recouvrent des exercices de grammaire qui évaluent les aspects suivants de la compétence grammaticale: le bon usage de l'auxiliaire *be*, des pronoms possessifs, des pronoms indéfinis et des prépositions. Plus précisément, il s'agit de vérifier la bonne distribution des pronoms possessifs en fonction de la personne et du nombre, le choix correct des pronoms indéfinis, ainsi que l'emploi des différentes formes de l'auxiliaire *be*. Un autre aspect grammatical évalué est la capacité à choisir la préposition qui convient dans une phrase.

Les items 11 à 20 portent sur le lexique. La compétence lexicale est évaluée en invitant les candidats à choisir le lexème qui convient parmi plusieurs formes appartenant à la même catégorie lexicale ou grammaticale, à savoir, des noms, des verbes, des adverbes de temps, des conjonctions de subordination ainsi que des adjectifs et des numéraux.[26] Il y a un seul item, le numéro dix-huit, dans lequel les options données appartiennent non pas à la même catégorie, mais à deux catégories différentes, les adjectifs et les numéraux ordinaux. Concernant la compétence verbale, le choix du verbe correct est à faire, une fois, parmi les formes verbales au gérondif et une autre fois, parmi les formes *au simple past.* Pour choisir le lexème qui convient dans une phrase, la

compréhension du sens de la phrase complète contenant le blanc à remplir est nécessaire. Au-delà de celle-ci, le cotexte fourni par d'autres énoncés aide à choisir l'option correcte.

Les items 21-40 recouvrent majoritairement des exercices de grammaire, bien qu'on puisse dénombrer aussi quelques exercices lexicaux, en l'occurrence les items 27, 36, 37, 38 et 39. Ces-derniers se distinguent par le fait que le bon choix ne repose pas sur un jugement de grammaticalité, mais sur une saisie correcte du sens. Il s'agit de choisir le lexème qui permettra d'interpréter sans difficulté la phrase. La phrase 38, par exemple, reste grammaticalement correcte, quel que soit le lexème inséré. En revanche, seul le choix du verbe *tell* (option 4) est le bon pour que cette phrase fasse sens. En ce qui concerne les exercices de grammaire, les options présentées appartiennent à l'une des catégories lexicales ou grammaticales suivantes : verbes, noms, prépositions, particules adverbiales ou déterminants. [27]

En ce qui concerne les exercices évaluant la compétence verbale, la distribution entre les différentes formes est la suivante. La première concerne *un infinitif*, une forme finie du verbe, une forme au *simple past* ainsi qu'un verbe associé à un modal (item 21). L'item 22 présente les formes verbales dont l'une est à l'infinitif et trois sont au gérondif. La première n'est précédée de rien, la seconde est accompagnée par une préposition et la troisième par la particule infinitive *to.* Dans le troisième cas, le choix entre les formes négatives des verbes *need*, *have* et *get* doit être fait (item 30).

Les items 41-50 proposent des exercices lexicaux évaluant la compétence lexicale à l'écrit. Il s'agit de choisir le lexème qui convient dans une phrase. Pour y parvenir, la compréhension du sens complet de la phrase est nécessaire. Au-delà de cette compréhension, la prise en compte du cotexte fourni par d'autres énoncés aide à choisir l'option correcte. Concernant les sous-types de compétences engagées, on relève la compréhension du sens des lexèmes appartenant à l'une des trois catégories lexicales suivantes : les noms, les adjectifs ou les verbes.

Les items 51-60 recouvrent à la fois des exercices lexicaux et grammaticaux. Ainsi, dans certains items, le bon choix est fondé sur la

grammaticalité de la phrase, tandis que dans d'autres cas, le choix est lié au sémantisme de la phrase. La répartition en deux catégories d'exercices ne dépend pas de la catégorie dont font partie les options. Cependant uniquement deux items de ce dernier passage du test évaluent la compétence lexicale. Le reste sert donc à évaluer la compétence grammaticale.

1.6.2 *Oxford Placement Test 2*

1.6.2.1 Description et évaluation des caractéristiques du test

L'Oxford Placement Test 2 a été publié par l'édition *Dave Allan* en 1992. Il ressemble à plus d'un titre à *l'Oxford Quick Placement Test*. Comme ce dernier, il ne peut pas être passé en ligne, mais uniquement de façon traditionnelle, sur papier. Le deuxième trait commun est le caractère commercial de ces deux tests, à cela près que l'*Oxford Placement Test 2* coûte plus cher encore. Puisqu'il s'agit d'un test de positionnement, sa fonction essentielle consiste à placer les candidats dans un cours d'anglais qui soit conforme à leur niveau. Le public visé par ce test est large puisqu'il s'agit de toutes les personnes souhaitant évaluer leur niveau de compétence en langue anglaise. Cette ouverture contraint le test à couvrir tous les niveaux de compétence distingués par le CECRL.

Concernant la conception du test, il se compose de trois parties. Les deux premières portent sur l'évaluation de la compétence grammaticale tandis que la troisième est centrée sur la compréhension de l'oral. Les deux premières parties comportent 50 items chacune, tandis que la troisième en comporte une centaine. Un autre élément de différenciation entre les trois parties est l'intégration de certains items des deux premières parties à un texte cohérent, contrairement à ceux de la troisième partie.

Concernant la compétence grammaticale, il nous paraît utile d'exposer brièvement le modèle de grammaire de Purpura (2004) avant d'aborder l'analyse des deux premières parties du test. La connaissance grammaticale englobe deux composantes : la forme grammaticale et le sens grammatical (Purpura 2004 : 61). Les formes grammaticales désignent toutes les formes linguistiques à tous les niveaux de la langue, notamment aux niveaux phonologique, lexical, morphosyntaxique, cohésif et interactionnel. La connaissance de la forme

grammaticale se réfère à la connaissance d'une ou de plusieurs de ces formes linguistiques. Le sens grammatical implique le sens littéral des phonèmes, morphèmes et phrases, de sorte que le sens d'une phrase est simultanément dérivé de celui de ses parties et de leur combinaison [28] (Purpura 2004 : 61). Purpura indique qu'il existe plusieurs termes pour désigner le sens grammatical et l'appelle le *sens littéral*, conformément à l'ouvrage de Grice (1975).

Dans ce test, l'ensemble des formes grammaticales est évalué. En effet, les candidats sont invités à choisir le morphème ou le syntagme qui est grammaticalement correct dans une phrase donnée. Le but de ces exercices est d'évaluer la connaissance des structures langagières par les candidats, comme signalé dans l'introduction théorique (*Oxford Placement Test 2,* 1992). L'évaluation du sens constitue le seul moyen de choisir les structures langagières grammaticalement justes. Une seule réponse est grammaticalement correcte parmi les trois proposées.

La compréhension orale testée dans la troisième partie consiste à reconnaitre le lexème ou le morphème entendu en appariant forme sonore et forme écrite. Ces items servent à évaluer la capacité de discrimination phonétique des candidats. La juxtaposition de deux unités lexicales partageant un même son articulé favorise la reconnaissance nette des phonèmes. Cette aptitude fait partie de la compétence phonologique des candidats, l'une des compétences linguistiques distinguées dans le CECRL, qui est définie comme « une connaissance de la perception et de la production » et qui comporte plusieurs sous-compétences, dont les plus pertinentes sont « une aptitude à percevoir et à produire » :

> -les unités sonores de la langue (phonèmes) et leur réalisation dans des contextes particuliers (allophones)
>
> -les traits phonétiques qui distinguent les phonèmes (traits distinctifs tels que, par exemple, sonorité, nasalité, occlusion, labialité)
>
> -la composition phonétique des mots (structure syllabique, séquence des phonèmes, accentuation des mots, tons, assimilation, allongements) (Conseil de l'Europe 2005 : 91).

On relève que les items inclus dans ce test ne permettent d'évaluer qu'une compétence phonologique partielle : l'aptitude à percevoir les phonèmes et leur

réalisation dans des contextes particuliers, à lier des phonèmes à une représentation orthographique, mais en laissant totalement de côté la capacité de production.

Le contenu des items dépend de la compétence évaluée. Les items ciblant la compétence grammaticale proposent de choisir le morphème ou le syntagme qui conviennent du point de vue grammatical. Les items se focalisant sur la compréhension de l'oral demandent aux candidats d'écouter des phrases contenant une alternative sonore et ensuite, de les comparer avec les versions graphiques. Dans tous les cas, le format mis en œuvre est celui du questionnaire à choix multiple.

Le mode d'évaluation est d'une remarquable transparence, et cela pour plusieurs raisons. Premièrement, le score pour les deux sections du test est affiché séparément. De plus, le nombre de points atteint pour la compétence grammaticale est calculé à partir du score obtenu dans les deux sous-parties évaluant cette compétence. Deuxièmement, le score maximal de 100 points, à gagner dans chacune de deux parties, évaluant la compétence grammaticale et la compréhension de l'oral, correspond à 100 items testés. Il en résulte que le score total gagné par le candidat reflète directement le nombre d'items auxquels il a répondu correctement. Le troisième point qui contribue à la transparence de l'évaluation est la haute densité de l'échelle de niveaux distingués. Il y a 16 niveaux qu'on peut atteindre, dont neuf pleins et sept intermédiaires qui se trouvent entre les niveaux proprement dits. Il en découle un étalonnage très précis de la compétence langagière de l'utilisateur. Le dernier facteur augmentant la transparence est l'affichage du *cut off* entre les niveaux. L'indication du score limite a pour conséquence que la corrélation entre le nombre des points atteint et le niveau attribué est transparente.

Le résultat individuel est fourni au candidat sous forme de tableau. Celui-ci montre les correspondances entre les niveaux de compétence à atteindre dans ce test et les certificats décernés par quelques autres tests en anglais langue étrangère, en l'occurrence, *Cambridge Exams, ARELS/Oxford Exams* ainsi que *LCCI Exams*. Cependant, ces données fournies au candidat ne constituent pas un bilan au sens propre en raison de l'absence d'évaluation fine de sa

compétence individuelle. Une évaluation fine est censée exposer les points forts et les points faibles de sa compétence langagière, en incluant des conseils sur la possibilité d'améliorer l'apprentissage de la langue. Or, le tableau donné sert essentiellement à informer le candidat sur son niveau de compétence, en le situant sur l'échelle du CECRL.

Les deux aspects évoqués, l'évaluation transparente à un haut degré et l'étalonnage très précis de la compétence langagière, représentent les deux aspects positifs de l'*Oxford Placement Test 2.* Certaines faiblesses sont toutefois décelables. La première est l'évaluation des compétences de réception uniquement. La non-évaluation des deux compétences productives, de l'expression écrite et de l'expression orale, implique que la compétence communicative n'est testée que partiellement. Or la réalisation de tâches de communication authentiques nécessite a minima d'avoir recours aux quatre activités langagières : deux activités de production et deux activités de réception (Conseil de l'Europe 2005 : 48). En tout état de cause, un nombre important de situations demandent aux utilisateurs de la langue de s'impliquer dans des activités communicatives mixtes (Conseil de l'Europe 2005 : 48). L'absence de bilan détaillé, déclinant les forces et les faiblesses, orienté vers les apprentissages futurs, est un autre aspect susceptible d'être jugé négativement.

1.6.2.2 Analyse de la typologie d'exercices et de compétences

Les deux parties du test ciblant la compétence grammaticale contiennent deux types d'exercices de grammaire : ceux qui évaluent la connaissance des formes grammaticales, d'une part, et ceux qui évaluent la compréhension du sens grammatical, d'autre part. Le premier segment de la première partie du test est constitué de 25 items non intégrés à un texte qui portent sur la forme grammaticale uniquement. Ces items invitent les candidats à choisir un morphème appartenant à l'une des catégories lexicales ou grammaticales suivantes: verbes, pronoms, adjectifs, déterminants, conjonctions ainsi que les syntagmes nominaux et verbaux. La variété des catégories montre que plusieurs sous-types de compétences grammaticales sont évalués. Le premier sous-type concerne la compétence verbale, où on trouve une distribution entre les formes suivantes :

Simple present, present progressive,infinitive progressive.

Will future, present progressive, verbe associé à un modal.

Be going to au présent et au passé, modal *could*

Gérondif précédé d'une préposition, infinitif, infinitif précédé d'une préposition.

Forme finie du verbe, *simple past, past progressive.*

Simple past, present perfect, past perfect.

L'auxiliaire *be* à des temps différents.

Le verbe causatif *make* à des temps différents.

Present perfect passive, simple past passive, past progressive.

Concernant la distribution entre les différents pronoms, on trouve deux cas de figure. Le premier est le choix à opérer entre *It* et *There* associés à *be* ou *have.* Le deuxième concerne les quantifieurs marquant une faible quantité et variant selon le nombre, à savoir, *little, few et less.* En ce qui concerne les déterminants, les items faisant partie de cette catégorie grammaticale présentent les distributions suivantes :

L'article défini, les déterminants indéfinis

Le déterminant *most* associé i) à un article défini ii) à une préposition.

L'article défini, les déterminants démonstratifs et possessifs.

Such a, l'article indéfini associé à l'adverbe de degré *so*.

Par rapport aux adjectifs, deux cas de figure se trouvent dans le segment du test concerné. Le premier présente des formes comparatives différentes tandis que le deuxième contient plusieurs formes superlatives.

Contrairement aux 25 premiers items, la partie suivante du test constitue un texte intégral. Une partie des 25 items évalue la connaissance des formes grammaticales elles-mêmes, tandis que le reste mesure la compréhension du sens grammatical. L'évaluation de la connaissance des structures grammaticales est effectuée par le choix de la juste forme appartenant à l'une des catégories lexicales ou grammaticales suivantes : les verbes, les déterminants, les pronoms, les prépositions et les adverbes. La compétence verbale des candidats est évaluée par des items présentant une multitude de distributions entre les formes, qui sont les suivantes :

Au passif : *present perfect, present progressive, simple past* ;

Past progressive, past perfect, verbe associé à un modal ;

Present progressive, past perfect, verbe précédé d'un modal ;

Infinitif, gérondifs précédés de prépositions différentes ;

Infinitif, gérondif, infinitif au parfait,

Forme finie, infinitif, gérondif,

L'auxiliare *be* à des temps et aspects différents.

Will , *would* ainsi que l'auxiliaire *did* associés à un infinitif.

La capacité à choisir le déterminant correct est testée par deux items présentant deux types de distribution. Le premier item met en option l'article défini, l'article indéfini et le déterminant possessif tandis que le deuxième présente les déterminants indéfinis variant selon le nombre et la quantité désignée. Les items centrés sur les pronoms présentent les types de distribution suivants: les pronoms relatifs, les pronoms relatifs et démonstratifs, les pronoms démonstratifs et les pronoms personnels variant selon le nombre et enfin les pronoms réfléchis et possessifs, précédés par des prépositions.

En ce qui concerne l'évaluation du sens grammatical, les items qui servent à cet usage sollicitent les candidats à choisir un morphème lexical ou grammatical faisant partie d'une des trois catégories. Il s'agit des verbes *make, have* et *let* intégrés dans les propositions causatives, des adverbes de degré et de temps au comparatif ainsi que des prépositions.

La deuxième partie d'*Oxford Placement Test 2* se compose de trois segments. Le premier comporte quinze items non-intégrés dont certains visent à évaluer la maitrise des formes grammaticales et d'autres la compréhension du sens grammatical. Comme dans tous les items de ce test qui portent sur l'évaluation des formes grammaticales, il s'agit de choisir la forme qui convient dans une phrase. Les formes appartiennent à l'une des catégories lexicales ou grammaticales suivantes: les verbes, les adjectifs, les pronoms, les déterminants, les prépositions, les conjonctions et les syntagmes. En ce qui concerne la compétence verbale, la distribution entre les formes suivantes est à noter :

Present perfect, present progressive, simple past à la voix passive;

Past progressive, past perfect, verbe associé à un modal ;

Present progressive, past perfect, verbe précédé d'un modal ;

Infinitif, gérondif précédé de prépositions différentes ;

Infinitif, gérondif, infinitif au parfait ;

Forme finie, infinitif, gérondif ;

Auxiliare *be* aux temps et aux aspects différents ;

Will, would ainsi que l'auxiliaire *did*, associés à un *infinitif* ;

Le deuxième type de distribution concerne les prépositions de manière, *as, like*, et le comparatif *than*. Le troisième cas de figure consiste à mettre le verbe *prefer* et les adverbes de degré en option. Cette distribution se distingue des précédentes en présentant des options qui relèvent de catégories différentes. Il en est de même pour un autre item qui propose un choix parmi trois syntagmes, deux nominaux et un adjectival. Finalement, la capacité à utiliser des déterminants indéfinis et des pronoms indéfinis est testée par deux items respectivement. Comme évoqué plus haut, la compréhension du sens grammatical est également évaluée par certains items de cette partie. Ainsi, les candidats doivent sélectionner un lexème qui convient dans la phrase sémantiquement. Les catégories qui apparaissent sont les noms et les adverbes composés, dont celui de lieu et celui de manière.

Le deuxième bloc de la deuxième partie du test se compose de 25 items intégrés dans un texte cohérent. Comme le précédent, ce bloc a pour objectif d'évaluer la connaissance des formes grammaticales ainsi que la compréhension du sens grammatical lui-même. Afin d'évaluer leur connaissance des formes grammaticales, les candidats sont invités à choisir la forme correcte du verbe, pronom ou déterminant ainsi qu'à sélectionner la bonne option parmi diverses propositions principales ou relatives, ainsi que des syntagmes nominaux. En ce qui concerne la distribution des formes verbales, on peut en trouver de nombreux, qui sont les suivantes :

Present progressive, will future, verbe associé à un modal;

Present progressive, simple present, une forme grammaticalement incorrecte de la négation ;

Infinitif, simple present, gérondif,

Simple past, present perfect, past perfect;

Would+infinitive, simple past, past perfect (progressive);

Would+infinitive, simple past,

Should+infitive, simple past, present perfect;

Simple past, simple past au passif; present perfect au passif;

Would+infinitive, would+ present perfect, past perfect.

Quant aux pronoms, on trouve deux types de distribution. La première présente des pronoms relatifs et la deuxième contient un pronom personnel, un pronom réfléchi et un déterminant possessif. En ce qui concerne les segments langagiers plus larges qu'un morphème, on en trouve trois types : les propositions principales, les propositions relatives (introduites par divers pronoms) et les propositions circonstancielles de temps. Pour ce qui est de l'évaluation du sens grammatical, cette partie du test propose un choix parmi divers syntagmes nominaux qui conviennent tous au niveau grammatical, mais dont seulement un seul donne à la phrase un sens cohérent.

Contrairement aux deux premiers blocs analysés, le troisième bloc constitutif de la deuxième partie est uniquement centré sur l'évaluation de la connaissance des formes grammaticales. Il comporte dix items non-intégrés qui constituent autant de phrases isolées. Ces items évaluent dans leur intégralité la connaissance du même aspect grammatical, qui est la reprise elliptique, désigné par *question tag* en anglais, et la capacité de choisir la forme qui convient grammaticalement à la proposition principale. L'exercice est facilité par la présentation des quatre phrases en amont servant d'exemples.

La deuxième moitié du test se compose de cent items non intégrés à un texte, qui portent sur la compréhension de l'oral. Comme évoqué ci-dessus, ils évaluent la compétence phonologique des candidats. Cette compétence est testée ici par le biais de la discrimination phonétique. Un tel procédé est à expliquer par le fait que la compétence phonologique se manifeste, entre autres, par «une aptitude à percevoir […] les traits phonétiques qui distinguent les phonèmes (traits distinctifs tels que, par exemple, sonorité, nasalité, occlusion, labialité)». (Conseil de l'Europe 2005 : 91). Les candidats sont invités à la discrimination phonétique de deux morphèmes similaires. Il est évident que cette instruction incite les candidats à percevoir les phonèmes non seulement en tant qu'unités sonores de la langue, mais également de reconnaitre les traits phonétiques qui distinguent les phonèmes les uns des autres, par exemple, la

sonorité et la nasalité. Quant aux options présentées, elles font partie, le plus souvent, de la même catégorie lexicale. Les catégories qui apparaissent sont les noms, les adjectifs, les verbes, les adverbes, les numéraux, les déterminants indéfinis, les syntagmes nominaux et les syntagmes adjectivaux. Les verbes apparaissent sous des formes multiples:

au participe passé ;

au présent, au *simple past* ; à l'infinitif ;

au présent et au *simple past ;*

à l'infinitif et au participe passé.

Dans un certain nombre d'items les options relèvent de différentes catégories. Ainsi, on trouve les juxtapositions d'un nom et d'un verbe à l'infinitif, d'un adjectif et d'un nom, d'une particule adverbiale et d'un participe passé, d'un numéral et d'un adverbe de temps, ainsi que les juxtapositions d'un nom ou d'un pronom avec un syntagme.

1.6.3 *Energy Placement Test*

1.6.3.1 Description et évaluation des caractéristiques du test

Energy Placement Test possède de nombreuses caractéristiques communes avec *Oxford Placement Test 2*. Ce test a été publié par *Pearson Education* en 2004.[29] Il se passe uniquement sur format papier. A la différence d'*Oxford Quick Placement Test* et d'*Oxford Placement Test 2*, sa passation est gratuite. Comme son nom l'indique, il s'agit d'un test de positionnement dont la fonction première est de « placer » le candidat dans le cours correspondant le mieux à son niveau. En marge de cet objectif principal, ce test possède trois fonctions annexes. La première est d'ordre diagnostique. Elle consiste à tester la compétence grammaticale des apprenants ainsi qu'à déterminer leurs problèmes éventuels dans ce domaine. La seconde utilisation possible est le regroupement d'étudiants par groupes de compétences. Enfin, le test peut être utilisé pour évaluer les acquis des apprenants au cours d'une période d'apprentissage. Cette dernière utilisation rapproche ce test de positionnement d'un test d'acquisition.

Le public officiellement visé par cet outil d'évaluation englobe toutes les personnes souhaitant apprendre l'anglais avec la méthode *ENERGY.* Les cours dispensés correspondent aux quatre niveaux de langue suivants: faux-débutant, élémentaire, intermédiaire et avancé. Malgré les quatre niveaux de compétences évalués, une entrée aux cours est possible aux niveaux faux débutant et élémentaire uniquement (*Teacher's notes* 2004 :7).

En ce qui concerne la conception du test, on note qu'il comporte deux versions identiques, désignées *student A test* et *student B test*. Chaque version couvre non seulement les quatre niveaux successifs, mais teste les mêmes aspects grammaticaux dans le même ordre. Le nombre d'items contenu dans les deux versions est rigoureusement identique. Il s'élève à soixante. Les compétences testées sont les compétences grammaticale et lexicale. La première est évaluée aux quatre niveaux, la seconde uniquement au niveau *faux-débutant*. Les notes d'accompagnement du test signalent que la compétence grammaticale d'un candidat peut ne pas refléter fidèlement sa compétence communicative, d'où le conseil donné aux formateurs de compléter le test de positionnement par une évaluation de l'expression orale et écrite (*Teacher's Notes* 2004 : 5).

Les tâches consistent à choisir un morphème ou un syntagme qui convienne grammaticalement à la phrase. Les formes appartiennent à l'une des catégories lexicales ou grammaticales suivantes: les verbes, les noms, les adjectifs, les adverbes, les pronoms et les déterminants. Par analogie avec les dispositifs déjà examinés, la méthode utilisée ici consiste à soumettre un nombre fixe de réponses possibles aux candidats parmi lesquelles il faut choisir. Le format appliqué est le questionnaire à choix multiples.

Le mode d'évaluation utilisé est transparent parce qu'il consiste à relier le score atteint à la performance individuelle ainsi qu'au niveau de compétence associé à ce score. La transparence est renforcée grâce à l'affichage du *cut-off* entre les quatre niveaux. Ainsi, le candidat est informé non seulement de son score et de son niveau de langue, mais aussi du score limite d'un niveau de compétence. Le seul élément affaiblissant la transparence du test est qu'il n'est pas indiqué comment le nombre de points attribué à un candidat a été obtenu,

c'est-à-dire, si ce nombre correspond exactement au nombre des bonnes réponses données. Les instructions fournies avec le test n'expliquent pas la composition du score attribué. Or, il est indispensable pour les candidats de pouvoir établir un lien entre le nombre de points obtenus et le nombre de réponses correctes ou incorrectes données pour percevoir l'évaluation comme complètement transparente.

Quant à la forme du bilan, il est établi sous forme d'un tableau qui affiche le score obtenu au test et le niveau de compétence correspondant. Toutefois, son contenu n'est pas d'une grande valeur pour le candidat, vu qu'il ne contient pas de points forts et faibles repérés dans la performance individuelle, n'offre pas la possibilité de revoir ses réponses et ne fournit pas de conseils sur l'apprentissage futur de la langue.

Les informations insuffisantes fournies dans le bilan constituent un point faible de ce test. Deux autres caractéristiques peuvent être aussi jugées négativement. Il s'agit, en premier lieu, de la non-évaluation des compétences communicatives, en réception ou en production. En deuxième lieu, on peut reprocher l'évaluation excessive de la compétence grammaticale, alors même que la compétence lexicale n'est évaluée que de façon fragmentaire et à un niveau relativement bas. Il faut souligner que ces deux caractéristiques ne constituent pas nécessairement des défauts puisque ce sont des aspects voulus par les concepteurs, assumés et évoqués dans les notes destinées aux professeurs (*Teacher's notes 2004 :5*). Il n'en demeure pas moins que ce déséquilibre est susceptible d'être perçu négativement pas les personnes non-impliquées dans la conception du dispositif.

Cela étant admis, force est de reconnaître que ce test se distingue par plusieurs aspects positifs. Ainsi, une utilisation facile grâce au format uniforme, une haute flexibilité grâce aux deux versions de même niveau, ainsi que la recension des réponses correctes dans une rubrique solutions. Ce dernier aspect implique la possibilité donnée aux candidats de vérifier leurs propres réponses. Un autre aspect positif d'*Energy Placement Test* est qu'il inclut plusieurs idées sur la façon de tester le niveau de compétence communicative à l'oral et à l'écrit.

1.6.3.2 Analyse de la typologie des exercices et des compétences

Comme il a déjà été signalé, les soixante items constitutifs de ce test évaluent la compétence grammaticale des candidats. Ils se focalisent dans leur totalité sur l'évaluation de l'une des deux composantes distinguées dans le modèle de grammaire de Purpura (2004) : les formes grammaticales. La compétence ciblée par les items est la capacité à choisir le morphème ou le syntagme grammaticalement correct. Les items comportent des options appartenant à une des catégories lexicales ou grammaticales suivantes **:** les verbes, les noms, les adjectifs, les adverbes, les pronoms, les déterminants et les syntagmes [30].

En termes des sous-compétences testées, il s'agit de l'évaluation de l'usage correct des verbes, des noms, des déterminants, des adjectifs, des adverbes et des syntagmes. Quant à la compétence verbale, on trouve les distributions des trois catégories de verbes, d'auxiliaires, de verbes lexicaux et de verbes modaux, parmi les items. En ce qui concerne les auxiliaires, on trouve les formes suivantes :

-Les formes de l'auxiliaire *be*

i) au présent, à des personnes et nombres différents. ;

ii) aux temps différents (*simple present* et *simple past*) au singulier et au pluriel ;

-Les formes de l'auxiliaire *do* au présent, variant selon les personnes et les nombres ;

-Les formes des auxiliaires *have, do* et *be* au présent, à des personnes et nombres différents ;

-Les formes des auxiliaires *have* et *do* au présent, au singulier et au pluriel, et au *simple past.*

-Les formes des auxiliaires *have* et *be* au présent, au singulier et au pluriel, ainsi qu'au *simple past*

En ce qui concerne la distribution des verbes lexicaux, on trouve des items contenant différentes formes d'un même verbe lexical aussi bien que ceux qui englobent plusieurs verbes lexicaux. Quant aux formes d'un même verbe lexical, on trouve les types de distribution suivants :

- au présent simple et progressif, au singulier et au pluriel ;

- au présent simple et progressif et au *past simple* et progressif, au singulier et au pluriel ;

-au present simple et progressif, au *past progressive* et au *present perfect progressive ;*

- au présent, *present perfect , past perfect* et au gérondif.

- au présent simple et progressif*, au will -future et au going to- future*;

-au *will future,* au *simple past* et au *would+infinitive ;*

- au *présent* et au *gérondif* ;

-à *l'infinitif* et au *gérondif* précédés de prépositions.

-aux formes finies au i) *will-future* et à *l'impératif* ou ii) au *simple past* et à l'impératif et aux formes non finies (i) à l'infinitif et au gérondif ou ii) au gérondif uniquement;

- au *simple present i*) à la voix passive uniquement ou ii) à la voix active et passive et au *simple past,* aux voix active et passive;

On trouve également des items qui soumettent les formes des différents verbes lexicaux comme options. Il s'agit de deux types de distribution:

-*le simple past* d'un verbe et le participe passé de trois verbes ;

-*le présent*, le *simple past* et le *past perfect* de deux verbes ;

Il faut noter la présence de plusieurs items contenant des verbes modaux. Les distributions suivantes entre les formes de cette catégorie verbale sont à repérer :

-deux verbes lexicaux au présent, un verbe lexical associé à un modal, un modal.

-deux modaux différents, à la forme affirmative et négative, et le verbe *have to,* sémantiquement proche d'un modal i) au singulier ii) au pluriel ;

-trois modaux, à la forme affirmative et négative ;

-le même modal aux formes affirmative et négative, le verbe *have to*, sémantiquement proche d'un modal, et un verbe lexical.

Pour évaluer la compétence à utiliser correctement les formes nominales, deux types de distributions apparaissent:

-Un nom irrégulier au singulier, au nominatif et au génitif, et au pluriel, au nominatif, ainsi que la formation régulière du pluriel de ce nom, qui est agrammaticale pour le nom irrégulier.

-Un nom régulier au singulier, au nominatif et au génitif, et au pluriel, au nominatif et au génitif.

L'usage correct des pronoms est évalué à l'aide des distributions des formes pronominales suivantes:

- Les pronoms personnels variant selon la personne, le nombre et le cas, le nominatif et l'accusatif ;

- Les pronoms non-référentiels *It* et *There* et les pronoms personnels variant selon les personnes ;

-Les pronoms démonstratifs se distinguant selon le nombre et la proximité.

-Les pronoms interrogatifs ;

-Les pronoms relatifs ;

- Les pronoms personnels et les pronoms possessifs variant selon les genres et les personnes respectivement.

En ce qui concerne les déterminants, l'usage correct des formes relevant de cette catégorie grammaticale est testé à l'aide des distributions suivantes :

-Deux quantifieurs indéfinis et i) deux articles indéfinis, en deux versions (*a* et *an*) ou ii) un article défini et un article indéfini,

- Un quantifieur indéfini, l'article indéfini en deux versions (*a* et *an),* et un numéral;

- Les déterminants possessifs variant selon les genres, les personnes et les nombres ;

La capacité à employer correctement des formes adjectivales est évaluée par le biais des deux distributions suivantes :

-Deux adjectifs au positif, un adjectif au comparatif et un adjectif au superlatif, formés sans adverbes *more* et *most* ;

-Un adjectif au positif, deux adjectifs au comparatif, formés i) avec et ii) sans adverbe *more,* et un adjectif au superlatif, formé à l'aide de *most.*

La connaissance des adverbes et la capacité à utiliser cette catégorie lexicale correctement sont testées à l'aide des trois distributions, à savoir :

- trois adverbes et un adjectif ;

- quatre adverbes de temps ;

- deux adverbes de degré ainsi que *such* i) avec ou ii) sans article indéfini ;

La capacité à utiliser le syntagme qui convient est évaluée par le biais des distributions suivantes:

- un sujet, qui est un pronom personnel, et un prédicat, qui est un verbe lexical soit au *simple présent* soit au *present progressive soit au simple past,* précédant ou succédant au sujet. Le prédicat est formé de manière suivante :

i) à l'aide de l'auxiliaire *do* au présent ou au *simple past* ou sans auxiliaire ou ii) à l'aide de l'auxiliaire *do* au présent ou au *simple past* ou de l'auxiliaire *be* au présent,

- à l'aide d'un même verbe lexical et d'un adverbe de fréquence précédant ou succédant au verbe;

-à l'aide d'un auxiliaire et d'un adverbe de temps ;

-à l'aide d'un même adjectif précédé ou succédé par un adverbe de degré ;

-à l'aide de l'auxiliaire *do*, du sujet et de l'adverbe intensificateur en position initiale;

- à l'aide de questions indirectes contenant un verbe lexical et un sujet, introduites par l'auxiliaire *do* ou par une conjonction conditionnelle ;

- à l'aide du verbe *make,* faisant partie d'une proposition causative, i) suivi par un pronom personnel en fonction du COD ou ii) situé seul.

L'analyse qui vient d'être menée d'*Energy Placement Test* montre que ce dispositif a recours à de nombreux types de distribution, majoritairement complexes, pour évaluer la capacité d'utilisation correcte des formes grammaticales. Bien que la compréhension du sens grammatical ne soit pas évaluée par ce dispositif, la complexité des types de distribution observés dans ce test est élevée.

1.6.4 *Upstream Enterprise Placement Test*

1.6.4.1 Description et évaluation des caractéristiques du test

Upstream Enterprise Placement Test est un test de positionnement polonais, publié par l'édition *Egis* en 2010. Il est accessible gratuitement sur le site suivant: http://de.scribd.com/doc/27535540/Placement-Test-Upstream-Enterprise. Ce test remplit deux fonctions typiques, étroitement liées pour cette catégorie de tests qui consistent, premièrement, à placer les apprenants dans la classe d'anglais qui correspond à leur niveau de compétence et, deuxièmement, à regrouper les étudiants des classes différentes ayant le même niveau de compétence. Le public visé est toute personne qui souhaite se positionner. Cet instrument d'évaluation couvre en effet sept niveaux de compétences dès A1 jusqu'au C1: *Upstream Beginner A1; Upstream Elementary A2; Upstream Pre-Intermediate B1 ; Upstream Pre-Intermediate B1+ ; Upstream Intermediate B2; Upstream Upper-Intermediate B2+; Upstream Advanced* C1. On constate l'apparition des mêmes niveaux de compétences que dans le CECRL tout en notant l'absence des niveaux A2+ et C2. Ce dispositif est composé de deux parties, A et B, qui ne sont pas conçues de la même manière, notamment du point de vue formel. La partie A aligne 60 items construits à partir de phrases isolées tandis que la partie B inclut 20 items intégrés à un texte. La différence formelle entre les deux parties est cependant aplanie par l'adoption du même format QCM. Les compétences évaluées sont d'ailleurs les mêmes, à savoir, la compréhension écrite ainsi que les compétences lexicale et grammaticale. La partie A est centrée sur la

compétence communicative. Le contenu des tâches dépend des compétences évaluées. Les tâches ciblant la compétence grammaticale consistent à choisir le bon morphème ou syntagme pour que la phrase soit grammaticalement correcte. Les tâches portant sur la compétence lexicale impliquent le choix d'un morphème en adéquation avec le sens de la phrase. Les tâches portant sur la compétence communicative consistent à sélectionner l'énoncé correct du point de vue de la convenance communicative. En ce qui concerne le mode d'évaluation, le niveau de compétence est attribué en fonction du score atteint, sans affichage du pourcentage de bonnes réponses. L'évaluation possède un haut degré de transparence néanmoins grâce à l'indication du score obtenu, du *cut-off* entre les niveaux et de l'affichage des réponses correctes. Les résultats sont fournis sous forme de tableau. On relève toutefois que ce test manque d'un bilan contenant des conseils sur la possibilité d'améliorer sa manière d'apprendre la langue. De même que tous les dispositifs évalués, ce test a ses qualités et ses défauts. Ses principaux points positifs sont l'évaluation de la compétence communicative, même si celle-ci n'est représentée que par quelques tâches seulement, ainsi que l'emploi uniforme du QCM qui est un format facile pour l'utilisateur. Le troisième atout est l'indication des réponses correctes. Les inconvénients *d'Upstream Enterprise Placement Test* sont l'absence d'évaluation des trois compétences communicatives, de la compréhension orale, ainsi que de l'expression écrite et orale, l'absence de bilan aussi bien que l'attribution d'un score global uniquement.

1.6.4.2 Analyse de la typologie des exercices et des compétences

Les exercices varient selon les compétences évaluées. Dans la partie A, les tâches se répartissent en exercices de grammaire, de lexique et de communication. Les exercices de grammaire reposent sur le choix d'une forme syntaxiquement correcte, appartenant à l'une des catégories lexicales ou grammaticales évaluées. Est évalué, en tant que sous-compétence, l'usage correct des verbes, y compris des verbes lexicaux, des auxiliaires et des modaux, des noms, des pronoms, des déterminants, des adjectifs et des adverbes, des prépositions et des particules adverbiales, des conjonctions et des syntagmes. En ce qui concerne les auxiliaires, les deux distributions suivantes y apparaissent:

Les formes de l'auxiliaire *be* au *simple present* et au *simple past* au singulier et au pluriel.
Les formes de l'auxiliaire *do* au *simple present* et au *simple past*, les modaux *will* au futur et *would* ;

Les verbes lexicaux se trouvent aux temps, aspects, modes et voix différents dans un grand nombre d'items:

-au *simple present,* au *present progressive* et au *present perfect* à la 3 Ps. Sg., ainsi qu'au *simple present* aux autres personnes
-au simple present, au *present progressive,* au *past progressive* et au *simple past ;*
-au present progressive, au *past progressive,* au *present perfect progressive* à la 3 Ps. Sg *et going to +infinitive* à la 3 Ps. Sg *;*
-au *will-future,* au *simple past,* au *present perfect* et au *simple present ;*
au *present progressive,* au *simple present ,* au *present perfect à la 3 Ps. Sg,* et au *simple past.*
-au *simple past* aux voix active et passive*,* au *past perfect* aux voix active et passive ;
- au *simple past*, au *past progressive* à la 3 Ps. Sg, au *past perfect* et au *present perfect progressive* ;
- au *simple present,* au *present progressive,* au *present perfect* et au *futur antérieur* (*will + present perfect*)
- au will-future, present perfect, simple past et au conditionnel (would + infinitive) ;
à l'infinitive, au simple present et au gérondif i) précédé d'une préposition ii) sans préposition ;
-au *will-future*, au *conditionnel,* au *present perfect* et au *past perfect* à la 3 Ps. Sg.
au *présent* aux voix simple et progressive, au *will+ infinitive* et au *will be + -ing* ;
- à l'infinitif, au *gérondif,* au *present perfect,* au *past perfect* ;
-au *simple present* à la 3 Ps Sg, au *simple past,* au *present perfect* et au *will future*
- au *will future,* au *simple present,* au *simple past et* au *past perfect ;*
-au *present perfect,* au *present perfect progressive,* au *past progressive* et au *past perfect ;*

L'usage des verbes modaux est testé moyennant les items dans lesquels ils apparaissent dans les constellations suivantes :

-must, should, can et l'auxilliaire *do*
-should, can et l'auxiliaire *do* au present et au *simple past*;
-les modaux *will* au futur et *would*, l'auxiliaire *do* au *présent* et au *simple past*
-must à la forme affirmative et négative, *can* et *have to*, sémantiquement proche d'un modal, et un verbe lexical
-must à la forme affirmative et négative, *can* et need, sémantiquement proche d'un modal.
- *can* et *could*, les deux à la forme affirmative et négative.
-would + infinitive, *might + infinitive*,*must + infinitive perfect.*

La capacité à utiliser correctement les noms est testée par les items qui contiennent les noms suivis par un groupe prépositionnel à choisir pour sa compatibilité avec la construction. Le bon usage des pronoms est évalué par le biais de deux constellations**:**

- pronoms indéfinis
- pronoms relatifs.

L'emploi correct des déterminants est testé à l'aide des distributions suivantes :

- quatre quantifieurs ;
- deux quantifieurs, un article défini et un article indéfini ;
-trois quantifieurs et un groupe nominal ayant la fonction d'un quantifieur.

Les adjectifs sont distribués de la façon suivante dans les items :

-le même adjectif aux trois formes : i) au positif, précédé par l'adverbe de comparaison « as », à la forme affirmative et négative, ii) au comparatif, formé à l'aide du suffixe *–er* et au superlatif, formé à l'aide du suffixe *–est* ;

-au positif précédé par l'adverbe de comparaison « as » à la forme affirmative et négative, au comparatif, formé à l'aide de l'adverbe *muc*h, et au superlatif, formé à l'aide du suffixe *–est*

La rectitude de l'usage des adverbes est évaluée par l'item qui comporte trois adverbes de degré et un adverbe de comparaison. L'emploi correct des catégories grammaticales, en l'occurrence, des prépositions, des particules adverbiales et des conjonctions, est également testé dans le but d'évaluer la compétence grammaticale:

-quatre prépositions suivies par un verbe prépositionnel ;
-quatre particules adverbiales;
-quatre conjonctions exprimant la négation

Enfin, les syntagmes interviennent dans les items centrés sur la compétence grammaticale. Ils se composent des catégories lexicales suivantes :

-d'un sujet (un pronom personnel) et d'un verbe lexical au *futur,* au *simple past* ; au *present perfect* et au *simple present,* formé à l'aide du modal *will* ou des auxiliaires *do* et *have,* selon les temps ;
-d'un sujet (un pronom personnel) et du verbe lexical *to be* au *present* ; à une forme future, au *present perfect* et au *simple past* ainsi que d'un même adverbe de fréquence ;
- d'une reprise elliptique composée d'un sujet et des auxiliaires *to do* et *to be* ;
-d'un sujet précédé par un déterminant et du verbe au présent et au futur, les deux à la forme affirmative et négative ;
-d'un sujet (pronom personnel) et d'un verbe lexical au *present* (*simple et progressive*), au *past perfect* et au *present perfect progressive,* tous à la forme interrogative ;

Quant aux exercices de lexique dans la section A, ils se répartissent en deux types, les exercices focalisés sur la forme lexicale et ceux qui ont pour l'objet le sens lexical. Le premier type d'exercices évalue la capacité de choisir le bon lexème pour former un bloc lexicalisé. La forme lexicale est mise en avant dans les items contenant les verbes, les noms et les adjectifs, avec les distributions suivantes:

-quatre verbes dont un forme un bloc lexicalisé avec le nom qui suit ;
-quatre noms dont un forme une collocation avec le verbe précédent ;
-quatre adjectifs suivis par un groupe prépositionnel dont il faut choisir celui compatible avec la préposition.
- trois adjectifs et un nom suivis par un groupe prépositionnel dont seulement un seul est utilisable avec la préposition qui suit ;

Le deuxième type d'exercices lexicaux propose de choisir un morphème lexical ou grammatical qui convient sémantiquement à la phrase. Les morphèmes impliqués sont des noms, des verbes, des adjectifs et des adverbes. Concernant l'usage des noms, on rencontre les trois distributions suivantes:

-quatre noms qui expriment une quantité suivie par la matière contenue ;
-quatre noms désignant des objets complètement distincts;
-quatre noms désignant la manière d'être payé. Le choix porte sur un synonyme de la première partie de la phrase ;

L'emploi correct des verbes du point de vue lexical est évalué moyennant les cinq distributions qui suivent:

- quatre verbes de mouvement différents exprimant tous un mouvement vers qqch et qui sont tous à la même forme (à l'indicatif présent)
- quatre verbes à particule dont les deux ont la même base verbale et se distinguent par leur particule adverbiale, et qui sont tous à l'impératif ;
- quatre verbes différents au *simple past* ayant un sens de perte ou d'éloignement
-quatre verbes à l'infinitif, avec un sens différent ;
- quatre verbes désignant un changement de vitesse ou de position dans l'espace. Il faut choisir celui qui est synonyme de la première partie de la phrase;

Quant aux adjectifs, trois répartitions surgissent dans le test afin de tester la compétence lexicale. Il s'agit de:

- deux adjectifs, un nom et une préposition de comparaison ;
- des adjectifs avec un sens différent parmi lesquels il faut choisir un synonyme de la première partie de la phrase ;
- quatre adjectifs, dont les deux mêmes, à la forme affirmative et négative;

`

En outre, la compétence lexicale est ciblée par quatre adverbes de fréquence, dont il faut choisir celui qui convient sémantiquement.

Les exercices ayant pour cible la compétence communicative consistent à choisir un énoncé adapté à un geste, à une situation particulière ou à un énoncé de l'interlocuteur, du point de vue de la convenance communicative. Trois options suivantes servent à cet effet :

-Choix d'un énoncé qui accompagne l'acte de donner quelque chose à quelqu'un
-Choix d'une bonne réponse au remerciement de quelqu'un.
-Sélection d'une bonne phrase pour décrire le temps pluvieux

A la différence de la partie A, les items dans la partie B d'Upstream *Enterprise Placement Test* apparaissent sous forme de phrases intégrées dans un texte cohérent. Cette section se compose de deux textes dont chacun comporte le même nombre d'items. Ces derniers évaluent soit la compétence grammaticale soit la compétence lexicale. En ce qui concerne la compétence grammaticale, celle-ci est testée par les items contenant des options qui relèvent des catégories

suivantes : verbes, prépositions, adverbes, déterminants et pronoms. Le bon usage des verbes du point de vue grammatical est testé par les items suivants :

-les verbes au *participe passé* suivis par deux groupes prépositionnels, dont il faut choisir celui compatible avec ces prépositions ;
-les verbes au *simple past* suivis par un groupe prépositionnel ;
- les verbes au présent suivis par le verbe à l'infinitif sans particule *to*

L'emploi grammaticalement correct des prépositions est évalué par deux items comportant quatre prépositions suivies d'un complément de lieu dans le premier cas et d'un complément de temps dans le second. Quant aux adverbes, ils sont proposés au choix par un item qui inclut deux adverbes de temps et deux adverbes de fréquence. Le choix du bon déterminant est évalué au moyen de deux items dont le premier offre au choix quatre déterminants indéfinis et le deuxième trois numéraux et un adjectif. En ce qui concerne les pronoms, ils se trouvent dans un item qui inclut quatre pronoms relatifs introduisant une proposition subordonnée se référant à un nom.

La compétence lexicale est évaluée dans la première partie de la section B au moyen de deux items suivants qui proposent uniquement les noms en tant qu'options :

-quatre noms qui désignent le monde physique en partie ou en entier ou le monde naturel, mais qui ne sont pas interchangeables.
-quatre noms qui désignent des catégories de personnes différentes

Quant aux items contenus dans le texte B, la compétence grammaticale y est testée à l'aide des items comportant des choix à opérer parmi des verbes, des adjectifs, des adverbes ou des conjonctions. Concernant les verbes, ils apparaissent dans les deux items dont le premier propose le même verbe au *present* et au *simple past*, à la voix active et passive. Le deuxième item relevant de cette catégorie contient les participes passés de quatre verbes différents. L'usage des adjectifs, des adverbes et des conjonctions est évalué par un seul item respectivement. Celui focalisé sur l'emploi de la premiére catégorie évoquée offre au choix un numéral et trois adjectifs. Celui centré sur les adverbes, contient deux adverbes et deux prépositions. L'emploi des conjonctions, enfin, est testé au moyen des quatre conjonctions introduisant une proposition subordonnée.

En ce qui concerne la compétence lexicale, les deux composantes de celle-ci, la forme et le sens lexical, sont évaluées par les items inclus dans le texte B. La forme lexicale est testée par un seul item proposant un choix à

effectuer parmi quatre noms pour former un bloc lexicalisé avec le verbe précédent. Quant au sens lexical, il est traité par deux items présentant uniquement un problème de distribution d'adverbes:

- adverbes d'ajout, d'illustration et de résumé
-trois adverbes de liaison et un adverbe de comparaison

1.6.5 *Cutting Edge Placement Test*

1.6.5.1 Description et évaluation des caractéristiques du test

Les dispositifs auxquels nous allons nous intéresser maintenant présentent une plus grande complexité que ceux que nous avons examinés jusqu'ici car ils se composent de plusieurs tests simples. Le premier de ces dispositifs complexes est le *Cutting Edge Placement Test*. Ce dernier a été publié par les éditions *Longman* et il a été conçu pour un usage gratuit. Conformément à la fonction habituelle des tests de positionnement, sa fonction consiste à évaluer les connaissances des candidats en anglais afin d'identifier le niveau individuel le plus favorable à la poursuite de l'apprentissage de la langue. Ce niveau de reprise est désigné par l'expression *cut-off point* dans la documentation adjointe au test (Notes & Answer Key). Contrairement à d'autres tests de positionnement qui remplissent également une fonction diagnostique, *Cutting Edge Placement Test* n'a d'autre fonction que de répartir les candidats par niveaux. Ceci est explicitement signalé aux utilisateurs (*Notes & Answer Key*). Le test ne doit pas non plus être utilisé à la manière d'un test d'acquisition, pour déterminer si les objectifs visés par un programme d'enseignement ont été atteints par les candidats au bout d'une certaine période de temps (Alderson, Clapham & Wall 1995:10). Bien que la recherche effectuée dans ce domaine, par exemple, par Brown (2010: 9-10), ait révélé que beaucoup de tests de positionnement et d'acquisition ont une fonction diagnostique secondaire, ce constat n'est pas accepté par les développeurs de *Cutting Edge Placement Test*. Le public visé par cet instrument d'évaluation est toute personne, quel que soit son niveau de langue, souhaitant évaluer son niveau de compétence.

Les cinq tests constitutifs du dispositif intégral recouvrent cinq niveaux successifs, à savoir, *Starter (débutant complet), Elementary (élémentaire), Pre-*

intermediate (pré-intermédiaire), Intermediate (intermédiaire) ainsi qu'Upper Intermediate (intermédiaire supérieur). Une contrainte temporelle est imposée qui s'élève à 1 heure 10 minutes pour l'ensemble de l'évaluation. La durée de la section ciblant la compréhension de l'oral est limitée à 20 minutes, tandis que la section évaluant les compétences écrites doit être parcourue en 50 minutes.

Le dispositif est constitué de cinq tests (A-E) qui évaluent les aspects de la langue enseignée aux niveaux couverts par la collection *New Cutting Edge*. Ces niveaux sont évoqués au-dessus *(Notes & Answer Key*). Chaque test est composé de la même manière et comporte deux grandes sections, dont la première évalue la compréhension de l'oral et la deuxième traite de la compréhension de l'écrit. Concernant le test de compréhension orale, ce dernier est constitué d'une seule partie renfermant un enregistrement. Variant en fonction des tests, l'enregistrement prend la forme d'un récit du passé, d'un reportage, d'une interview, d'une conversation et d'un entretien téléphonique.

La deuxième section, intitulée « test de lecture » (*Reading section*), est beaucoup plus complexe aussi bien par sa structure formelle que par le répertoire des compétences évaluées. Cette section est composée de quatre parties dont les trois premières contiennent quatre items et la dernière huit. La compétence évaluée par la première partie de la deuxième section est la compréhension écrite tandis que l'habileté testée par la deuxième partie varie entre la compétence lexicale et grammaticale, selon le test. Quant à la troisième partie de la deuxième section, la compétence communicative y est évaluée tandis que la dernière partie se focalise sur l'évaluation de la compétence grammaticale des candidats.

Le contenu des items dépend de leur fonction. Il est donc déterminé par la compétence à évaluer. Les items servant à évaluer la compréhension orale et écrite consistent à comprendre des textes oraux ou écrits et à choisir soit une seule réponse correcte parmi les trois ou quatre possibilités données, soit à retenir quatre énoncés corrects parmi huit ou dix options. Les items qui testent la compétence grammaticale sont de deux types différents.[31] Les représentants du premier type invitent les candidats à compléter des phrases à trous par un morphème ou par un syntagme grammaticalement correct, à choisir parmi les

trois ou quatre options données. Les items de deuxième type évaluant la compétence grammaticale contiennent deux phrases isolées qui sont similaires au niveau du lexique et de la grammaire, mais qui révèlent tout de même certaines différences lexicales ou grammaticales. Ces items conduisent les candidats à décider si le sens de deux phrases est le même ou différent. Il faut noter que les items de deuxième type évaluent aussi bien la compétence lexicale que la compétence grammaticale, même si ce fait n'est pas indiqué dans l'explication jointe (Notes & Answer Key). Bien que la compétence lexicale soit également nécessaire pour insérer le morphème grammatical ou le syntagme correct dans une phrase à trou, elle a une plus grande importance dans les items de deuxième type. La compréhension du sens de la phrase est encore plus importante ici pour les candidats, puisqu'il est impossible de répondre en se focalisant uniquement sur le sens des options proposées.

Les items évaluant la compétence lexicale procèdent par sélection d'un morphème lexical qui se trouve en relation sémantique avec un mot donné. Cette relation sémantique est soit celle d'antonymie soit de synonymie soit d'hyponymie.[32] La compétence lexicale est évaluée dans ces items au travers de la compréhension de morphèmes isolés et de la capacité qu'a le candidat à les utiliser. Le deuxième type d'items évaluant la compétence lexicale consiste à choisir un mot ayant la même voyelle orale qu'un autre mot signalé dans la phrase de référence. Selon le répertoire des compétences linguistiques présentées dans le CECRL, ce type d'item évalue la compétence phonologique, car trouver un mot contenant un même son vocalique présuppose « une aptitude à percevoir et à produire les unités sonores de la langue (phonèmes et leur réalisation dans des contextes particuliers » (CECRL 2005 : 91). L'analyse de *Cutting Edge Placement Test* effectuée dans ce chapitre indique que la compétence évaluée est *lexicale* afin de simplifier le classement d'items et afin de rester conforme au classement entrepris dans la clé des notes et des réponses de ce dispositif.

Les items évaluant la compétence communicative consistent à tester la capacité à utiliser les énoncés en contexte. Le contexte est fourni par l'énoncé déclaratif ou interrogatif servant de stimulus. Dans le test A, il ne faut pas choisir l'énoncé qui convient au stimulus, mais déterminer le lieu typique où cet énoncé

serait prononcé. La compétence communicative évaluée par ce type d'items concerne la *compétence sociolinguistique* évoquée dans le CECRL (Conseil de l'Europe 2005 : 84, 93), car la capacité de choisir la réponse qui convient à un énoncé déclaratif ou interrogatif donné indique l'aptitude à « faire fonctionner la langue dans sa dimension sociale » (Conseil de l'Europe 2005 : 84, 93). Cette aptitude implique la capacité de participer aux relations sociales de façon efficace en utilisant des expressions courantes ou toutes faites, et ainsi, de mener à bien différents types d'échanges sociaux (Conseil de l'Europe 2005 : 95). Il est clair que la compétence communicative testée par ces items est différente des compétences communicatives langagières qui sont linguistiques.

La méthode utilisée par le test consiste à présenter les items au format QCM et à demander aux participants de choisir la réponse correcte parmi deux, trois ou quatre options données. Le nombre d'options disponibles varie en fonction des items. En outre, le format *QCM* est parfois combiné avec le format *questionnaire à trous*. Le mode d'évaluation consiste à calculer et à afficher un score séparé pour la performance à chacun des cinq tests (A-E). Cette manière de mesurer la performance rend son mode d'évaluation transparent. La transparence est encore renforcée par la corrélation entre le score obtenu à chaque test et le nombre de bonnes réponses données, ainsi que l'indication du score limite de validation de niveau dans les cinq tests, qui s'élève à 19 points sur 24. Le score atteint dans chaque test est affiché sous forme de tableau. La structure, très claire, renforce l'impression de transparence dans l'évaluation.

En résumé, le *Cutting Edge Placement test* possède de nombreux aspects positifs mais aussi quelques points négatifs. Parmi les points forts figurent de toute évidence une évaluation de la compétence communicative, un haut degré de transparence, une orientation facile pour l'utilisateur grâce à l'affichage d'un score séparé pour chaque test sur un même tableau. Un autre atout est l'indication des solutions dans la liste des réponses attendues qui permet aux candidats de vérifier leurs propres réponses. Néanmoins, on relève aussi certains éléments critiques. En premier lieu, les scores attribués ne sont pas appariés avec les niveaux communs de référence du CECRL. En deuxième lieu, les deux compétences productives, l'expression écrite ainsi que l'expression orale, ne sont pas évaluées. De même, il manque des commentaires sur les

compétences des participants, par exemple, en indiquant les points forts et faibles de leur performance. Il manque également des conseils sur l'amélioration de l'apprentissage de la langue.

1.6.5.2 Analyse de la typologie des exercices et des compétences

La première partie du test, centrée sur l'évaluation de la compréhension de l'oral, contient deux types d'items au format *QCM*. Le premier consiste à choisir plusieurs réponses correctes parmi huit ou dix options données. Le deuxième consiste à sélectionner un seul énoncé convenable parmi les quatre alternatives proposées. Les items inclus représentent les différents types d'activités de compréhension de l'oral. Les enregistrements A et D renvoient à des récits personnels, situés dans le passé pour le premier message, et dans le présent pour le deuxième. La tâche B est organisée autour d'une interview, la tâche C représente une conversation téléphonique tandis que celle cataloguée E une conversation en face à face. Ces activités de compréhension de l'oral évaluent plusieurs sous-compétences de la compréhension de l'oral, en l'occurrence « comprendre une interaction entre locuteurs natifs » et « comprendre en tant qu'auditeur » (Conseil de l'Europe 2005 :55, 56).

La deuxième partie, composée d'activités de compréhension écrite, englobe quatre sous-parties. Celles-ci contiennent des items de cinq types différents, qui se répartissent en fonction des parties. La première partie des cinq tests (A-E) contient des items qui évaluent la compréhension écrite ainsi que plusieurs sous-compétences, en l'occurrence, la compréhension de textes cohérents ainsi que la compréhension de questions et de réponses à choix.

La deuxième partie des tests, composés des items numérotés 5 à 8, englobe soit des exercices de grammaire soit de vocabulaire. Il faut noter que tous sont au format QCM. Concernant les items de grammaire, ils se focalisent sur la compétence grammaticale qui est testée par le biais de phrases lacunaires auxquelles manquent un pronom personnel ou possessif, ou encore une préposition.

Les items qui évaluent la compétence lexicale, dans le test B, procèdent de manière directe en testant la compréhension du sens des lexèmes. La compréhension du sens est testée par la connaissance des mots ayant un lien sémantique particulier au mot donné : antonymie, synonymie ou hyponymie. Les

items du test D évaluent la compétence phonétique en invitant les candidats à choisir le mot qui rime avec un mot particulier. Ces items testent bien « la connaissance de la perception et de la production des unités sonores de la langue » des candidats (Conseil de l'Europe 2005 :91).

Les items numérotés 9 à 12 sont centrés sur la compétence communicative. La compréhension de l'écrit est testée également car il faut comprendre le sens des énoncés donnés et des options proposées. Cependant, la sémantique textuelle n'est pas seule en cause. Pour choisir la réponse correcte, la compétence sociolinguistique est nécessaire. Celle-ci, rappelons-le, porte sur « la connaissance et les habiletés exigés pour faire fonctionner la langue dans sa dimension sociale » (Conseil de l'Europe 2005 : 93). Pour donner un aperçu des différentes sous-compétences communicatives évaluées dans cette troisième partie, nous dirons que les connaissances sociolinguistiques de base sont dominantes dans le test A, tandis que les autres tests, B à E, évaluent la connaissance des syntagmes et des phrases typiquement utilisés dans le cadre d'échanges sociaux courants. Parmi les tournures langagières incluses dans les items figurent celles qui marquent des relations sociales ou la politesse.

La quatrième partie de chaque test est composée d'items évaluant la compétence grammaticale. Cette compétence est testée au travers de deux types d'items. Le premier type, apparaissant dans les tests A, B et C, consiste à compléter les phrases par un morphème lexical ou grammatical, ou encore par un syntagme adéquat. La deuxième catégorie d'items, utilisée dans les tests D et E, contient deux phrases juxtaposées dont il faut comprendre et comparer le sens.

1.6.6 *Success Placement Test*

1.6.6.1 Description et évaluation des caractéristiques du test

Comme le *Cutting Edge Placement*, le *Success Placement Test* est également un instrument d'évaluation composé de plusieurs tests simples. En l'occurrence, il s'agit d'un recueil de trois tests, dont chacun est prévu pour une durée d'une heure. Ce dispositif a été édité par *Pearson Longman* en 2007.[33] Puisqu'il s'agit d'un test de positionnement, sa première fonction consiste à évaluer le niveau de compétence en langue de l'utilisateur. La deuxième fonction de ce test est de

placer les étudiants dans la classe d'anglais de la série *Success* correspondant au niveau individuel du candidat. Les trois tests constitutifs du recueil couvrent quatre niveaux de compétence: élémentaire, pré-intermédiaire, intermédiaire et intermédiaire supérieur. Le premier test couvre les niveaux *élémentaire* et *pré-intermédiaire tandis que* le deuxième se focalise sur les niveaux *pré-intermédiaire* et *intermédiaire*. Le troisième test contient les items des niveaux *intermédiaire* et *intermédiaire supérieur*. La conception des trois tests et le nombre d'items contenus sont identiques, à savoir une centaine.

Les compétences évaluées par le recueil de tests *Success* sont la compétence grammaticale et lexicale ainsi que la compréhension de l'écrit. La compétence grammaticale est testée au travers du choix de morphèmes lexicaux ou grammaticaux qui sont syntaxiquement corrects. Les options contenues dans les items destinés à tester cette compétence relèvent de la classe des noms, pronoms, verbes, adjectifs, adverbes, prépositions ou encore déterminants. Dans un nombre important d'items, les options présentées portent sur les propositions. Quant aux formes verbales, elles se repartissent dans les catégories suivantes : verbes lexicaux à des temps et modes différents ; verbes modaux à des temps différents ; verbes et expressions sémantiquement proches des modaux. La compétence lexicale est évaluée au travers de choix d'unités sémantiquement pertinentes. Cette tâche implique la compréhension de morphèmes isolés. La compréhension de l'écrit est évaluée au moyen de phrases lacunaires à compléter. Il faut noter que tous les items sont au même format, qui est le questionnaire à choix multiples.

Le résultat global est fourni au candidat sous forme de pourcentage des bonnes réponses. Le pourcentage des réponses correctes a une fonction décisive pour le placement des apprenants. Si le résultat est inférieur à 60 %, les apprenants sont orientés vers un cours de langue de niveau immédiatement inférieur : par exemple, cours « élémentaire » en cas de passation du test de positionnement « élémentaire-pré-intermédiaire ». En cas de résultat supérieur à 70%, il est préconisé de placer les apprenants dans la classe de langue correspondant à deux niveaux supérieurs. Si le résultat obtenu au test est situé entre 60% et 70%, la décision sur le placement dans le cours de langue est à prendre selon le niveau du reste de la classe (*Success Placement Test 2007 :1*).

Malgré l'utilité évidente de l'affichage du pourcentage pour le placement des candidats, le degré de transparence de ce mode d'évaluation reste moyen. En effet, l'affichage du pourcentage des bonnes réponses est réalisé sans indication du score véritablement atteint dans l'échelle d'évaluation. Ce manque de transparence est renforcé par l'absence de bilan de compétences d'un candidat. On pourrait imaginer, par exemple, de brefs commentaires sur ses atouts et ses points faibles. De même, des conseils manquent sur la possibilité d'améliorer la manière d'apprendre la langue. On constate encore deux autres points négatifs propres à ce test, dont le premier est l'absence d'alignement du résultat atteint, affiché sous forme du pourcentage, sur les niveaux communs de référence du Cadre Européen. Enfin, si l'évaluation de la compétence grammaticale et lexicale est bel et bien réalisée, tout comme celle de la compréhension écrite, trois activités langagières fondamentales (compréhension orale, expression écrite et orale) sont mises à l'écart. Malgré ces lacunes, il faut reconnaître à ce test plusieurs points positifs. En premier lieu, l'utilisation d'une seule méthode, QCM, dans tous les items. Cette uniformité facilite la passation du test par les candidats. En deuxième lieu, l'indication des réponses correctes dans la clé de réponses, qui offre la possibilité au candidat de vérifier les options choisies.

1.6.6.2 Analyse de la typologie des exercices et des compétences

Le premier test du recueil, *Success Elementary-Pre-intermediate placement test*, contient deux types d'items : grammaticaux et lexicaux. Les items qui ciblent la compétence grammaticale consistent à compléter une phrase lacunaire par un représentant des catégories lexicales ou grammaticales, en l'occurrence, les verbes, les adjectifs, les adverbes, les déterminants, les pronoms, les prépositions et les particules adverbiales. En outre, certains items mettent les propositions au choix. Quant à la compétence verbale, on trouve les distributions suivantes entre les verbes lexicaux:

> -les formes des verbes *be, do* et *have* aux temps différents : au présent, au *simple past*, au *present perfect,* et aux *formes infinites* (l'infinitif et le gérondif) ;
>
> -les verbes *have*, *have got* et *have to* aux formes déclaratives et interrogatives, à des personnes et à des nombres différents;
>
> - les formes non finies (*infinitif, gérondif* et *participe passé)* et finies d'un même verbe, utilisées à des personnes, temps et aspects différents ;

En ce qui concerne les verbes auxiliaires, on trouve les distributions suivantes :

- les formes différentes des verbes auxiliaires *be* et *do* qui varient selon le nombre, la personne et le temps ;

- les formes des verbes auxiliaires *do*, *be* et *have* au présent et le modal *will* au futur.

- les reprises elliptiques;

-les modaux et les constructions verbales sémantiquement proches des modaux comme *have to*.

En ce qui concerne les items contenant les adjectifs et les adverbes, on trouve les variations suivantes:

-les formes d'un même adjectif au positif, comparatif et superlatif;

- les adjectifs au positif et au comparatif, les adjectifs associés aux adverbes de manière ainsi que les adverbes au positif et au comparatif ;

- les formes d'un même adverbe au positif, au comparatif et au superlatif.

Par rapport à l'occurrence des déterminants et des pronoms dans les items, on constate les variations suivantes :

- articles indéfinis, définis ou aucun article ;

- articles indéfinis, définis, déterminants démonstratifs ;

- pronoms et déterminants possessifs, pronoms personnels

- pronom non-référentiel *there* combiné avec les formes conjuguées de l'auxiliaire *be*.

- quantifieurs (déterminants) indéfinis ;

- pronoms indéfinis ;

- pronoms interrogatifs liés à l'auxiliaire *do* au présent et aux auxiliaires *be* et *do* au *simple past* ;

En ce qui concerne la distribution des prépositions et des particules dans ce premier test constitutif du recueil, on trouve:

- les prépositions qui se réfèrent soit à un verbe prépositionnel ; soit à un complément circonstanciel de temps.

- les particules adverbiales dont une seule convient au verbe régissant.

Pour clore la liste des items évaluant la compétence grammaticale, il convient d'évoquer ceux contenant des propositions. On trouve deux types de propositions qui se composent respectivement:

- du pronom non-référentiel *there*, du verbe lexical *be* et d'un déterminant indéfini, aux formes affirmative et interrogative ;

- d'un sujet et d'un prédicat (d'un seul verbe au *simple présent* ou au *present progressive*);

Ce premier test dans le recueil contient également les items qui évaluent la compétence lexicale. Cette compétence est testée par le choix de l'option correcte, c'est-à dire, de celle qui convient sémantiquement dans la phrase donnée. On y trouve les distributions suivantes :

- les verbes aux formes finies et à l'infinitif;

- les prépositions ;

- les adverbes de temps;

- les morphèmes lexicaux employés pour compléter les collocations (- *fail the exam, go for a walk, go sailing etc.)*

Le deuxième test de positionnement dans le recueil, couvrant les niveaux *pré-intermédiaire* et *intermédiaire*, évalue également les compétences grammaticales et lexicales. Les items qui testent la compétence grammaticale des candidats soumettent les catégories lexicales et grammaticales suivantes en tant qu'options: les verbes, les adjectifs, les adverbes, les pronoms, les déterminants, les prépositions, les particules adverbiales, les syntagmes adjectivaux et verbaux et les propositions. Contrairement au premier test dans le recueil, on constate la présence de syntagmes parmi les items de ce deuxième test.

En ce qui concerne les items contenant des verbes, on y trouve des auxiliaires, des verbes lexicaux ainsi que des modaux. Pour les auxiliaires, on constate les distributions suivantes :

-Les formes différentes de l'auxiliaire *do* qui varient selon le temps (*simple present vs. simple past)* et la personne;

-les formes des trois auxiliaires, *have*, *be* et *do*, au présent et au *simple past ;*

-l'auxiliaire *will* et l'expression *be going to*, aux formes déclarative et négative ;

-l'auxiliaire *will* aux formes déclarative et négative, précédé ou suivi d'un adverbe ;

-les reprises elliptiques contenant des verbes variant suivant la personne et le temps ;

Pour ce qui est des verbes lexicaux, on constate des distributions variées parmi les items, en l'occurrence:

- les formes non finies *(infinitif* et /ou *gérondif*) et/ou finies d'un même verbe associées à diverses constructions impliquant la personne, le temps, l'aspect (*simple vs. progressive*), la voix (active ou passive) et/ou le mode (indicatif s conditionnel) ;

- les formes verbales finies, variant en fonction du temps, de la personne, du mode (*indicatif* ou *conditionnel*) ;

- l'expression *used to* utilisée avec ou sans *did* et *was*, aux formes déclarative et interrogative ;

Les verbes modaux apparaissent également parmi les items, ainsi que *have to*, au *présent*, avec ou sans *do*.

Les items contenant les déterminants affichent les distributions suivantes :

-deux articles indéfinis, l'article défini et aucun article ;

-deux déterminants indéfinis et deux adjectifs, tous précédés par un adverbe ;

-l'article indéfini, l'article défini, le déterminant indéfini,

-deux déterminants interrogatifs et deux adjectifs, tous précédés par un adverbe ;

Quant à l'usage des noms, des adjectifs et des adverbes, on trouve les distributions suivantes :

-quatre formes d'un même adjectif, dont i) trois formes au comparatif et une forme au superlatif ; ii) une forme au comparatif et trois formes au positif, dont deux formes précédées par les adverbes ;

-trois formes d'un même adjectif, dont une au positif, une au comparatif et une au superlatif, ainsi que d'un nom ;

-un adjectif, le même nom au singulier et au pluriel, et le participe passé, ayant tous le même radical;

- un adjectif et trois noms, ayant le même radical, dont un nom inexistant;

- un même adjectif i) suivi par les prépositions ou par un pronom interrogatif ; ii) précédé ou suivi par les adverbes, ou précédé par la particule infinitive.

Les adverbes apparaissent non seulement en combinaison avec les adjectifs dans un item, mais également associé à un numéral et à la particule infinitive. Concernant les pronoms, on trouve quatre pronoms relatifs et quatre pronoms indéfinis à distribuer. Il n'y a pas d'items dans ce deuxième test qui affichent une combinaison des différents types de pronoms. En ce qui concerne les prépositions et les particules adverbiales, on constate l'occurrence des distributions suivantes parmi les items :

-quatre prépositions qui se réfèrent i) au verbe prépositionnel ; ii) à l'adjectif prépositionnel ; iii) au complément d'objet ;

- trois prépositions qui se réfèrent au complément circonstanciel et un adverbe ;

- quatre particules adverbiales qui conviennent au verbe précédent ;

Les propositions présentées dans les items se composent d'un sujet et d'un verbe qui varie soit en fonction du temps (*simple present, simple past*) soit en fonction de l'aspect, soit en fonction du temps et du mode (*indicatif et conditionnel*). Les syntagmes qui apparaissent dans ce test se composent d'un verbe qui varie soit selon le temps (*simple present, simple past, present perfect, past perfect*) soit selon le temps et la voix (*active ou passive*), ou encore selon la présence d'un adverbe.

Dans ce deuxième test constitutif du recueil *Success*, on trouve aussi des items évaluant la compétence lexicale. La réponse correcte consiste à choisir le lexème qui convient sémantiquement à une phrase donnée. Les lexèmes qui apparaissent appartiennent à l'une des quatre catégories suivantes: les adverbes, les verbes, les noms et les adjectifs. Dans certains items contenant des verbes il faut choisir celui qui fait partie d'une collocation, par exemple, le verbe to *pass* afin de former la collocation *to pass a law (item 94)*

Les trois tests étant conçus de manière identique dans le recueil *Success*, le troisième contient également des items évaluant la compétence grammaticale et la compétence lexicale. La compétence grammaticale est testée par le biais d'items qui consistent à compléter des phrases à trous par un représentant des catégories lexicales ou grammaticales suivantes : verbes, noms, adjectifs, adverbes, déterminants, pronoms, prépositions, particules adverbiales et propositions. On trouve un grand nombre d'items qui servent à évaluer la compétence verbale, affichant de nombreuses distributions de formes verbales:

> -les formes d'un même verbe à des aspects différents (simple ou progressif), précédées ou non des auxiliaires *do* ou *be* ;
>
> -les formes d'un même verbe à des aspects différents (*simple* et *progressif*) et à des temps variés i) *simple past, present perfect, et past perfect*, précédés ou non des auxiliaires *be* et *have*, ii) present, *will*-future, going to-future ; *would*+ infinitive ;
>
> -les formes d'un même verbe à des temps variés i) (*present, simple past, will-future, would+infinitive*) ; ii) (*simple past, past perfect, would + infinitif, would+infinitif passé* ;) iii) (*present, simple past, past perfect, would+ infinitive*) ; iv) (*present, simple past, present perfect, would+infinitive*) ; v) (*simple past, present perfect, past perfect, would+infinitive)* ;
>
> -les formes d'un même verbe à des voix différentes (active ou passive) et i) à des temps différents i) *(simple past, present perfect, past perfect)* ; ii) aux deux aspects (*simple* et *progressif)* et temps (*present, past*) ;
>
> -quatre verbes différents i) au présent suivis d'un adverbe de temps ; ii) à l'infinitif suivis d'un complément d'objet;

- quatre verbes différents i) au *simple past*, dont deux auxiliaires, un verbe proche d'un auxiliaire et un verbe modal; ii) au *simple past* qui forment les verbes à particules avec la particule adverbiale qui suit ; iii) au *gérondif* **;**

- i) quatre verbes modaux ; ii) quatre verbes modaux suivis par un *infinitif présent* ou un infinitif passé ; iii) deux verbes modaux, l'expression *had better*, et l'adjectif « best » ; iv) trois verbes modaux et l'expression « is going to » **;**

- i) les formes non finies d'un même verbe (*l'infinitif et le gérondif*) précédées ou non de la particule *to* ; ii) les formes *finies* et *non finies* d'un même verbe (*simple present, simple past* et *gerondif*);

- les différentes formes de l'expression « is going to » à des temps variés (*simple present, simple past*,) pourvue de la particule négative située aux différents lieux de la phrase **;**

En ce qui concerne les noms, on constate les distributions suivantes :

- le même nom employé au singulier, précédé d'un article défini, d'un article indéfini, sans article, ainsi qu'au pluriel ;

- deux noms différents, dont le premier est précédé d'un article défini et indéfini; et dont le deuxième est précédé d'un article défini ou sans article

Ce dispositif d'évaluation inclut un grand nombre d'items qui contiennent des morphèmes grammaticaux, en l'occurrence, des prépositions, des particules adverbiales, des conjonctions et des pronoms. On trouve deux types de prépositions réparties selon les items : celles qui se réfèrent au complément d'objet qui suit et celles qui se rapportent au verbe prépositionnel précédent. Il n'y a qu'une seule catégorie de conjonctions parmi les items : celles de subordination. On relève la présence de trois types de pronoms, à savoir :

- deux pronoms personnels, un pronom possessif et un pronom réfléchi,

- un pronom personnel, un pronom possessif et deux adverbes ;

- quatre pronoms indéfinis, dont deux suivis par les prépositions ;

Concernant les syntagmes, on en rencontre deux distributions, celles des syntagmes verbaux et des syntagmes nominaux. Chaque syntagme verbal est constitué d'un verbe complété soit d'un adverbe soit d'un adjectif attribut (s'il s'agit d'un verbe-copule). Les syntagmes nominaux sont composés soit d'un nom, d'un adjectif et d'un article, soit d'un nom, d'un numéral et d'un adverbe.

Pour conclure l'analyse des items servant à évaluer la compétence grammaticale, il faut noter la présence de nombreuses propositions. Celles-ci se composent de la manière suivante :

-d'un sujet et d'un verbe i) à des aspects différents *(simple* et *progressif)*; ii) à des aspects différents (*simple* et *progressif)* et à des temps différents (*present* et *present perfect); iii)* à des temps différents *(simple present, simple past,);*

- d'un pronom sujet à des personnes différentes et d'un verbe à des temps variés (*simple past, past perfect*) ;

- d'un sujet et de verbes différents dont i) un auxiliaire au *simple past*, un auxiliaire associé à un verbe lexical au *simple past* et un modal au conditionnel ; ii) de deux verbes à des temps différents (*simple present* et *simple past)*;

- d'un sujet et d'un verbe i) à des temps différents (*simple present, simple past*), précédés ou non d'une conjonction de subordination ; ii) d'un verbe au présent (simple) précédés ou non d'une conjonction de subordination ; iii) d'un verbe au passé (simple), contenant ou non des auxiliaires.

La variété de distribution des propositions montre l'importance de les comprendre et de savoir les utiliser à des niveaux intermédiaire et intermédiaire supérieur, car le nombre et la complexité des propositions sont beaucoup plus élevés dans ce test que dans les deux tests recouvrant les niveaux inférieurs. Cela est valable aussi pour les syntagmes, qui n'apparaissent pas du tout dans le premier test du recueil, mais la différence est encore plus marquante pour les propositions.

Les items construits autour de l'évaluation de la compétence lexicale présentent les catégories lexicales et grammaticales suivantes: noms, verbes, les adjectifs, particules adverbiales et conjonctions. Il est frappant de constater qu'à l'exception des conjonctions, il faut souvent choisir le morphème qui forme une expression figée avec le mot précédent ou le mot suivant de la phrase, par exemple, le verbe *to board*, qui entre dans la collocation to *board the plane (item 49).* Ce type de procédé restait isolé dans le deuxième test de ce recueil. Or dans ce troisième test, la méthode s'applique à un grand nombre d'items testant la compétence lexicale.

1.6.7 Le test du CAREL

1.6.7.1. Description et évaluation des caractéristiques du test

Contrairement aux tests présentés ci-dessus, le test de positionnement du CAREL est d'origine française. Il est conçu et administré par le Centre Audiovisuel de Royan pour l'Etude des Langues. Ce test est gratuit et accessible sur le site suivant : http://www.carel-royan.fr/testez-votre-niveau.html. Puisque le centre audiovisuel de Royan offre des formations intenses, immersives, à distance et mixtes, le but de ce test est d'évaluer le niveau linguistique des

candidats souhaitant suivre une de ces formations au centre. [34] Les langues proposées à l'évaluation sont l'anglais, l'allemand, l'espagnol et le français. Une contrainte temporelle est imposée car il est préconisé de réaliser ce test en 45 minutes maximum. Toutefois, il n'existe aucun contrôle. Ce test couvre sept niveaux de compétence dont six relèvent des niveaux distingués par le CERCL et portent les mêmes noms que dans ce référentiel : *Introductif-Intermédiaire-Seuil-Avancé-Autonome-Maitrise*. Le niveau inférieur, désigné *Vrai débutant*, est ajouté à six niveaux du CECRL, pour désigner le niveau de compétences au-dessous du niveau introductif A1, évoqué dans le CECRL.

Le test CAREL anglais est composé des trois étapes : parcours linguistique des candidats, auto-évaluation et test à proprement parler. Dans la rubrique « parcours linguistique » on demande aux candidats des renseignements sur leur apprentissage préalable de l'anglais, notamment la durée d'étude, les séjours éventuels dans un pays anglophone ainsi qu'une estimation a priori de leurs connaissances. Lors de la deuxième étape, centrée sur l'auto-évaluation, les candidats sont invités à évaluer dans quelle mesure ils sont capables d'effectuer les activités décrites dans les énoncés. Quatre icônes sont à leur disposition pour auto-évaluer cette capacité. Contrairement au CECRL, les compétences sont regroupées par deux lors de l'auto-évaluation. Bien que les compétences orales et écrites apparaissent sur le même écran, les énoncés portent sur une seule compétence à la fois. Lors de la troisième étape, huit tâches sont soumises aux candidats qui correspondent au niveau déterminé. Ces tâches, à effectuer dans l'ordre, évaluent trois compétences langagières communicatives : compréhension écrite, compréhension orale et expression écrite. S'y ajoutent deux compétences linguistiques, lexicale et grammaticale. Ces compétences sont évaluées de façon intégrée, c'est-à-dire que plusieurs compétences sont ciblées par la même tâche.

Le résultat est fourni sous forme de pourcentage de réponses correctes. Un niveau de compétence correspondant à ce pourcentage est attribué. Il faut noter que seules les activités autocorrectives servent de base au calcul du pourcentage et à l'attribution du niveau de référence car les activités d'expression écrite doivent être corrigées par le personnel du centre. L'attribution du niveau de compétences est présentée dans une seule phrase et manque de

transparence car le calcul du pourcentage de bonnes réponses n'est pas explicité. Aucun détail n'est donné sur les réponses, correctes et incorrectes. Le bilan final ne permet pas aux candidats de se faire une idée précise sur leurs compétences en langue, encore moins d'identifier des zones nécessitant un apprentissage renforcé. Cependant, à la fin du test, un questionnaire *objectifs et besoins* est proposé à ceux qui voudraient suivre une formation au CAREL. Ce questionnaire aide à définir les besoins et les objectifs poursuivis. Il constitue clairement un point fort du test. Le deuxième atout du test CAREL est la variété et l'intégration des activités permettant d'évaluer les compétences langagières. Les points faibles sont le manque de transparence dans l'attribution du résultat et la non-prise en compte des activités d'expression lors du calcul du score.

1.6.7.2 Analyse de la typologie des exercices et des compétences

Le test se compose de huit tâches qui évaluent les compétences de façon intégrée. La première tâche est un exercice de compréhension de l'oral, au format QCM, qui implique deux compétences: la compréhension de l'oral et la compréhension de l'écrit. Il faut comprendre un texte oral et établir les correspondances entre celui-ci et les quatre porte-paroles qui y apparaissent. La deuxième tâche est un exercice de lecture qui combine l'évaluation des quatre compétences, aussi bien communicatives que linguistiques. La compréhension de l'écrit est testée moyennant la lecture de l'article donné tandis que l'expression écrite est sollicitée par le biais des réponses aux questions ouvertes. La compétence lexicale est testée par le repérage de métaphores et d'allitérations dans le texte, ainsi que par la recherche de synonymes et d'antonymes. La compétence grammaticale est évaluée par le repérage des fautes de grammaire dans le texte. La troisième tâche est réalisée au travers d'un exercice de compréhension de l'oral, qui, outre cette compétence, évalue deux autres compétences communicatives. La compréhension écrite est testée par le choix d'un titre pertinent pour l'enregistrement. L'expression écrite est évaluée, comme dans la tâche précédente, par la formulation de réponses à des questions ouvertes. La quatrième tâche est un exercice de lecture qui vise deux habiletés : la compréhension de l'écrit et la compétence lexicale. La première de celles-ci est évaluée par la lecture d'un texte alors que la seconde est testée via la recherche des idiomes pertinents. La cinquième tâche est un exercice de

compréhension orale qui évalue non seulement cette compétence, mais également la compréhension écrite. Cette dernière est ciblée par le besoin d'indiquer pour chacune des phrases données si elle est vraie par rapport au texte enregistré. La sixième tâche consiste à lire une interview qui, de ce fait, évalue la compréhension de l'écrit. A part celle-ci, cette tâche cible la compétence lexicale car il faut compléter les trous dans le texte par les mots qui conviennent le mieux. La septième tâche consiste à écouter plusieurs enregistrements et à produire des énoncés libres adaptés à la description des textes sonores. Ces activités visent à évaluer la compréhension de l'oral et l'expression de l'écrit. La dernière tâche est centrée sur l'évaluation des deux compétences communicatives. La compréhension de l'écrit est testée par la lecture d'un texte auquel il faut répondre par la rédaction d'une lettre. Cette activité sert à évaluer l'expression de l'écrit.

1.6.8 Bilan de l'évaluation des tests

Comme on a pu le constater, il existe une grande variation d'adossement des tests aux niveaux de compétence définis par le CECRL. Certains tests sont étroitement adossés aux niveaux communs de référence, tandis que d'autres instruments d'évaluation en dévient de façon significative. Des tests comme *DIALANG*, mais aussi *Oxford Quick Placement Test, Oxford Placement Test 2 et CAREL* font partie de la première catégorie. En revanche, d'autres instruments d'évaluation ne correspondent pas aux niveaux de compétences du CECRL, soit parce qu'ils couvrent un spectre de niveaux plus limité que le *Cadre européen commun de référence pour les langues,* soit parce qu'ils ne distinguent pas les mêmes niveaux de compétence que le référent officiel. Les dispositifs comme le *Cutting Edge Placement Test* et le *Success Placement Test* illustrent ces deux cas de figure à la fois, car ils n'incluent pas les niveaux C1 et C2 d'un utilisateur expérimenté. De plus, ils ajoutent un niveau de compétence supplémentaire à deux niveaux intermédiaires distingués par le CECRL, désigné *pré-intermédiaire*. Comme celui-ci est situé entre les niveaux A 2 et B 1, il correspond au niveau A2+ du CECRL, mais contrairement à ce dernier, il a le même statut que les autres niveaux distingués dans les deux tests.

Pour conclure, force est de constater que l'usage des deux instruments d'analyse s'est avéré utile pour analyser de nombreux éléments des tests. La grille permet de repérer et de décrire les caractéristiques formelles et fonctionnelles des dispositifs. Le tableau, quant à lui, permet d'identifier les types d'exercices et de compétences évalués. De ce fait, ces deux instruments apportent une description complémentaire des caractéristiques des tests évalués.

Les qualités repérées varient d'un dispositif à l'autre dans une large mesure. En effet, les tests ont uniquement en commun la fonction d'être des tests de positionnement. Plusieurs dispositifs remplissent également une autre fonction, notamment diagnostique. Le choix de tests présentant une grande diversité de caractéristiques est intentionnel pour montrer le large répertoire des tests de positionnement disponibles sur le marché. La décision d'élaborer un test appartenant à cette catégorie n'a pas pour effet de déterminer les autres qualités de ce dispositif. Malgré la variété des formats des tests de positionnement et des fonctions remplies par ceux-ci, tous les dispositifs partagent le même objectif : placer les candidats dans le ou les groupes de niveaux qui correspond à leurs compétences langagières.[35] (Enquête Grenoble Universités : 3).

1.7 L'utilisation des tests de positionnement en anglais dans L'Enseignement Supérieur

Jusqu'à l'enquête entreprise par l'Université de Grenoble en 2009, il n'existait aucune étude sur l'utilisation des tests de positionnement en anglais dans l'enseignement supérieur (Enquête Grenoble Universités 2009 : 3). L'usage des tests de positionnement, aussi bien en matière de tests utilisés que d'objectifs poursuivis, variait au sein d'une même université (ibid. :3). De ce constat est née la volonté d'en apprendre davantage sur les pratiques menées par les établissements d'enseignement supérieur en France en termes de positionnement des étudiants en anglais. Cette volonté s'est concrétisée en objectif de recension des tests de positionnement en anglais effectivement employés. Les initiateurs de l'étude espéraient que ce recensement constituerait

une aide au choix du test de positionnement le mieux adapté aux besoins des étudiants spécialistes d'autres disciplines en anglais (ibid. : 3).

Pour atteindre cet objectif, une enquête auprès des 72 établissements a été organisée. Dans le cadre de cette enquête, un questionnaire contenant 31 questions a été conçu et envoyé aux institutions concernées par le positionnement des étudiants en anglais (ibid. : 3). Les questions étaient, en partie, au format QCM et, en partie, ouvertes. Elles portaient sur les différents aspects liés aux tests en anglais, par exemple, sur le format du test et les résultats, les compétences évaluées, l'usage du test, le nombre de sessions organisées chaque année et le public concerné (ibid. : 3). Sur les 72 établissements contactés, 32 ont complété le questionnaire, ce qui constitue un retour de 44, 4 %. Environ deux tiers des institutions participant à l'enquête ont indiqué utiliser un test de positionnement, à savoir, 21 établissements. Ceux qui n'emploient pas de test ont fourni des raisons différentes dont la plus fréquente était de ne pas avoir de « nécessité de positionner les étudiants » (ibid.: 8). Seul un établissement a justifié la non-utilisation d'un test de positionnement par le « manque des moyens financiers » (ibid.: 8). Bien que les tests utilisés par les établissements soient divers, on reconnaît tout de même une régularité. L'*Oxford Quick Placement test* est le plus fréquemment employé, suivi de *DIALANG*. Plus de la moitié des institutions utilise des dispositifs non référencés dans le questionnaire, le plus souvent un « test maison » (ibid.:8).

En ce qui concerne le format des tests, la majorité des établissements, approximativement 62%, ont recours à un test informatique et 19% à un test mixte, combinant ce format avec un support papier. Ces chiffres démontrent l'usage très important des technologies numériques dans le domaine d'évaluation des compétences sur le terrain (ibid. : 11). Quant aux habiletés évaluées dans les tests, les compétences de réception sont ciblées beaucoup plus souvent que les compétences de production. La compréhension orale et écrite des candidats est testée avec la même fréquence dans un tiers des institutions (ibid. : 11). Par rapport à ce chiffre, les deux compétences de production sont évaluées beaucoup moins lors du positionnement des candidats, mais il existe un décalage considérable entre la production écrite et orale. Tandis que la production écrite est testée par environ 17% d'établissements, la production orale

est ciblée dans environ 6% de cas. L'interaction souffre également d'un taux d'évaluation faible. Ce résultat est lié á l'usage des technologies numériques qui favorisent le format QCM et, à l'heure actuelle, ont toujours des limites dans l'évaluation des tâches construites (Douglas 2010 :117). Bien que la recherche souligne que cet inconvénient puisse un jour être maîtrisé, l'utilisation des technologies numériques a eu pour effet de raréfier le format de la réponse construite (Douglas 2010 : 117). Or, l'emploi de ce format est plus naturel lors de l'évaluation des compétences productives, tant écrite qu'orale. Toutefois, il est possible de tester l'expression écrite en recourant aux tâches de format QCM. Dans le test POSILANG, l'expression écrite est également testée par des tâches au format QCM. En revanche, ce dernier ne convient pas du tout pour évaluer l'expression orale. Ces explications démontrent un lien direct entre l'usage d'un test informatique par la majorité d'institutions et le faible taux d'évaluation des compétences productives, notamment de la production orale, ainsi que de l'interaction. En revanche, il n´existe pas de corrélation entre le nombre de candidats passant un test et les compétences visées dans celui-ci (ibid: 11). Le nombre de participants aux tests de positionnement varie largement selon l'établissement, entre 40 et 5000. Toutefois, la production orale est uniquement évaluée par un seul établissement parmi les cinq dont le nombre de candidats dépasse 1500 personnes. L'interaction orale est également testée dans un seul établissement avec un grand nombre d'étudiants (ibid. : 11).

Quant à l'usage d'un test de positionnement, il varie en fonction des établissements. Ce type de test est utilisé pour former des groupes de niveau dans 34% de cas (ibid. : 12). Dans d'autres universités, le test de positionnement sert à des usages divers : fournir des informations pour l'apprenant ou pour l'institution, plus rarement servir de base pour une certification ou pour valider un niveau. Dans 3% des universités, les résultats à ce test sont utilisés pour la recherche (ibid. : 29)

De même, le format des résultats varie selon les universités. Ceux-ci sont donnés sous la forme d'un score ou d'un pourcentage de réussite par quasiment la moitié des établissements. Le reste des institutions fournit des résultats adossés au CECRL. Ils sont calculés pour chaque compétence par presque un quart d'institutions, tandis que 18 % informent les candidats de leur niveau global

uniquement. Le niveau global et les niveaux pour chaque compétence à la fois sont indiqués par environ 12% des universités (ibid. 14). L'adossement des résultats au CECRL par plus de la moitié des universités n'est pas étonnant, car il reflète l'impact important de ce référentiel sur l'évaluation en langues (Westhoff 2007 :676).

Cette étude menée par le département LANSAD de l'Université de Grenoble révèle deux choses importantes. Premièrement, l'usage des tests de positionnement préoccupe la grande majorité des universités françaises (Enquête Grenoble Universités : 3). Deuxièmement, les établissements cherchent à collaborer dans ce domaine avec d'autres universités. Ceci montre que le positionnement des candidats à l'aide des tests actuellement disponibles ne les satisfait pas complètement (ibid. : 3).

Notes:

[1] Ces critères seront présentés plus bas dans le chapitre.

[2] La définition de Carrol peut être traduite de la manière suivante : « Un test psychologique ou éducatif est une procédure conçue pour déduire un certain comportement à partir duquel des inférences peuvent être opérées concernant certaines caractéristiques d'un individu » (Carrol 168: 46).

[3] Le CECRL établit une distinction entre les différents types des compétences. Les deux grandes catégories distinguées sont les activités de communication langagière d'une part et les compétences linguistiques d'autre part.

[4] Ces tests sont accessibles sur le site suivant :

a) *tolearnenglish sur http:/tolearnenglish.com/test-de-niveau-anglais-grammaire.php),*

b) *e-anglais sur http://www.e-anglais.com/tests/index.html*

c) *ifg langues sur http://www.ifglangues.net*

[5] Voir la grille d'évaluation. La grille constitue un recueil de paramètres choisis selon des critères pertinents pour décrire et évaluer divers tests de positionnement présentés dans la suite de ce chapitre. Il existe d'autres grilles d'évaluation contenant des critères de description variables, dont deux seront décrites dans les sous-chapitres suivants. La grille présentée constitue un modèle parmi d'autres utilisés pour analyser et comparer les tests.

[6] Ce modèle des spécifications sera présenté dans le troisième chapitre, consacré à la conception du test de positionnement POSILANG.

[7] Cette deuxième composante des méthodes est désignée par « mode d'évaluation » dans notre grille.

[8] Le mode d'évaluation de DIALANG et son impact sur la transparence de l'évaluation sera expliqué dans la partie consacrée à ce système d'évaluation, dans le deuxième chapitre.

[9] Voir la note d'orientation, rédigée dans le cadre du Projet Innova-Langues (document de présentation, 2011). INNOVA est l'acronyme de : «INNOVAtion et transformation des pratiques de l'enseignement-apprentissage des langues dans l'enseignement supérieur ».

[10] Ces qualités sont appelées « principes » par certains linguistes, par exemple, Brown (2010 :25), qui en distingue cinq.

[11] *Standards for educational and psychological testing* constituent le code officiel de la pratique professionnelle aux Etats-Unis. Depuis les années 1950, ce document est publié toutes les dix ans environ par les trois associations professionnelles suivantes : *American Educational Research Association*, *American Psychological Association* et *The National Council on Measurement in Education* (Chapelle 1999 : 265).

[12] *Affect* renvoie au degré d'anxiété excessif causé par une procédure d'évaluation (Madsen 1983 : 179).

[13] L'évaluation directe implique des tâches ouvertes. Celles-ci permettent la mise en œuvre d'habiletés complexes, qui ne sont pas limitées par des formats de réponse restrictifs.

[14] La position de Messick se distingue de celle d'autres spécialistes qui considèrent cette forme de validité comme séparée. Les dénominations de celle-ci incluent « consequential validity » (Brown 2010 : 34) ou « washback validity » (Morrow 1986 :5). Ces dénominations soulignent non seulement le caractère spécifique de cette validité mais incorporent une dimension causale.

[15] Tandis que l'influence des pratiques d'évaluation des compétences sur l'enseignement et l'apprentissage est connue sous le terme « washback » dans la linguistique appliquée, ce phénomène est résumé par le terme « backwash » dans le domaine éducatif général. Alderson & Wall (1993) notent qu'il n'y a pas de raisons de donner préférence à l'un de ces deux termes (Alderson & Wall 1993 : 115).

[16] Alors que Fulcher & Davidson (2007) considèrent la validité en tant que **propriété** liée à l'**usage** d'un test, Chapelle (1999) établit une distinction entre les deux conceptions de cette notion (1999 : 258). La conception de validité en tant que **propriété** d'un test est ancienne tandis que, selon la conception moderne, la validité est considérée en tant qu'**argument concernant l'interprétation et l'usage d'un test**. Il n'est pas important pour la compréhension de l'argumentation exposée ici que Fulcher & Davidson (2007) ne fassent pas la distinction entre les conceptions ancienne et moderne de validité dans son argumentation.

[17] Le rapport établi entre validité et authenticité est très ancien. L'interprétation du lien unissant ces deux notions est partagée par les chercheurs à l'heure actuelle. La grande majorité des linguistes assimile ces deux concepts l'un à l'autre, comme évoqué ci-dessus (Chapelle 1999 : 256).

[18] http://www.iltaonline.com/

[19] http://www.iltaonline.com/index.php?option=com_content & view=article & id=57 & Itemid=47

[20] http://www.iltaonline.com/index.php?option=com_content & view=article & id=122 & Itemid=133

[21] http://www.iltaonline.com/index.php?option=com_content & view=article & id=122 & Itemid=133

[22] La conception de validité, mise en évidence par ILTA dans la ligne directrice citée, est traditionnelle car elle est considérée comme une caractéristique inhérente au test. Cette conception de la validité a été suivie par une conception de la validité en tant qu'argument concernant les inférences tirées à partir des scores : dans quelle mesure les interprétations et les usages des tests sont-ils justifiés ? (Chapelle 1999 : 258). Une conception plus récente de la validité est à la base de la paraphrase qui suit la ligne citée. Pour cette raison il y est question des inférences valables et non pas d'un test valable (Chapelle 1999 : 258).

[23] http://www.iltaonline.com/index.php?option=com_content & view=article & id=122 & Itemid=133

[24] La réponse correcte s'appelle la clé, tandis que les autres options sont appelées des distracteurs car elles détournent temporairement du bon choix (Purpura 2001: 1).

[25] Il y a une seule exception à ce constat, l'item 4, où se trouvent deux phrases sémantiquement liées : « CLOSED FOR HOLIDAYS. Lessons start again on 8th January ».

[26] On établit une distinction entre les catégories lexicales et grammaticales. Les premières ont un contenu sémantique tandis que les secondes constituent un apport de type fonctionnel. Les noms, les verbes, les adjectifs et les adverbes désignent des catégories lexicales (ou parties du discours) tandis que les déterminants, les conjonctions et les auxiliaires désignent des catégories grammaticales (Crowgey 2012 : 5).

[27] Il existe plusieurs termes synonymes pour désigner le sens grammatical qui sont utilisés par d'autres linguistes. Ainsi, Jaszczolt (2002) désigne ce dernier comme le sens sémantique et le sens de l'énoncé. Grice (1975), par exemple, le nomme le sens littéral, le sens de la phrase ou le sens conventionnel. Le terme « sens grammatical » sera utilisé par la suite pour rester conforme au modèle de la connaissance grammaticale de Purpura (1961).

[28] Les phonèmes sont non seulement les unités sonores de la langue, mais ce sont également les unités les plus petites sémantiquement distinctives.

[29] Ce test est accessible sur le site suivant :

http://www.pearson.pl/pub/angielski/uploaddocs/placements_tests/Energy_Placement_Test.pdf.

[30] Un syntagme est généralement un groupe de mots, ordonné autour d'un noyau, qui constitue une unité sémantique ou fonctionnelle.

[31] Les items de deuxième type apparaissent dans les tests D et E tandis que les tests A-C comportent les items de premier type.

[32] L'hyponymie est une relation d'inclusion entre deux mots dont l'un représente une classe incluse dans l'autre. http://www.cnrtl.fr/definition/hyponymie

[33] Le test est accessible gratuitement sur le site suivant : http://www.scribd.com/doc/54165945/Placement-Tests

[34] La formation mixte est également désignée par le terme *Blended Learning*. Ce type de formation hybride englobe la formation en présentiel et à distance à la fois.

[35] Selon le rapport sur l'état des lieux dans l'utilisation des tests de positionnement en anglais dans l'enseignement supérieur, le terme « test de positionnement » n'apparaît pas comme clairement défini (Enquête Grenoble Universités : 3).

2 Adosser un test au Cadre Européen Commun de Référence pour les Langues

2.1 Enjeux

Les gouvernements européens des pays occidentaux ont compris dans les années cinquante que les nouveaux développements sociétaux tels que la globalisation, la migration ainsi que le multilinguisme et le multiculturalisme imposaient la résolution d'enjeux éducatifs complexes (Byrnes 2007 : 641). La conception et la mise en place d'une politique efficace d'apprentissage langagier était l'un de ces enjeux (Byrnes 2007 : 641). Cette politique avait pour but de rendre les personnes concernées capables de participer au système éducatif d'un pays. Il s'agissait d'un enjeu poursuivi par le conseil de l'Europe, alors récemment établi.

Depuis sa fondation en mai 1949, le Conseil de l'Europe a décidé d'adopter une politique linguistique commune, bien que cette organisation supranationale se compose, dès l'origine, de dix pays se distinguant par leur langue et leur culture[1] (Byrnes 2007 : 641). Depuis le début des années 1970, cette institution a particulièrement encouragé et soutenu l'apprentissage des langues vivantes afin de promouvoir la communication et les échanges entre les États membres du Conseil de l'Europe (Little 2007 :646). L'engagement de cette organisation dans le domaine des langues vivantes a donné lieu à l'établissement de deux principes. Il s'agit, d'abord de la nécessité d'analyser les besoins communicatifs des apprenants et ensuite de décrire les répertoires langagiers relatifs aux besoins communicatifs respectifs (Little 2007 :646).

La politique linguistique adoptée par le Conseil de l'Europe est celle du plurilinguisme, qui non seulement est définie comme un choix de la politique linguistique mais aussi comme une aide apportée à chaque individu d'un État membre lui permettant de construire son identité linguistique et culturelle (Conseil de l'Europe 2005 : 105). En outre, en ce qui concerne la construction de l'identité linguistique, le Conseil de l'Europe s'est fixé pour but de développer «une compétence plurilingue » des apprenants (Conseil de l'Europe 2005 : 11). Celle-ci implique que les apprenants disposent d'un « répertoire langagier » (Conseil

de l'Europe 2005 : 11) engageant toutes les capacités linguistiques des individus (Conseil de l'Europe 2005 : 11). Il faut noter que le plurilinguisme est considéré comme une compétence complexe qui n'englobe pas de compétences superposées ni juxtaposées : « On considérera qu'il n'y a pas là superposition ou juxtaposition de compétences distinctes, mais bien existence d'une compétence complexe, voire composite, dans laquelle l'utilisateur peut puiser » (Conseil de l'Europe 2005 : 128). Dans le modèle du plurilinguisme, les différentes langues et cultures peuvent être maîtrisées à des degrés variés par un individu : « […] la compétence à communiquer langagièrement et à interagir culturellement d'un acteur social qui possède, à des degrés divers, la maîtrise de plusieurs langues et l'expérience de plusieurs cultures » (Conseil de l'Europe 2005 : 129). Dans ce modèle, les différentes langues maîtrisées par une personne interagissent et sont sollicitées selon la situation donnée (Conseil de l'Europe 2005 : 11). Il faut distinguer entre le plurilinguisme et le multilinguisme, le terme plus commun, en mettant en avant la valeur communicative de toutes les langues maîtrisées par un individu, ce qui accroit la compétence communicative de la personne concernée dans le milieu plurilingue de l'Europe (Byrnes 2007 : 642).

La politique de plurilinguisme menée par le Conseil de l'Europe pendant plusieurs décennies a donné lieu à l'élaboration et à la publication du *Cadre Européen Commun de Référence pour les Langues*, édité en 2001 (Byrnes 2007 : 641).[2] L'importance accordée par ce document au plurilinguisme constitue l'une de ses propriétés remarquables (Byrnes 2007 : 642). Conformément à cette politique linguistique, ce référentiel se propose :

> -de promouvoir et faciliter la coopération entre les établissements d'enseignement de différents pays
>
> -d'asseoir sur une bonne base la reconnaissance réciproque des qualifications en langues
>
> -d'aider les apprenants, les enseignants, les concepteurs de cours, les organismes de certifications et les administrateurs de l'enseignement à situer et à coordonner leurs efforts (Conseil de l'Europe 2005 : 12).

Ces trois buts différents poursuivis par le *Cadre Européen Commun de Référence pour les Langues* reposent sur l'idée commune de coopération internationale entre les professionnels des langues vivantes ainsi qu'entre les institutions d'enseignement des langues (Conseil de l'Europe 2005 : 9). La coopération sert notamment à lutter contre les problèmes de communication

rencontrés dans le domaine de l'enseignement des langues vivantes, imputables aux différences entre les systèmes éducatifs des différents pays en Europe (Conseil de l'Europe 2005 : 9). Afin de résoudre ces difficultés de communication, le CECRL fournit « une base commune pour l'élaboration de programmes de langues vivantes, de référentiels, d'examens, de manuels, etc. en Europe » destinée aux spécialistes des langues vivantes (Conseil de l'Europe 2005 : 9). La base commune contient les critères objectifs qui servent à décrire la compétence langagière et rend donc possible la description explicite d'objectifs, de contenus et de méthodes d'enseignement et de qualifications. La mise à disposition d'une base commune est censée augmenter la transparence des programmes d'enseignement, des cours et des qualifications, dont on espère un impact positif sur la coopération entre les États européens (Conseil de l'Europe 2005 : 9).

Alors que les trois buts poursuivis par le CECRL reposent sur une idée commune, chacun éclaire un aspect différent de la coopération internationale. Tandis que le premier objectif cible la coopération entre les établissements d'enseignement des langues vivantes, le deuxième vise à contribuer à la coopération par la mise en place d'une reconnaissance réciproque des qualifications en langues. Enfin, le troisième objectif examine la dimension humaine de l'idée de coopération en soulignant la nécessité pour tous les acteurs d'enseignement et d'apprentissage des langues vivantes de parvenir à coordonner leurs efforts (Conseil de l'Europe 2005 : 9).

Les buts déclarés dans le *Cadre Européen Commun de Référence pour les Langues* reflètent les trois principes adoptés par le Conseil de la Coopération Culturelle du Conseil de l'Europe. Cette organisation, qui est une composante du Conseil de l'Europe, est chargée de mener des activités dans le domaine de l'éducation et de la culture (Van Ek 2001 : ii). Comme le CECRL, le Conseil de la Coopération Culturelle met en avant l'objectif de coopération entre les pays européens :

> […] les États membres, en adoptant ou en développant une politique nationale dans le domaine de l'enseignement et de l'apprentissage des langues vivantes, pourraient parvenir à une plus grande concertation au niveau européen grâce à des dispositions

ayant pour objet une coopération suivie entre eux et une coordination constante de leurs politiques. [3]

L'objectif de coopération évoqué dans ce principe concerne les institutions politiques des pays puisqu'il n'est question ni de la coopération entre les établissements, ni des certifications, ni des administrateurs, ni des acteurs de l'enseignement. Il n'est donc pas identique au but de coopération entre les systèmes éducatifs européens poursuivi par le CECRL. Néanmoins, le deuxième principe, inclus dans la Recommandation R (82) n°18 du Comité des Ministres du Conseil de l'Europe, postule que la «meilleure connaissance des langues vivantes européennes » est la seule possibilité susceptible de « faciliter la communication et les échanges entre Européens de langues maternelles différentes » et de « favoriser la compréhension réciproque et la coopération en Europe ».[4] L'apprentissage – puis l'usage des langues étrangères – ne sont pas considérés comme des buts en soi par le Conseil de l'Europe, mais comme des moyens censés permettre aux Européens de surmonter les barrières linguistiques et culturelles afin de se comprendre réciproquement, d'améliorer leur accès à l'information et leur mobilité personnelle (Van Ek 2001 : ii). Les objectifs évoqués reflètent l'intention sociale envisagée par cette institution qui consiste à rapprocher les peuples européens en leur faisant prendre conscience de l'identité européenne commune et de ses valeurs (Van Ek 2001 : ii).

Les objectifs sociétaux et individuels évoqués expliquent les activités menées par la Coopération Culturelle du Conseil de l'Europe qui consistent à promouvoir la mise en œuvre des réformes en matière d'enseignement et d'apprentissage des langues vivantes et d'évaluation des compétences (Van Ek 2001 : ii). Le but principal des activités mises en place par cette institution est de développer une éducation qui satisfasse aux besoins des sociétés européennes à l'heure actuelle aussi bien qu'aux besoins et aux motivations des apprenants individuels au sein des sociétés contemporaines (Van Ek 2001 : ii).

Le CECRL a une grande influence sur la promotion du plurilinguisme en Europe, dans les domaines de l'élaboration du curriculum aussi bien que dans la conception et l'évaluation des examens dans un grand nombre de pays européens. Les personnes concernées par l'apprentissage et l'enseignement des

langues étrangères reconnaissent l'utilité de ce document, de même que les personnes impliquées dans l'organisation de ce processus et ses bénéficiaires, en l'occurrence, les employeurs (Hulstijn 2007 : 1). Les chercheurs s'accordent sur le fait que, grâce à l'élaboration du CECRL, un discours commun transcendant les barrières culturelles a été créé pour la première fois. Ce discours commun sert à débattre des enjeux centraux dans l'apprentissage, l'enseignement et l'évaluation en langues, et ce en dépit de différences culturelles considérables (Byrnes 2007 : 642).

2.1.1 La perspective philosophique du Cadre Européen Commun de Référence

Le Cadre Européen Commun de Référence pour les Langues est un document détaché du contexte, mais n'est pas indifférent à celui-ci. En effet, ce document doit être transférable dans différents contextes et être "traduit" selon la situation donnée dans une forme pertinente au contexte afin d'être implémenté (North 2007 : 656). Cette idée est exprimée dans le CECRL lui-même en termes plus concrets :

> Chacun des constituants majeurs et chacune des composantes particulières du modèle proposé peut, si il ou elle est retenu(e) comme objectif privilégié de l'apprentissage, entraîner des choix variés de contenus, de démarches, de moyens pour mener à bien cet apprentissage. (Conseil de l'Europe 2005 : 130).

Selon Byrnes, deux versions du CECRL, celle détachée du contexte et celle spécifique à ce dernier, sont nécessaires pour prouver la validité revendiquée par ce document (Byrnes 2007 : 642). De même que le CECRL n'est pas propre à un contexte, il n'est pas attaché à une langue particulière. Ceci implique qu'il décrit les fonctions communicatives que les apprenants doivent maîtriser à différents niveaux de compétence sans spécifier comment ces fonctions pourraient être réalisées dans une langue particulière, par exemple, en anglais ou en français (Little 2007 : 646). L'indication des fonctions communicatives repose sur l'hypothèse, admise par le CECRL, que toute tâche communicative demande le même niveau de compétence, indépendamment de la langue en question (Little 2007 : 646).

La perspective favorisée par ce document est actionnelle. Celle-ci constitue la deuxième qualité remarquable du CECRL, en plus du plurilinguisme (Byrnes 2007 : 642). Cette approche est remarquable en raison de sa nouveauté (Byrnes 2007 : 642). Elle constitue l'heuristique derrière le schéma descriptif du CECRL. En effet, cette perspective est définie de la façon suivante :

> Elle considère avant tout l'usager et l'apprenant d'une langue comme des acteurs sociaux ayant à accomplir des tâches (qui ne sont pas seulement langagières) dans des circonstances et un environnement donnés, à l'intérieur d'un domaine d'action particulier. Si les actes de parole se réalisent dans des activités langagières, celles-ci s'inscrivent elles-mêmes à l'intérieur d'actions en contexte social qui seules leur donnent leur pleine signification. Il y a «tâche» dans la mesure où l'action est le fait d'un (ou de plusieurs) sujet(s) qui y mobilise(nt) stratégiquement les compétences dont il(s) dispose(nt) en vue de parvenir à un résultat déterminé (Conseil de l'Europe 2005 : 15).

La perspective actionnelle se focalise ainsi sur l'accomplissement de tâches qui sont conceptualisées comme « actions en contexte social » (p.15). Celles-ci englobent les activités langagières aussi bien que non-langagières (Tagliante 2005 : 36). Les apprenants accomplissent des tâches et non pas des activités langagières pures, car ces dernières sont toujours situées dans « des circonstances et un environnement donnés » (Conseil de l'Europe 2005 : 15). De ce fait, les activités langagières sont sémantiquement déterminées par des « actions en contexte social » (Tagliante 2005 : 36). Puisqu'ils réalisent des tâches et non uniquement des activités langagières, les apprenants d'une langue sont considérés comme des « acteurs sociaux » (Conseil de l'Europe 2005 : 15). Dans cette perspective, le développement des stratégies verbales et non-verbales joue un rôle primordial, car celles-ci doivent être mobilisées pour mettre en œuvre les tâches sociales (Conseil de l'Europe 2005 : 15).

Une action en contexte social peut être considérée comme une tâche seulement si l'action poursuit un objectif bien déterminé, qu'il soit personnel ou établi par la situation d'apprentissage (Goullier 2005 : 21). Cependant, il ne suffit pas qu'un objectif soit fixé et poursuivi au cours d'une action en contexte social; deux autres conditions doivent également être remplies (Goullier 2005 : 21). En effet, l'objectif doit être clairement perçu par les apprenants ou les usagers de la langue et l'action doit donner lieu à « un résultat déterminé » (Conseil de l'Europe 2005 : 15). Cette caractéristique de la tâche est une composante essentielle de sa définition donnée dans le *Cadre européen commun de référence* car : « Est

définie comme tâche toute visée actionnelle que l'acteur se représente comme devant parvenir à un résultat donné en fonction d'un problème à résoudre, d'une obligation à remplir, d'un but qu'on s'est fixé » (Conseil de l'Europe 2005 : 15).

Il faut distinguer entre deux catégories de tâches : les tâches complexes d'une part et les tâches communicatives d'autre part (Bourguignon 2010 : 27). Les tâches complexes intègrent plusieurs activités langagières (Bourguignon 2010 : 27). En revanche, les tâches communicatives font appel à une seule activité langagière et, par conséquent, sont également appelées « tâches simples » (Bourguignon 2010 : 27). La lecture d'un texte, par exemple, est une tâche simple parce que sa réalisation fait appel à une seule activité langagière, à savoir, la compréhension de l'écrit (Bourguignon 2010 : 27). Une tâche complexe serait, par exemple, la réservation d'un séjour de vacances car elle exige l'usage intégré de plusieurs activités de communication langagière, au moins celles de compréhension de l'écrit et de production de l'écrit. À côté des activités langagières, les capacités non langagières sont nécessaires afin de remplir l'objectif de cette tâche (Bourguignon 2010 : 27). Les deux catégories de tâches ont en commun la poursuite d'un but non langagier (Bourguignon 2010 : 27). Une tâche simple, par exemple, la lecture d'un texte, est accomplie par rapport à un objectif non langagier, et ce, même s'il s'agit d'une lecture pour le plaisir (Bourguignon 2010 : 35).

Les tâches définies par les descripteurs du CECRL se déclinent en de nombreuses catégories car elles varient en fonction de plusieurs facteurs. Il s'agit des domaines dans lesquels les tâches sont situées, de leur nature, de leur complexité, des stratégies mises en œuvre, de leur type, des activités langagières déployées ainsi que de l'évaluation, liée à la tâche donnée (Tagliante 2005 : 38). Chacun de ces six niveaux de compétences, exposés dans le CECRL, se distingue par la quantité et la qualité d'actes de communication à maîtriser à ce niveau particulier (Tagliante 2005 : 38). Le nombre et la qualité des actes de communication sont déterminés par le nombre de tâches que l'apprenant peut accomplir correctement sur les plans linguistique et pragmatique (Tagliante 2005 : 38).

2.1.1.1 Comparaison entre la perspective actionnelle et l'approche communicative

La perspective actionnelle est voisine de l'approche communicative. L'une et l'autre partagent un certain nombre de caractéristiques communes (Tagliante 2005 : 36). Premièrement, elles se focalisent sur la communication entre les personnes. Deuxièmement, elles accordent à l'apprenant un rôle primordial dans le processus d'apprentissage en le rendant autonome et responsable de son progrès (Tagliante 2005 : 36). En complément de ces concepts communs, l'approche actionnelle intègre la notion de « tâche » à effectuer au sein des multiples contextes sociaux dans lesquels un apprenant devra agir (Tagliante 2005 : 36). Comme nous l'avons évoqué, selon les principes du CECRL, les activités langagières doivent être situées dans le cadre d'actions sociales pour avoir du sens (Conseil de l'Europe 2005 : 15). À cet effet, les activités langagières doivent également être déployées par rapport à un objectif non-langagier (Conseil de l'Europe 2005 :15, Bourguignon 2010 : 35). Ceci concerne aussi bien l'usage de la langue que son apprentissage.

La perspective actionnelle demande la mise en œuvre d'un cadre d'apprentissage qui place l'apprenant dans une situation lui permettant de réaliser des activités langagières dans la poursuite d'un objectif qui est non-langagier (Bourguignon 2010 : 35). La démarche d'apprentissage acquiert du sens grâce au lien entre le développement des capacités à communiquer et le but visé qui est clairement identifié (Bourguignon 2010 : 35). Ce cadre d'apprentissage est appelé « l'unité d'action » (Bourguignon 2010 : 35). Il s'agit donc d'une démarche qui met l'apprenant en action au moyen de l'accomplissement de tâches (Bourguignon 2010 : 38). Cette démarche doit inciter l'apprenant à faire face aux contraintes contenues dans les tâches pour l'obliger à choisir parmi ses connaissances et ses capacités celles qui sont pertinentes par rapport aux contraintes perçues (Bourguignon 2010 : 38). Puisque toutes les tâches autres que purement communicatives contiennent des contraintes, il est essentiel de les prendre en compte pour réussir une tâche. Selon Bourguignon (2010 : 36), les contraintes contenues dans les tâches transforment l'approche communicative en une approche « communic'actionnelle », qui associe donc les approches communicative et actionnelle. De même que dans l'approche communicative, « le développement

des capacités à communiquer est toujours un objectif majeur » dans l'approche communic'actionnelle, cet objectif est poursuivi « à travers les tâches communicatives qui sont au service de la réussite d'une tâche qui met l'apprenant en action » (Bourguignon 2010 : 38). Cette définition renvoie à la distinction entre les tâches communicatives et les tâches complexes mentionnée ci-dessus (Bourguignon 2010 : 38). Il est important de prendre conscience du fait que, dans la perspective actionnelle, les tâches communicatives sont loin d'être secondaires bien qu'elles soient au service de la tâche complexe. L'importance des tâches communicatives dans l'approche communic'actionnelle s'explique par l'objectif commun aux deux approches, qui consiste à développer les capacités de communication des apprenants (Bourguignon 2010 : 38). Ceci n'empêche pas que la tâche complexe soit également essentielle dans cette approche puisqu'elle sert à développer la compétence langagière. La comparaison entre les approches communicative et communic'actionnelle montre que le paradigme actionnel en langues n'implique pas une rupture avec les pratiques antérieures, mais autorise une évolution des usages pédagogiques mis en place par l'approche communicative (Bourguignon 2010 : 101). Dans les deux approches, la connaissance de la langue est au service du développement des capacités à communiquer, mais le but ultime de l'apprentissage a évolué. Alors que dans l'approche communicative, la finalité de l'apprentissage était la communication, dans l'approche communic'actionnelle la finalité est l'action menée dans le cadre de l'accomplissement des tâches qui ne sont pas seulement langagières (Bourguignon 2010 : 101).

2.1.1.2 L'évaluation des compétences dans la perspective actionnelle

L'approche communic'ationnelle ne doit pas se limiter aux choix des objectifs et des méthodes d'enseignement et d'apprentissage, mais également avoir de l'influence sur le domaine d'évaluation des compétences langagières (Tagliante 2005 : 36). La perspective actionnelle demande la conception et la mise en œuvre de tests qui permettent de réaliser des tâches « dans des circonstances et un environnement donnés, à l'intérieur d'un domaine d'action particulier » (Conseil de l'Europe 2005 : 36). Cette exigence implique que les tâches d'évaluation incluent les informations sur les acteurs, les lieux, les événements et

les objets créant les circonstances dans lesquelles la tâche se situe (Tagliante 2005 : 36).

La nécessite de fournir ces informations lors de la réalisation des tâches dans un test peut donner l'impression que les tâches à accomplir en contexte réel sont identiques à celles à effectuer dans une situation d'évaluation. Or ce n'est pas le cas car les deux types de tâches se distinguent par plusieurs caractéristiques. En ce qui concerne les tâches en situation d'évaluation, leur première différence est l'impossibilité pour l'apprenant de choisir la tâche à effectuer étant donné que cette fonction revient à l'enseignant (Tagliante 2005 : 36). La deuxième particularité est la conscience du fait que la performance de la tâche sera évaluée. Troisièmement, les apprenants savent que l'objectif n'est pas un simple accomplissement de la tâche mais également une performance linguistiquement correcte (Tagliante 2005 : 36). Le quatrième aspect qui distingue les tâches d'évaluation est la conscience des candidats de leur manque de « caractère vital ». Cette caractéristique implique qu'il est possible de réessayer en cas d'erreur (Tagliante 2005 : 36). La dernière qualité propre à ce type de tâches est la possibilité de se faire aider dans beaucoup de cas, soit par l'enseignant soit par un autre apprenant (Tagliante 2005 : 36).

2.1.1.3 Les qualités inhérentes au Cadre européen commun de référence pour les langues

Pour parvenir à réaliser les buts décrits, le CECRL doit être « aussi transparent, cohérent et exhaustif que possible » (Conseil de l'Europe 2005 : 15). Ces trois critères auxquels le document doit satisfaire sont autonomes, car ils ne dépendent pas l'un de l'autre. L'exhaustivité implique la spécification de « toute la gamme des savoirs linguistiques, des savoir-faire langagiers et des emplois de la langue » (Conseil de l'Europe 2007 : 12). Ce référentiel doit permettre à tout utilisateur de décrire ses objectifs en matière de compétences langagières à l'aide des échelles incluses. Le CECRL répond à ce critère en présentant les nombreux paramètres de la compétence langagière d'une part, et en distinguant une série de niveaux communs de référence qui servent à étalonner la compétence langagière et à calibrer les progrès d'apprentissage d'autre part (Conseil de l'Europe 2005 : 12). Il est évident que le but de ce document n'est pas et ne peut pas être la prévision de tous les emplois possibles de la langue dans toutes les situations. L'intention est cependant de rendre tous ses

utilisateurs capables de décrire leurs objectifs en faisant référence à ce document (Alderson 2004 : 2).

Le critère de transparence renvoie à la clarté des informations incluses, tant au niveau du contenu transmis qu'au niveau de la formulation. En plus d'être claires, les informations doivent être explicites afin d'être immédiatement disponibles et comprises par les utilisateurs (Alderson 2004 : 2). La cohérence implique la présence d'informations et d'analyse libres de toute contradiction (Conseil de l'Europe 2005 : 13). Ce critère exige les rapports d'équilibre entre les différents éléments constitutifs des systèmes éducatifs qui sont les besoins, les objectifs, les contenus, les curricula, les matériaux, les méthodes d'enseignement et les types d'évaluation.

Le respect de ces trois critères par le CECRL n'équivaut pas à sa transformation en un « système unique et uniforme » (Conseil de l'Europe 2005 : 13). Au contraire, il doit rester « ouvert et flexible » pour être adaptable aux situations d'enseignement, d'apprentissage et d'évaluations individuelles (Conseil de l'Europe 2005 :13). Bien que ces deux qualités soient nécessaires pour que le CECRL puisse remplir ses fonctions, elles ne sont pas synonymes. L'ouverture du CECRL implique sa capacité à être « étendu et affiné » pour pouvoir s'appliquer aux nouveaux emplois de la langue (Conseil de l'Europe 2005 :13). En revanche, la flexibilité ou la souplesse demandée signifie son adaptabilité à des conditions différentes (Alderson 2004 : 2).

Conformément aux critères évoqués, le CECRL n'a pas l'intention de dicter ni les objectifs ni les méthodes aux professionnels des langues (Tagliante 2005 : 35):

> Il ne s'agit aucunement de dicter aux praticiens ce qu'ils ont à faire et comment le faire. […]. La fonction du CECRL n'est pas de prescrire les objectifs que ses utilisateurs devraient poursuivre ni les méthodes qu'ils devraient utiliser» (Conseil de l'Europe 2005 : 35).

Les programmes et les examens d'apprentissage des langues doivent être conçus et adaptés au contexte éducatif dans lequel ils seront utilisés. Les concepteurs du CECRL sont très clairs à ce propos :

> Les utilisateurs du *Cadre de référence* envisageront et expliciteront selon le cas
> – quels sont les types d'évaluation parmi ceux présentés qui sont
> – les mieux appropriés aux besoins des apprenants dans leur système
> – les plus appropriés et les plus réalisables dans la culture pédagogique de leur système (Conseil de l'Europe 2005 : 145).

Non seulement le CECRL n'est pas prescriptif mais il n'est pas non plus dogmatique car il ne favorise aucune théorie traitant des méthodologies de l'enseignement, bien qu'il se situe dans la perspective actionnelle (Tagliante 2005 : 36).

2.2 Origine des échelles de descripteurs du CECRL

2.2.1 Origine des niveaux communs de référence

Le *Cadre Européen Commun de Référence pour les Langues* a été développé par une équipe internationale d'experts qui ont travaillé sous la direction de la Division des Politiques linguistiques du Conseil de l'Europe (Little 2007 : 646). [6] Publiée par le Conseil de l'Europe en 2001, la version complète qui fait aujourd'hui référence représente la dernière étape dans le processus de développement du CECRL (Alderson 2002 : 2). La première maquette de ce document a été diffusée en automne 1995 et envoyée aux institutions européennes d'enseignement supérieur, qui ont été invitées à la commenter. Cette demande a été suivie puisque 200 questionnaires sur 1000 ont été remplis et renvoyés au Conseil de l'Europe (Alderson 2002 : 2). Sur la base de ces commentaires, la deuxième maquette du CECRL a été élaborée et soumise à l'évaluation de la Conférence Finale du Projet en Langues Modernes qui eut lieu en avril 1997. Lors de cette conférence, il a été recommandé de piloter le Cadre européen commun de référence dans la prochaine étape de développement (Alderson 2002 : 2).

Il est évident qu'avant l'élaboration des descripteurs du CECRL, les niveaux communs de référence devaient être définis. Le premier de ces niveaux a été défini pour l'apprentissage de l'anglais par le professeur John Trim dans son ouvrage *Threshold level*, publié en 1975 par le Conseil de l'Europe. *Threshold level,* traduit par « niveau seuil » en français, doit son nom au fait que

l'apprenant accède à un seuil de communication lorsqu'il acquiert les compétences attendues à ce niveau (Tagliante 2007 : 42). Lors de sa définition, on a attribué au niveau seuil la fonction de définir la compétence langagière minimale qui rend l'apprenant capable d'utiliser la langue étrangère au quotidien, à des fins privées et professionnelles, alors qu'il se trouve dans une communauté de la langue cible (Breakthrough 2001 : 1). Selon la définition fournie par le concepteur du niveau, l'apprenant se situant au niveau seuil peut se débrouiller en voyage dans le pays de la langue cible, dans toutes les situations quotidiennes. Il est notamment capable d'entrer en contact, de maintenir les relations avec autrui et d'échanger des informations et des idées (Van Ek&Trim 1975 : 1).

La publication de l'ouvrage *Threshold level* a profondément influencé la conception de l'enseignement des langues vivantes et a encouragé son évolution. En déterminant les fonctions attribuées aux compétences langagières au niveau seuil, cet ouvrage a défini « l'approche notionnelle/fonctionnelle » d'enseignement et d'apprentissage des langues qui est à l'origine de l'approche actionnelle représentée par le CECRL aujourd'hui (Tagliante 2005 : 41). Au-delà de cette influence majeure exercée sur la définition de l'approche actionnelle du CECRL, *Threshold Level* est un modèle très utilisé depuis son apparition pour définir les objectifs d'apprentissage des langues et les spécifications de son contenu. Il est à l'origine de l'élaboration de programmes plus concrets, globaux et cohérents que ceux qui existaient jusqu'alors (van Ek 2001 : 1).

Lors de la définition du niveau seuil, ce dernier a été considéré comme le niveau le plus bas pour lequel il était possible de formuler des objectifs de compétences spécifiques (Conseil de l'Europe 2010 : 1). Or, la compétence langagière à ce niveau s'est avérée relativement avancée et a donc rendu nécessaire le ciblage des stades d'apprentissage antérieurs (Tagliante 2005 : 41). Pour cette raison et afin de satisfaire la demande des organismes d'enseignement et d'évaluation, un niveau inférieur, désigné « Waystage », a été défini dans l'ouvrage qui lui est consacré en 1991. Après la définition de celui-ci, le besoin de spécifier le niveau le plus bas de la compétence langagière semble être satisfait. Cependant, dans les années 1990, on a reconnu la nécessité de déterminer un niveau inférieur à « Waystage » ainsi qu'un niveau supérieur au

niveau « Seuil » (Conseil de l'Europe 2010 : 2). En ce qui concerne ce dernier, désigné « Vantage », il a été défini par van Ek et Trim dans un nouvel ouvrage, publié en 2001 (Tagliante 2005 : 41).

La comparaison des objectifs d'apprentissage de langue aux trois niveaux évoqués doit avant tout s'appuyer sur une définition que fournit l'introduction au « Vantage »:

> In all, *Waystage, Threshold* and *Vantage* now offer to all practitioners a description of the language needed to assure a learner's ability to deal effectively with the challenges presented by everyday life, presented at three levels rising from a minimal equipment to deal with the highest priority needs, through the minimum needed to deal with the full range of requirements for a visitor or temporary resident, to an enriched equipment adequate to deal effectively with the complexities of daily living (van Ek&Trim 2001: 6).

L'extrait cité montre que les tâches demandées aux apprenants reflètent les besoins émergents dans la vie quotidienne. Ce passage révèle aussi que la différence entre les besoins langagiers aux trois niveaux successifs concerne leur complexité. Celle-ci est minimale au niveau « Waystage » qui renvoie à des besoins langagiers très élémentaires. La complexité est moyenne au niveau « Seuil » où il est demandé aux apprenants de faire face aux besoins qui apparaissent lors de l'usage de la langue cible dans le pays où celle-ci est parlée. Enfin, le niveau « Vantage » permet de répondre à des besoins plus complexes qui sont toujours ancrés dans la vie quotidienne.

Concernant la définition du niveau inférieur au « Waystage », la *Division langues modernes* au Conseil de l'Europe a chargé le professeur Trim de procéder à la spécification de ce niveau, désigné « Breakthrough», en complément à la série des trois niveaux déjà définis. La première version non publiée à l'époque a été soumise à la *Division des Politiques linguistiques* du Conseil de l'Europe (Breakthrough 2001 : 3). L'ouvrage consacré au niveau « Breakthrough » est la dernière addition à la série des spécifications des objectifs d'apprentissage de langue inaugurée par la publication du niveau « Seuil » par le Conseil de l'Europe en 1975 (Breakthrough 2001 : 3). Ce niveau est considéré comme le niveau le plus bas auquel l'apprenant est capable de produire et de comprendre les questions et les énoncés simples, et ainsi d'interagir dans le domaine des besoins immédiats ou sur des sujets très familiers (Conseil de l'Europe 2005 : 25). La compétence au niveau *Breakthrough*

se distingue de celle aux niveaux supérieurs car elle englobe les formules langagières dans une large mesure. Ces formules sont les mots et les phrases brèves toutes faites qui sont utilisées en fonction de différentes situations. La capacité à lier les phrases les unes avec les autres est très limitée à ce niveau. Par conséquent, les capacités réceptives ne permettent pas de comprendre les textes habituels cohérents, mais seulement de reconnaître les mots et les phrases isolées, à condition de comprendre l'idée principale (Breakthrough 2001 : 4). Ce niveau peut théoriquement être atteint après 60 à 100 heures d'apprentissage, en fonction de la langue maternelle, du niveau d'éducation et de l'expérience individuelle d'apprentissage des langues (Tagliante 2005 : 42).

Les ouvrages évoqués se sont avérés très utiles dans l'enseignement des langues depuis leur apparition et ils continuent à l'être à l'heure actuelle. Ainsi, les descriptions des niveaux de compétence entreprises par ces ouvrages servent souvent encore de base à la conception des nouveaux programmes d'enseignement. Ils permettent d'élaborer des manuels plus motivants et d'introduire des systèmes d'évaluation plus réalistes et plus transparents que ceux en vigueur avant la publication de ces ouvrages. En outre, ces descriptions des niveaux de compétence ascendants constituent l'une des origines des échelles à six niveaux du CECRL, qui font l'objet du sous-chapitre suivant.[7]

2.2.2 Origine des échelles des descripteurs

Le *Cadre européen commun de référence pour les langues* et le *Portfolio européen des langues* (PEL) sont liés dès leur conception. Celle-ci fut décidée au Symposium de Rüschlikon, organisé en 1991 par les autorités suisses (North 2000 : 7) autour du thème « Transparency and Coherence in Language Learning in Europe : objectives, evaluation and certification ». Les participants au Symposium décidèrent de collaborer avec le Conseil de l'Europe afin d'encourager l'élaboration d'un « cadre européen commun de référence » et de former un groupe de travail pour examiner les formes et les fonctions possibles d'un « portfolio européen de langues ».[8] L'impulsion était donnée.

C'est lors du Symposium de Rüschlikon que l'idée d'une échelle européenne commune de compétences langagières émergea et qu'il fut décidé de calquer celle-ci sur celle suggérée pour le PEL. Cette échelle, on le sait, est

devenue l'élément central du CECRL (North 2007 : 656). En effet, l'échelle esquissée pour le PEL contenait une échelle globale et six sous-échelles. Les six niveaux adoptés par le *Cadre européen commun de référence* étaient identiques à ceux utilisés par *l'Association européenne d'évaluation en langues*. Le niveau A1 y a été ajouté (North 2007 : 656). [9]

L'échelle commune utilisée par *l'Association européenne d'évaluation en langues* contenait cinq niveaux communs de référence à l'époque de la conception de l'échelle européenne commune de compétences langagières (ALTE 2002 : 7). Par la suite, le niveau inférieur *Breakthrough* a été inséré dans l'échelle ALTE, mais il n'en faisait pas encore partie à l'époque de l'élaboration des niveaux communs de référence du CECRL (ALTE 2002 : 7). Le tableau ci-dessous illustre le lien entre les deux cadres communs de référence. [10]

Council of Europe Levels	A1	A2	B1	B2	C1	C2
Alte levels	ALTE Breakthrough Level	ALTE Level1	ALTE Level2	ALTE Level3	ALTE Level4	ALTE Level5

Les échelles de compétences contenues dans le CECRL sont fondées sur les résultats de la recherche effectuée entre 1993 et 1996 dans le cadre d'un projet du Fonds national suisse de recherche scientifique (Conseil de l'Europe 2005 : 155). En l'occurrence, deux études ont été menées, d'une durée d'un an chacune. La première, réalisée entre 1994 et 1995, se focalise sur deux compétences, l'interaction et la production orale en anglais langue étrangère, et se limite à l'évaluation par les enseignants (Conseil de l'Europe 2005 : 155). La deuxième étude, menée entre 1995 et 1996, évalue en complément les compétences réceptives, non seulement en anglais mais aussi en français et en allemand (Conseil de l'Europe 2005 : 155). Outre l'évaluation par les enseignants, cette deuxième étude englobe l'auto-évaluation par les candidats et les informations provenant d'examens différents, par exemple, le DELF et le DALF (Conseil de l'Europe 2005 : 155).

Les concepteurs des échelles de compétences ne se sont pas contentés d'élaborer des échelles. Ils se sont attachés à fournir aux descripteurs autant de catégories descriptives que possible (North 2007 : 656). Les échelles de compétences et les descripteurs sont structurellement liés depuis leur origine. Ils ont été développés simultanément pendant les étapes qui ont conduit à la production des échelles de compétences relevant de différentes catégories (North 2000 : 124-125).

Premièrement, une étude exhaustive des échelles de compétences existantes a été menée. Trente échelles de compétences ont été choisies en tant que sources pour le développement de la banque des descripteurs (North 2000 : 181-182). Les échelles sélectionnées font partie de plusieurs catégories, en l'occurrence, des échelles d'interaction orale et de production orale en différents contextes (North 2000 : 21-22).

La banque d'items a été produite en étudiant le contenu de chaque échelle de compétences pour la production orale et pour l'interaction. Le contenu a été divisé en phrases isolées. Ensuite, chaque phrase a été analysée séparément afin de déterminer la catégorie décrite par celle-ci et a été attribuée à la même catégorie dans la banque des descripteurs au cours du développement (North 2000 : 182). Les catégories conçues ont été regroupées en catégories hiérarchiquement organisées qui décrivent les compétences et les activités communicatives (North 2000 : 124)[11] (North 2000 : 182-183). En ce qui concerne les catégories pour les activités communicatives, celles-ci n'ont pas été complètement élaborées dans cette version, étant donné que la première étude s'est concentrée sur l'interaction et la production orale. Les descripteurs entrés dans la banque d'items ont été regroupés en six niveaux, sur la base des échelles d'ALTE et de l'échelle maquette du *Portfolio Européen de Langues*, évoquées précédemment. La banque d'items ainsi élaborée contient 1679 descripteurs potentiels qui relèvent de la production orale, de l'interaction et de la compétence langagière globale (North 2000 : 184). Ceux-ci ont été complétés par les descripteurs des échelles de la production écrite et de la compréhension de l'oral, ce qui a fait passer leur nombre à 2047 descripteurs (North 2000 : 184). Par la suite, les descripteurs ont été édités, premièrement, afin d'éviter les répétitions et d'en réduire le nombre, et deuxièmement, afin de les reformuler de façon positive. Ceci s'est révélé nécessaire pour le processus de validation

qualitative inscrit à l'étape suivante et en vue de l'utilisation des descripteurs en tant que critères pour l'évaluation des étudiants (North 2000 : 185). Ce travail d'édition a abouti au maintien d'environ 1000 descripteurs.

L'étape suivante fut celle des ateliers de pilotage effectués par les professeurs (North 2000 : 185). Les objectifs de ces sessions étaient de contrôler le choix des catégories descriptives aussi bien que des descripteurs. En l'occurrence, il s'agissait de vérifier l'usage adéquat des catégories descriptives, la formulation transparente des descripteurs, leur pertinence pour la catégorie à laquelle ils appartiennent et les secteurs éducatifs concernés (North 2000 : 185). Pour atteindre ces objectifs, la nécessité de sélectionner des descripteurs de meilleure qualité se posa à nouveau. A cet effet, onze workshops furent organisés en 1994 auxquels participèrent des professeurs de tous les secteurs éducatifs impliqués[12] (North 2000 : 185). Dix sessions se concentrèrent sur les catégories descriptives tandis que la onzième séance eut pour objectif de répartir les descripteurs selon les différents niveaux de compétence provisoires (North 2000 : 185).

Deux techniques furent utilisées au cours de tous ces ateliers. La première consistait à commenter des enregistrements vidéo présentant des interactions en anglais, entre paires de jeunes adultes sur des sujets variés. S'agissant d'une étude suisse, les langues maternelles des participants étaient l'une des quatre langues suivantes : le français, l'allemand, l'espagnol ou l'italien, la langue choisie devant toujours être différente de celle de l'interlocuteur (North 2000 : 185). Les discussions menées au sujet de ces interactions eurent pour but de vérifier que leur contenu était bien couvert par les descripteurs et par les catégories incluses dans la banque élaborée et, le cas échéant, de déterminer si certaines descriptions de la performance d'interlocuteurs pouvaient être éditées sous forme de descripteurs (North 2000 : 187). Les objectifs visés par cette technique furent atteints car l'usage de ces catégories descriptives lors des discussions a depuis été démontré (North 2000 : 185). La deuxième technique consista à répartir les descripteurs contenus dans la banque selon les différentes catégories. A côté de cette tâche, les descripteurs peu clairs ou situés dans une catégorie non-pertinente furent signalés (North 2000 : 189). La catégorisation initiale des descripteurs fut suivie par un travail de reformulation de ceux qui étaient jugés peu clairs, mal formulés ou trop longs, ce qui concernait tout de

même la moitié des descripteurs analysés. Cette procédure servit à vérifier que les catégories, attribuées aux descripteurs lors de la conception de la banque d'items, étaient pertinentes (North 2000 : 191).

A la suite de cette procédure, l'organisation des descripteurs en niveaux de référence provisoires fut contrôlée lors du dernier atelier, auquel participèrent 25 professeurs. A cet usage, chacun des 400 descripteurs maintenus fut réparti selon trois niveaux généraux qui furent eux-mêmes subdivisés en deux niveaux plus fins, ultérieurement (North 2000 : 191). Cette dernière séance servit moins à exclure les descripteurs qu'à préparer les questionnaires pertinents pour l'enquête ultérieure prévue (North 2000 : 191). Le choix des descripteurs retenus pour les questionnaires s'appuyait sur le niveau présumé des élèves impliqués dans l'enquête (North 2000 : 191-192).

Pour collecter les données empiriques, 7 questionnaires composés de 50 descripteurs chacun furent élaborés. Les questionnaires étaient situés à des niveaux de compétences différents, du niveau *Breakthrough* jusqu'au niveau *Effectiveness,* ce dernier correspondant au niveau C1 du CECRL (Conseil de l'Europe 2005 : 24). Puisque l'élaboration des questionnaires devait se faire selon les niveaux de compétences, la sélection des descripteurs fut effectuée sur la base de leur attribution aux niveaux de compétence provisoires dans la banque des descripteurs, et en tenant compte de la confirmation de leur niveau de difficulté lors des ateliers (North 2000 : 193). Le deuxième facteur, pris en compte lors du choix des descripteurs pour les questionnaires, était constitué par les jugements des professeurs à propos de la pertinence des descripteurs particuliers pour leur secteur éducatif. Les descripteurs furent regroupés selon les aspects thématiques suivants : les tâches parlées, les qualités de performance parlée, la compréhension, les stratégies d'interaction ainsi que les tâches écrites (North 2000 : 193). Les questionnaires furent attribués aux apprenants en fonction du nombre d'années et d'heures hebdomadaires consacrées à l'apprentissage formel de l'anglais. Chaque professeur devait choisir dix apprenants pour les évaluer à l'aide des questionnaires (North 2000 : 208).

La deuxième procédure mise en œuvre, la conférence d'évaluation, avait pour objectif de fournir des estimations de la sévérité des professeurs. Ces estimations devaient être prises en compte au cours des évaluations du niveau

de compétences des apprenants dans l'enquête questionnaire (North 2000 : 193). Les descripteurs qui couvrent la production et l'interaction orales, aussi bien préparées que spontanées, furent pilotés à l'aide de performances vidéo. Les sujets proposés aux candidats des niveaux élémentaire et intermédiaire relevaient de la vie quotidienne tandis que les enjeux controversés furent mis à la disposition des apprenants avancés (North 2000 : 200). À cet effet, un mini questionnaire, contenant une sélection de cinq à sept descripteurs jugés suffisamment clairs, fut établi pour chaque performance vidéo. Ces descripteurs relevaient du questionnaire utilisé pour l'enquête, au niveau de compétence correspondant. Une partie des descripteurs choisis définissait les tâches accomplies lors des performances, dont celles de production et d'interaction orales. La deuxième partie des descripteurs déterminait les aspects qualitatifs de la performance, en l'occurrence l'étendue, la correction, l'aisance et la prononciation (North 2000 : 205). La procédure d'évaluation consista, en premier lieu, à marquer le(s) descripteur(s) pertinent(s) de façon provisoire en comparant ce(s) dernier(s) à la performance en cours. En deuxième lieu, on porta le jugement définitif sur l'évaluation, établi à partir de la première impression (North 2000 : 199). Ces deux procédures empiriques permirent d'analyser les descripteurs à l'aide du programme statistique FACETS, de déterminer la valeur de facilité des items, l'erreur standard de mesure, le degré d'adaptation des items à la performance attendue ainsi que les niveaux de compétence des candidats (North 1999 : 208).

2.3 Présentation des échelles illustratives du CECRL

L'un des objectifs du *Cadre Européen Commun de Référence pour les Langues* est de rendre « la comparaison entre les différents systèmes de qualifications » plus facile et par conséquent, de promouvoir « la reconnaissance réciproque des qualifications en langues » (Conseil de l'Europe 2005 : 23, 12). Il est évident que la mise à disposition de critères objectifs pour la description des niveaux de compétence langagière est indispensable à cette fin. Pour fournir de tels critères dans les examens existants, les concepteurs du CECRL ont élaboré un schéma descriptif et des niveaux communs de référence (Conseil de l'Europe 2005: 23).

2.3.1 Le schéma descriptif

Le schéma descriptif contient toutes les compétences impliquées dans l'usage et l'apprentissage langagiers (Little 2007 : 646). Il comporte, premièrement, quatre compétences générales, en l'occurrence savoir, savoir-faire, savoir-être et savoir-apprendre. Deuxièmement, il englobe quatre compétences langagières communicatives ; linguistiques, pragmatiques, sociolinguistiques et socioculturelles (Little 2007 : 646). Quant aux compétences linguistiques, celles-ci comprennent les habiletés lexicale, grammaticale, sémantique, phonologique, orthographique et ortho épique (Conseil de l'Europe 2005 : 81). La troisième composante du schéma descriptif est constituée par les quatre catégories de l'activité langagière; la réception, la production, l'interaction et la médiation (Little 2007 : 646).[13] Son quatrième élément se compose des quatre domaines d'usage langagier; privé, public, professionnel et éducatif (Little 2007 : 646). La composante finale du schéma descriptif englobe les types de paramètres qui déterminent l'usage de la langue : le contexte situationnel, le type de texte ou le thème, ainsi que les conditions ou les contraintes dans lesquelles l'usage langagier se déroule (Little 2007 : 646).

Le schéma descriptif présenté révèle la perspective du CECRL selon laquelle non seulement les catégories de l'activité langagière et les compétences langagières communicatives sont impliquées dans l'usage et l'apprentissage d'une langue, mais également les quatre compétences générales. Cette vue globale sur l'usage et l'apprentissage de la langue implique que ces activités engagent toutes les connaissances et les capacités d'un être humain, y compris non-linguistiques. Les catégories de l'activité communicative et les compétences langagières communicatives sont les « paramètres » de la compétence langagière qui forment la « dimension horizontale » du *Cadre commun de référence* (Conseil de l'Europe 2005 : 19). Chacun de ces paramètres s'organise en six niveaux communs de référence qui forment la « dimension verticale » en fonction de leur ordre ascendant (Conseil de l'Europe 2005 : 19). Les désignations « dimension horizontale » et «dimension verticale» proviennent du fait qu'il est courant de présenter les paramètres de la compétence langagière et

les six niveaux de référence sous la forme d'une grille (Conseil de l'Europe 2005 : 19).

Le CECRL présente 34 échelles illustratives pour les activités de compréhension de l'oral et de l'écrit, de production de l'oral et de l'écrit, pour l'interaction orale et écrite, ainsi que pour la prise de notes (Little 2007 : 646). En revanche, les échelles ne sont pas disponibles pour la médiation (Little 2007 : 646). Les échelles illustratives sont résumées par deux échelles. La première, globale, décrit brièvement la compétence communicative générale à chacun des six niveaux (Little 2007 : 646). La deuxième échelle, qui a pour fonction de résumer les définitions de compétence contenues dans les échelles illustratives, est la grille d'auto-positionnement (Little 2007 : 646). Celle-ci présente la compétence de manière concise, distincte selon les différents niveaux et selon les cinq activités de communication langagière (Little 2007 : 646, Conseil de l'Europe 2005 : 26-27).[14]

Conformément à l'approche actionnelle adoptée par le CECRL, des échelles sont également proposées pour les stratégies de production, à savoir la « planification », la « compensation », le « contrôle » et la « correction » (Conseil de l'Europe 2005 : 53-54). Deuxièmement, une échelle est disponible pour les stratégies de réception, à savoir « reconnaitre des indices et faire des déductions », et troisièmement pour les stratégies d'interaction que sont les « tours de parole », « coopérer » et « faire clarifier » (Conseil de l'Europe 2005 : 70-71). Enfin, les échelles existent pour les 13 paramètres de compétences communicatives langagières. La première catégorie de compétence communicative langagière – les compétences linguistiques – contient des échelles pour les paramètres suivants : « étendue linguistique générale », «étendue de vocabulaire », «maîtrise de vocabulaire », « correction grammaticale », «maîtrise du système phonologique » et «maîtrise de l'orthographe » (Conseil de l'Europe 2005 : 88-90, 92-93). La deuxième composante de ces compétences, la compétence sociolinguistique, présente une échelle pour le paramètre de « correction sociolinguistique » (Conseil de l'Europe 2005 : 95). Le CECRL souligne la difficulté d'étalonner la compétence sociolinguistique selon les niveaux car les utilisateurs élémentaires (A1-A2) ne maîtrisent pas encore toutes les composantes intégrées dans cette compétence

(Conseil de l'Europe 2005 : 95). Quant à la compétence pragmatique, les échelles existent pour ses deux composantes, à savoir la compétence discursive et la compétence fonctionnelle (Conseil de l'Europe 2005 : 81). Concernant la compétence discursive, il s'agit de quatre paramètres que sont la « souplesse », les « tours de parole », le «développement thématique » ainsi que la « cohérence et cohésion » (Conseil de l'Europe 2005 : 97- 98). Pour ce qui est de la compétence fonctionnelle, on trouve les échelles pour ces deux paramètres : l'« aisance à l'oral » et la « précision » (Little 646 : 2007).

Les échelles de compétence ne sont pas arrangées dans un ordre hiérarchique. Ceci implique qu'il appartient aux utilisateurs du CECRL de déterminer quelles échelles sont pertinentes pour leur(s) objectif(s) particulier(s), dans le sens que celles-ci permettent de décrire une activité communicative ciblée. En outre, un lien doit être établi entre les échelles de compétence pertinentes (Little 646 : 2007).

2.3.2 Les niveaux de compétence

Le CECRL parle d'un accord sur « le nombre et la nature des niveaux appropriés pour l'organisation de l'apprentissage en langues et une reconnaissance publique du résultat » (Conseil de l'Europe 2005 : 24). Le consensus qui existe sur ce point détermine que les six niveaux communs de référence couvrent entièrement « l'espace d'apprentissage pertinent pour les apprenants européens en langues » (Conseil de l'Europe 2005 : 24). Les six niveaux de compétences en langues vivantes étrangères ou régionales forment la « dimension verticale » évoquée dans le CECRL (Conseil de l'Europe 2005 : 31). Les niveaux communs de référence suivants sont identifiés et décrits dans ce référentiel et forment une série ascendante : A1, A2, B1, B2, C1 et C2 (Conseil de l'Europe 2005 : 31). Ces six niveaux se regroupent en trois tranches de compétence A, B et C (Byrnes 2007: 642). La première de ces tranches recouvre les niveaux A1 et A2 et constitue le domaine de compétences de l'utilisateur élémentaire, tandis que la deuxième tranche, contenant les niveaux B1 et B2, est maîtrisée par un utilisateur indépendant. La troisième tranche de compétences quant à elle recouvre les niveaux C1 et C2 qui dénotent un utilisateur expérimenté (Little 2007 : 646).

L'appartenance de deux niveaux subordonnés à une même tranche de compétences ne les empêche pas d'être indépendants et situés dans un ordre ascendant. Ce fait est reflété par leur désignation. Ainsi, le domaine de compétences de l'utilisateur « élémentaire » se compose du niveau A 1 nommé « Introductif ou découverte », ainsi que du niveau A2, appelé « Intermédiaire ou de survie » (Conseil de l'Europe 2005 : 25). La tranche intermédiaire de compétences englobe le niveau B1, appelé « Niveau Seuil », et B2 intitulé « Avancé ou indépendant » (Conseil de l'Europe 2005 : 25). En ce qui concerne la tranche supérieure de compétences, son niveau inférieur, C1, est appelé « autonome » et celui situé au-dessus, C2, « Maîtrise » (Conseil de l'Europe 2005 : 25). Les six niveaux de compétence, aussi appelés « points communs de référence » et les « niveaux communs de référence » composent une échelle de compétences (Conseil de l'Europe 2005 : 25)

Bien qu'il existe un consensus sur la répartition de « l'espace d'apprentissage pertinent pour les apprenants européens en langues » sur une échelle de six niveaux de référence, il ne s'agit que de la répartition adoptée et utilisée en pratique et non pas de la distinction empiriquement confirmée (Conseil de l'Europe 2005 : 24). En effet, les résultats empiriques du projet suisse préconisent l'adoption d'«une échelle sur neuf niveaux cohérents et à peu près égaux » (Conseil de l'Europe 2005 : 30). En supplément des niveaux conventionnels, cette échelle comprend des étapes situées entre les niveaux A2 et B1, entre B1 et B2 ainsi qu'entre B2 et C1 (Conseil de l'Europe 2005 : 30). Tandis que les six niveaux conventionnels sont qualifiés de « niveaux critériés », les étapes intermédiaires s'appellent les « niveaux avancés » (Conseil de l'Europe 2005 : 30). Ces derniers peuvent être marqués de deux manières à l'écrit : soit par un « + », soit par le chiffre « 2 », qui sont ajoutés au niveau dit « critérié » (Conseil de l'Europe 2005 : 30). Ainsi, le niveau situé entre A2 et B1 peut s'écrire soit « A2+ » soit « A2.2 » (Conseil de l'Europe 2005 : 31).[15] Ces niveaux se caractérisent par une performance plus forte par rapport aux mêmes qualités trouvées aux niveaux critériés, en plus des indications de qualité qui deviennent saillantes au niveau supérieur (North 2007 : 657). L'existence des « niveaux avancés » révèle que la délimitation entre les niveaux individuels implique toujours des décisions subjectives par leurs utilisateurs. Ceci est prévu

par le CECRL qui, dans le but de répondre aux besoins locaux de ses différents utilisateurs, propose un ensemble de niveaux susceptibles d'être distingués selon des degrés de finesse différents. Ainsi, l'option pour les niveaux conventionnels larges ou pour les niveaux étroits dépend des besoins et des préférences des institutions locales. La souplesse proposée par le « système d'arborescence » décrit se manifeste aussi par les deux autres possibilités offertes aux utilisateurs (Conseil de l'Europe 2005 : 35). La première consiste en un emploi combiné des différents niveaux de finesse par la même institution selon les différentes situations qui se présentent.[16] La deuxième possibilité, mise à disposition des institutions par ce « système d'arborescence », est d'affiner les niveaux de compétence qui conviennent à leur contexte particulier d'enseignement des langues et au public cible (Conseil de l'Europe 2005 : 35). En développant l'ensemble des niveaux présentés, les concepteurs du CECRL ne prétendent pas recouvrir la totalité de la compétence langagière possible des apprenants. Ainsi, le CECRL souligne que certaines tâches peuvent être réalisées efficacement au niveau inférieur à A1, malgré un répertoire linguistique très limité à ce stade d'apprentissage (North 2007 : 657). Par conséquent, ce document attire l'attention sur l'utilité d'énumérer ces tâches « simples et globales », en cas d'enseignement à de jeunes débutants (Conseil de l'Europe 2005 : 30). [17]

Les échelles de compétences, incluses dans le CECRL, ont volontiers été adoptées par les professionnels de l'enseignement des langues vivantes (Little 2007 : 648). Leur succès s'explique par la combinaison entre ses qualités traditionnelles d'une part et ses caractéristiques innovatrices d'autre part (Little 2007 : 648). Les six niveaux distingués au sein de chaque échelle ne sont en effet pas complètement nouveaux, puisqu'ils renvoient à la distinction traditionnelle en « niveau de base », « niveau intermédiaire » et « niveau avancé » (Hulstijn 2007 : 1). Le trait novateur des échelles de compétences tient à leur potentiel plus large que celui des échelles traditionnelles. Les échelles de compétences du CECRL offrent ainsi un moyen beaucoup plus fin de décrire les niveaux de compétence en langue seconde que ce qui était possible avant l'existence de ce référentiel (Little 2007 : 648). Le succès de ces échelles de compétences n'empêche pas qu'elles posent un grand nombre de défis aux professionnels de l'enseignement des langues, défis dont il sera question ultérieurement (Little 2007 : 648).

La deuxième raison du succès de ce document est que les niveaux communs de référence ne sont pas spécifiques à une langue donnée, mais restent les mêmes pour toutes les langues (Little 2007 : 646). Pour cette raison, les traductions du *Cadre européen commun de référence* ont été effectuées et distribuées en 18 langues jusqu'à l'année 2005 (Tagliante 2005 : 35). La validité des niveaux de compétences pour toutes les langues est la condition essentielle pour pouvoir atteindre l'objectif déclaré par ce document qui est de fournir « une base commune pour l'élaboration de programmes de langues vivantes, de référentiels, d'examens, de manuels, etc. en Europe » (Conseil de l'Europe 2005 : 9).

2.3.2.1 Les types d'échelles de compétences

Une distinction a été établie entre trois types d'échelles de compétences selon la fonction qu'elles remplissaient (Alderson 1991 : 72). Le premier type est constitué par les échelles centrées sur l'utilisateur, le deuxième type englobe celles prévues pour l'examinateur. Quant au troisième type d'échelles, il se concentre sur le concepteur de tests et d'examens (Conseil de l'Europe 2007 : 5). Ces trois types d'échelles de compétences sont présents dans le CECRL.

Les échelles centrées sur l'utilisateur contiennent les descripteurs qui définissent les compétences des apprenants. Ces échelles sont globales car elles comportent un seul descripteur par niveau (Conseil de l'Europe 2005 : 35). Leur trait caractéristique est la simplicité. L'échelle globale des niveaux communs de référence, présentée dans le premier tableau du CECRL, constitue un exemple de ce type d'échelle (Conseil de l'Europe 2005 : 25). [18]

Les échelles centrées sur l'examinateur ont pour fonction de guider la notation. Par conséquent, elles se focalisent sur « la qualité de la performance de l'apprenant » (Conseil de l'Europe 2005 : 35). Les échelles de ce type peuvent être soit globales et présenter un seul descripteur par niveau, soit analytiques. Les échelles analytiques décrivent les différents aspects de la performance. Le tableau 3 du CECRL qui décrit les aspects qualitatifs de l'utilisation de la langue parlée par niveau de compétences, est un exemple d'échelle analytique (Conseil de l'Europe 2005 : 28). Conformément au destinataire de ce type d'échelles, elles servent à diagnostiquer les compétences des apprenants.

Les échelles centrées sur le concepteur servent de base à la conception des tests aux niveaux respectifs. Leurs énoncés définissent les tâches communicatives spécifiques que les participants aux tests doivent pouvoir effectuer à chaque niveau de compétences. Comme pour les échelles centrées sur l'utilisateur, ce type d'échelles se focalise aussi sur ce que l'apprenant est « capable de faire » (Conseil de l'Europe 2005 : 35). Les échelles illustratives des activités langagières et des compétences communicatives langagières sont des exemples de ce type d'échelle, car elles peuvent servir à la conception de tests visant l'évaluation de capacités spécifiques.

Avant d'emprunter des descripteurs au CECRL pour les utiliser, par exemple, dans l'élaboration d'un portfolio européen en langues, il est essentiel de prendre en compte le type d'échelles auquel ils appartiennent (Conseil de l'Europe 2007 : 4). Cela est nécessaire pour se décider quant à l'usage possible du futur portfolio européen en langues et, plus particulièrement, quant à son potentiel pour l'auto-évaluation (Conseil de l'Europe 2007 : 4). Afin d'être utilisables pour l'auto-évaluation, les descripteurs doivent être déclinés à la première personne du singulier, à l'instar des descripteurs contenus dans la *Grille pour l'auto-évaluation*. Par ailleurs, ils doivent parfois être simplifiés ou fractionnés (Conseil de l'Europe 2007 : 4). Ces procédures d'adaptation des descripteurs ont souvent lieu lors de l'élaboration des portfolios européens des langues, présentés ultérieurement (Conseil de l'Europe 2007 : 4).

2.3.3 La critique du système basé sur les échelles

Bien que le CECRL jouisse d'une grande reconnaissance au sein de l'Union Européenne, y compris chez les linguistes, certains spécialistes critiquent ce document, lui reprochant un échelonnage des niveaux de compétences (Narcy-Combes 2007 : 9). Ces chercheurs, qui adhèrent au concept de « non-linéarité des apprentissages », reprochent donc au CECRL d'encourager une conception linéaire du processus d'apprentissage (Narcy-Combes 2007 : 9). Ils dénoncent en l'occurrence la structuration de la compétence en différents niveaux qui, selon eux, présentent l'enseignement et l'apprentissage en mode discontinu. La discontinuité est constituée par l'avancement des apprenants d'un niveau de compétence à un autre, alors que ces niveaux ne sont pas explicitement liés. Un

autre point de critique de cette « pédagogie par palier ou par stade » (Narcy-Combes 2007 : 9) est que le retour en arrière d'un niveau supérieur à un niveau inférieur n'est pas possible (Narcy-Combes 2007: 9). La critique du CECRL est partagée par les adeptes des théories émergentistes et socioconstructivistes, selon lesquelles l'émergence des phénomènes langagiers, par exemple, la syntaxe, le lexique, la morphologie ou la phonologie, se déroule de manière non-linéaire et par conséquent, non-prévisible, en raison de l'influence des contextes sociaux et psychologiques passés et présents sur l'apprentissage des langues vivantes (Narcy-Combes 2007 : 5). Selon ce principe, l'émergence de nouvelles pratiques langagières présuppose un certain nombre de ruptures comme les retours en arrière, qui ne sont pas prévisibles, mais seulement analysables après leur apparition (Narcy-Combes 2007 : 9). Puisque dans cette perspective les ruptures dans l'apprentissage sont aussi imprévisibles qu'inévitables, l'individualisation des apprentissages demande la mise en œuvre d'un dispositif qui favorise l'émergence de ces ruptures que les apprenants modifient avec le temps (Narcy-Combes 2007 : 9).

La particularité essentielle des perspectives émergentiste et socioconstructiviste est qu'elles tiennent compte du déclenchement social des processus neurocognitifs dont font partie les phénomènes langagiers (Narcy-Combes 2007 : 9). Cette idée reflète les théories socio-interactionnistes qui représentent une approche sociale des phénomènes langagiers (Narcy-Combes 2007 : 4). La théorie socio-interactionniste la plus connue est le socio-cognitivisme de Vygotski (1934) qui insiste sur le « déclenchement social » des opérations cognitives (Vygotski 1934 : 156). Les théories émergentiste et socioconstructiviste ne sont bien sûr pas identiques, mais leurs apports sont complémentaires, c'est-à-dire nécessaires pour prendre en compte tous les aspects du processus d'apprentissage. Il est essentiel de garder à l'esprit, lors de la conception de dispositifs, que les apprentissages ne sont efficaces que si leur sens social et affectif est compris par un apprenant (Narcy-Combes 2007 : 12).

A première vue, la nécessité de comprendre le sens des dispositifs d'apprentissage ne paraît pas propre à ces deux théories, car elle est aussi exigée par le *Cadre européen commun de référence.* La perspective actionnelle, inhérente au CECRL, établit une démarche d'apprentissage déterminée par

« l'accomplissement des tâches à effectuer » (Conseil de l'Europe 2005 : 15). Cette démarche d'apprentissage demande à l'apprenant de prendre connaissance de la tâche à réaliser ainsi que des activités langagières nécessaires à son accomplissement afin de comprendre pourquoi il a besoin d'apprendre (Bourguignon 2010 : 35). Selon ce cadre d'apprentissage, la prise de connaissance d'une tâche par un apprenant implique la compréhension des raisons pour ses besoins individuels d'apprendre (Bourguignon 2010 : 35). Puisque les raisons motivant la nécessité d'apprendre résultent de la tâche à effectuer, leur saisie n'implique pas nécessairement la compréhension du sens social et affectif des tâches, exigée par les théories émergentiste et socioconstructiviste. On peut très bien comprendre les raisons motivant un apprentissage pour être en mesure d'accomplir une tâche donnée, sans forcément comprendre le sens social ou affectif de la tâche en question. La tâche perd-elle pour autant de son efficacité ? Pour les émergentistes et les socioconstructivistes la perte est certaine.

2.4 Analyse des descripteurs

Au sein des 34 échelles illustratives de compétences et des deux échelles servant de les résumer, contenues dans le CECRL, se trouvent les descripteurs qui sont définis comme « les descriptions de la compétence langagière à des niveaux différents» (Conseil de l'Europe 2005 : 150). Conformément à l'approche actionnelle, les descripteurs définissent la compétence langagière en termes de ce que l'apprenant et l'usager d'une langue peuvent faire en langue seconde (L2) (Little 2007 : 646). De ce fait, les descripteurs renvoient aux « actes » que les apprenants doivent pouvoir réaliser moyennant l'usage de la langue (Lallement 2007 : 21). Comme il a été expliqué dans la sous-partie consacrée à l'approche actionnelle, les actes décrits par les descripteurs sont « les actions en contexte social » qui demandent aux usagers et aux apprenants d'une langue des activités langagières aussi bien que non-langagières (Conseil de l'Europe 2005 : 15, Tagliante 2005 : 36).

Les descripteurs contenus dans le CECRL forment un système complexe car ils fournissent le contenu communicatif pour les six niveaux de compétence distingués (Hulstijn 2007 : 1). Selon North, les descripteurs font référence aux

types de textes, aux thèmes, aux conditions et aux contraintes d'usage de la langue, mais ils évitent d'identifier un domaine quelconque (North 2007 : 656). La grande majorité des descripteurs n'évoquent cependant pas la notion de domaine, mais elle apparaît néanmoins dans un certain nombre de définitions de compétences. Pour donner un exemple, citons le descripteur du niveau B1 de l'échelle de compétences « Interaction orale générale » :

> Peut communiquer avec une certaine assurance sur des sujets familiers habituels ou non en relation avec ses intérêts et son domaine professionnel. Peut échanger, vérifier et confirmer des informations, faire face à des situations moins courantes et expliquer pourquoi il y a une difficulté. Peut exprimer sa pensée sur un sujet abstrait ou culturel comme un film, des livres, de la musique, etc. (Conseil de l'Europe 2005 : 61).

La mention du domaine dans le descripteur cité n'est pas liée à l'activité d'interaction, car cette notion apparaît également au sein de l'échelle « Compréhension générale de l'écrit ».

> (Celle-ci) peut comprendre dans le détail des textes longs et complexes, qu'ils se rapportent ou non à son domaine, à condition de pouvoir relire les parties difficiles (Conseil de l'Europe 2005 : 57).

Même sans se référer à un domaine de façon explicite, une grande partie des descripteurs permettent de déduire le contexte grâce aux informations concernant le thème, le type de texte, et/ou la situation d'usage de la langue. Le descripteur de niveau C1, situé au sein de l'échelle « Comprendre en tant qu'auditeur », par exemple, relève clairement du domaine professionnel bien qu'il ne l'évoque pas explicitement : « Peut suivre la plupart des conférences, discussions et débats avec assez d'aisance. » (Conseil de l'Europe 2005 : 56).

Contrairement aux échelles de compétences, le système des descripteurs représente une nouveauté, à plusieurs points de vue (Hulstijn 2007 : 1). En premier lieu, le procédé d'élaboration des descripteurs sera décrit avant d'aborder les problèmes révélés par les méthodes de conception et d'étalonnage. Ensuite, les dimensions qualitative et quantitative des descripteurs seront examinées.

2.4.1 Procédé d'élaboration des descripteurs

Les descripteurs ont été conçus sur la base des résultats de la recherche menée entre 1993 et 1996 dans le cadre d'un projet du *Fonds national suisse de*

recherche scientifique. L'objectif de ce projet était de déterminer avec précision les niveaux de compétence des différents paramètres au sein du « schéma descriptif » du CECRL (Conseil de l'Europe 2005 : 155). Grâce à un procédé d'étalonnage, appliqué de façon cohérente, les descripteurs sont situés à un niveau de compétence particulier à l'intérieur des échelles du CECRL (Conseil de l'Europe 2005 : 150). Le placement des descripteurs aux différents niveaux a été entrepris moyennant une combinaison de méthodes intuitive, quantitative et qualitative (Conseil de l'Europe 2005 : 150). Ce projet comporte quatre étapes distinctes et successives, dont la première est intuitive, la deuxième qualitative, la troisième quantitative. La quatrième phase constitue l'étape d'interprétation (Conseil de l'Europe 2005 : 156).

Avant d'expliquer la manière dont les trois méthodes ont été combinées, il convient de définir ces dernières. Les méthodes intuitives ne nécessitent aucune collecte méthodique de données. Elles se fondent sur un principe d'interprétation de l'expérience (Conseil de l'Europe 2005 : 150). Les méthodes qualitatives impliquent des séances de travail avec des groupes d'informateurs et une interprétation qualitative et non pas statistique des informations collectées. En ce qui concerne les méthodes quantitatives, l'analyse statistique est centrale à celles-ci et les résultats d'analyse doivent être interprétés avec prudence (Conseil de l'Europe 2005 : 150).

Dans un premier temps, lors de la première étape qualifiée d'intuitive, les échelles de compétences existantes utilisées dans le domaine public sont analysées en détail. Dans un deuxième temps, ces échelles sont déconstruites et les descripteurs classés selon les catégories descriptives élaborées par le CECRL. Cette étape est menée dans le but d'obtenir un nombre initial de descripteurs en rapport avec les catégories du « schéma descriptif » du CECRL (Conseil de l'Europe 2005 : 156 ; Little 2005 : 648).

La deuxième étape, qualitative, consiste à analyser les enregistrements d'enseignants lors de l'évaluation des capacités des apprenants, manifestées lors des enregistrements vidéo. Cette analyse est effectuée dans le but de vérifier que le métalangage des enseignants correspond bien à celui des descripteurs. Ensuite, les descripteurs sont répartis dans les catégories qu'ils décrivent, ce qui permet de déterminer la qualité des descripteurs. Ensuite, les

descripteurs sont classés selon les niveaux communs de référence (Little 2005 : 648).

Dans la troisième phase, quantitative, les apprenants représentatifs sont évalués par les enseignants grâce à une série de questionnaires qui contiennent chacun 50 descripteurs. Les descripteurs inclus dans les questionnaires ont été choisis, lors de l'étape précédente, sur la base des jugements qualitatifs, c'est-à-dire de leur clarté et de leur pertinence (Conseil de l'Europe 2005 : 156). Les questionnaires, utilisés lors de deux années consécutives, ont en commun l'utilisation des mêmes descripteurs d'interaction orale (Conseil de l'Europe 2005 : 156). Cette évaluation des apprenants poursuit deux objectifs. Le premier est d'établir « l'indice de difficulté » de chaque descripteur (Conseil de l'Europe 2005 : 156). Le deuxième cherche à déterminer statistiquement la variation de l'interprétation des descripteurs par rapport aux différents secteurs éducatifs, aux régions linguistiques et aux langues cibles, pour reconnaitre les descripteurs avec un « indice de stabilité élevé » dans des contextes différents (Conseil de l'Europe 2005 : 156). Les descripteurs satisfaisant à ce critère ont été utilisés dans l'élaboration des échelles globales de compétences qui servent à résumer les échelles illustratives. Ensuite, les variations de sévérité des jugements des enseignants sont quantifiées (Conseil de l'Europe 2005 : 156). Celles-ci sont calculées en faisant évaluer les performances des apprenants dans les enregistrements vidéo par tous les enseignants participant au projet de recherche. Les différences de sévérité sont prises en compte lors de l'interprétation des résultats selon les secteurs éducatifs en Suisse (Little 2007 : 648 ; Conseil de l'Europe 2005 : 156).

Lors de la quatrième étape, celle d'interprétation, les « seuils fonctionnels sur l'échelle des descripteurs » ont été déterminés, ce qui a permis d'élaborer la série des niveaux communs de référence (Conseil de l'Europe 2005 : 156). Ensuite, les niveaux ainsi établis ont été résumés dans les trois échelles, également incluses dans le CECRL : l'échelle globale, la grille pour l'auto-évaluation et la grille pour l'évaluation de la performance orale (Conseil de l'Europe 2005 : 156). En outre, les échelles illustratives pour les catégories, susceptibles d'être étalonnées en niveaux de compétences, ont été présentées. Enfin, les descripteurs ont été adaptés à l'auto-évaluation qui a été incluse dans

la version expérimentale du Portfolio européen des Langues (Conseil de l'Europe 2005 : 156).

On peut conclure de ces quatre étapes d'élaboration des descripteurs que la collecte des données a suivi une recherche qualitative intense au cours de laquelle les groupes d'enseignants avaient la fonction d'informateurs (North 2007 : 657). L'analyse statistique a été effectuée après la phase de recueil des données qualitatives (North 2007 :657). Les descripteurs ont été étalonnés selon le modèle de Rasch pour constituer les échelles de compétences (Conseil de l'Europe 2005 : 151).[19]

2.4.2 Critique des descripteurs contenus dans les CECRL

Les descripteurs inclus dans le CECRL ont fait l'objet de critiques de la part de plusieurs experts et ce pour diverses raisons. Le premier point de critique concerne l'étalonnage des descripteurs du CECRL car il ne se fonde pas sur les données empiriques fournies par les apprenants de la langue seconde. Plusieurs experts considèrent l'absence de ce type de données comme un problème crucial des échelles de compétences (North& Schneider 1998: 238-239, Hulstijn 2007 :7). L'étalonnage des descripteurs possède cependant bien une base empirique au niveau des jugements faits par des enseignants de langues et d'autres experts par rapport à la meilleure façon de décrire les différents niveaux de performance des apprenants de manière cohérente (Byrnes 2007 : 643). La validité et la fiabilité de ces jugements ne sont pas remises en question par les critiques. Certains dénoncent le fait que les jugements des professeurs constituent un fondement trop faible pour un document de référence ayant des implications aussi importantes que le CECRL pour les politiques éducatives en Europe (North& Schneider 1998 : 238-239, Hulstijn 2007 :8). Pour cette raison, Hulstijn plaide pour des études empiriques auprès d'apprenants de langue vivante étrangère, en vue de fonder les échelles du CECRL sur de véritables données d'expérience (Hulstijn 2007 :7). Il faut souligner que l'étalonnage des descripteurs sur la base de jugements émis par des enseignants ne représente pas un argument de remise en cause de la validité empirique des descripteurs pour tous les experts. Ainsi, dans certains travaux, la recherche qualitative

effectuée par les professeurs apparaît comme un argument en faveur de leur validité empirique (North 2007 : 657). En effet, les descripteurs contenus dans le CECRL forment une banque d'items empiriquement calibrés dont les caractéristiques statistiques ont été calculées (North 2007 : 657).

Même les experts qui ne remettent pas en question le fondement empirique de toutes les échelles du CECRL soulignent pourtant le manque de base empirique d'un certain nombre de descripteurs qu'il contient (Little 2007 : 649, North 2007 : 657). Les échelles illustratives ne sont que partiellement validées par la recherche empirique car « les descripteurs de la production écrite du Chapitre 4 ont été essentiellement élaborés à partir de ceux de la production orale » (Conseil de l'Europe 2005 : 156). Ce procédé pousse Little à conclure que les descripteurs de cette catégorie ne sont pas empiriquement validés (Little 2007 :648). Selon d'autres experts, les descripteurs qui définissent cette activité de communication langagière ont pourtant un haut degré de validité (North 2007 : 657). North appuie sa thèse sur l'étude menée par Kaftandjeva & Takala (2002), selon laquelle les descripteurs utilisés par les apprenants du finlandais dans l'auto-évaluation de leur production écrite avaient une corrélation élevée avec les valeurs de l'échelle du CECRL (North 2007 : 657). North souligne qu'un certain nombre de descripteurs manquaient de base empirique parce qu'ils n'étaient pas empiriquement validés dans le projet suisse. Il s'agit des descripteurs faisant partie des quatre catégories suivantes. Premièrement, beaucoup de descripteurs du niveau C2 sont concernés, notamment ceux qui définissent les activités de communication langagière (North 2007 : 657). La deuxième catégorie manquant de base empirique est l'échelle « Maîtrise du Système Phonologique » car ce domaine de compétence a fait l'objet de différences d'interprétation considérables en fonction de la langue enseignée par le professeur (North 2007 : 657). La troisième catégorie de descripteurs empiriquement non validés est regroupée au sein de l'échelle « Maîtrise de l'orthographe » car ils ont été inclus dans le CECRL pour des raisons de complétude (North 2007 : 657). La quatrième catégorie est composée des descripteurs de l'échelle « Correction sociolinguistique » qui ont été ajoutés pour l'édition du CECRL en 2001 (North 2007 : 657). La faiblesse des descripteurs qui définissent la compétence sociolinguistique est également soulignée par d'autres chercheurs qui ne remettent pas pour autant en cause leur base empirique

(Tagliante 2007 : 41). Les concepteurs du CECRL mentionnent les difficultés d'étalonnage des descripteurs de la compétence sociolinguistique selon les différents niveaux communs de référence. Pourtant, selon ce document, les descripteurs inclus dans l'échelle « Correction sociolinguistique » ne sont pas concernés par les difficultés d'étalonnage (Conseil de l'Europe 2005 : 95).

L'étalonnage des descripteurs ne représente pas la seule raison pour mener des études empiriques avec les apprenants de langue vivante étrangère. Le besoin urgent de résultats de ces études s'explique également par la présence de trois assertions, contenues dans le CECRL, qui ne s'appuient pas sur une base empirique véritable (Hulstijn 2007 :7). La première assertion, qui est en même temps la plus explicite, est la linéarité d'acquisition de la compétence langagière qui se manifeste par la répartition de la compétence langagière en une série ascendante au sein des niveaux communs de référence.[20] La deuxième affirmation, exprimée par le CECRL, est que les apprenants à un niveau de compétence particulier peuvent effectuer toutes les tâches communicatives faisant partie des niveaux inférieurs. À cause de cette série ascendante de niveaux communs de référence, les apprenants du niveau B2, par exemple, devraient pouvoir effectuer toutes les tâches communicatives aux trois niveaux inférieurs à celui-ci. La troisième suggestion du CECRL également en manque de soutien empirique est la correspondance admise entre le niveau d'un apprenant dans les compétences fonctionnelles[21] et son niveau dans les compétences linguistiques (Hulstijn 2007: 8). Bien que cette correspondance ne soit pas explicitement affirmée dans le CECRL, l'utilisation des mêmes symboles pour marquer les niveaux communs de référence, à savoir les lettres A, B ou C, associés aux chiffres 1 ou 2, suggèrent que c'est bien le cas. Cette suggestion implique qu'un apprenant qui se trouve à un niveau particulier dans les compétences fonctionnelles se situe au même niveau dans les compétences linguistiques (Hulstijn 2007 :8).

Malgré le scepticisme exprimé à l'égard de ces assertions plus ou moins explicites contenues dans le CECRL, les experts ne demandent pas de renoncer à son utilisation. Loin de là, les critiques reconnaissent l'utilité de ce document de référence qui devrait selon eux continuer à servir à la communauté éducative et à remplir ses fonctions. Son maintien n'empêche cependant pas d'effectuer les

études nécessaires afin de fournir la base empirique manquante à ce document à l'heure actuelle (Hulstijn 2007 :8).

Les aspects qualitatifs d'un certain nombre de descripteurs au sein de plusieurs échelles de compétences ont également fait l'objet de critiques (Little 2007 : 648), à commencer par les descripteurs qui définissent la qualité « Aisance » au sein de la grille décrivant les aspects qualitatifs de la langue parlée (Conseil de l'Europe 2005 : 26). Cette qualité est largement conceptualisée en termes d'hésitation dans la communication orale, typique pour les niveaux bas de compétence, qui diminue progressivement avant de disparaître aux niveaux supérieurs. Pour illustrer ce constat, nous développerons l'exemple des descripteurs aux niveaux A1 et C1. Le descripteur A1 définit le niveau d'aisance de façon suivante : « Peut se débrouiller avec des énoncés très courts, isolés, généralement stéréotypés, avec de nombreuses pauses pour chercher ses mots, pour prononcer les moins familiers et pour remédier à la communication.» (Conseil de l'Europe 2005 : 26). Les pauses sont centrales dans la définition fournie par ce descripteur. En revanche, le descripteur au niveau C1 définit que l'apprenant « Peut s'exprimer avec aisance et spontanéité presque sans effort. Seul un sujet conceptuellement difficile est susceptible de gêner le flot naturel et fluide du discours. » (Conseil de l'Europe 2005 : 26). La fluidité figure en tant que qualité essentielle de l'aisance du discours à ce niveau de compétence. Les définitions, données par les descripteurs, ne tiennent pas compte du fait que les locuteurs natifs peuvent aussi hésiter fréquemment en communiquant, ce qui n'empêche pas leur production fluide du discours (Little 2007 : 648).

Les descripteurs au sein de l'échelle « Maîtrise du système phonologique » ont été critiqués à cause de leur approximation progressive concernant les normes de prononciation des locuteurs natifs (Little 2007 : 648). Citons en exemple les descripteurs aux niveaux A1, B1 et C1. La compétence au niveau le plus bas est définie ainsi : « La prononciation d'un répertoire très limité d'expressions et de mots mémorisés est compréhensible avec quelque effort pour un locuteur natif habitué aux locuteurs du groupe linguistique de l'apprenant/utilisateur » (Conseil de l'Europe 2005 : 92). L'expression centrale ici est « compréhensible avec quelque effort par un locuteur natif ». Au niveau B1, la notion d'effort n'apparaît plus et « la prononciation est clairement intelligible »

malgré « un accent étranger […] quelquefois perceptible » et malgré « des erreurs de prononciation » occasionnelles (Conseil de l'Europe 2005 : 92). Au niveau C1, l'apprenant « peut varier l'intonation et placer l'accent phrastique correctement afin d'exprimer de fines nuances de sens » (Conseil de l'Europe 2005 : 92). La proximité avec les normes phonologiques des locuteurs natifs, révélée par ces descripteurs, va à l'encontre du principe pédagogique selon lequel il est indispensable d'assurer la correction phonologique dès le début de l'apprentissage d'une langue (Little 2007 :649).

En outre, les utilisateurs ont dénoncé la définition des compétences en langue générale, malgré la présence des sous-échelles pour les activités de communication spécifiques dans toutes les activités productives et réceptives de communication langagière (North 2007 : 657). De même, le manque de descripteurs pour les aspects socioculturels, pour la lecture littéraire ainsi que pour la médiation ont été critiqués. Alors qu'on a essayé d'élaborer des descripteurs pour les deux premières catégories évoquées, on n'a pas tenté de les développer pour les activités de médiation (North 2007 : 657).

2.4.2.1 Opacité des descripteurs

Un autre point de critique concerne l'opacité d'un certain nombre de descripteurs. On note ainsi une spécification insuffisante des connaissances lexicales et grammaticales relatives aux différents niveaux communs de référence (Westhoff 2007 : 676). En effet, le manque de ce genre de spécification est manifeste au sein de toutes les échelles de compétences, aussi bien de celles qui définissent les compétences fonctionnelles que les échelles se focalisant sur les compétences linguistiques. L'absence de spécifications au sein des échelles de compétences fonctionnelles paraît logique, étant donné que leur but est de définir les tâches communicatives susceptibles d'être maîtrisées aux différents niveaux de compétence langagière. En revanche, ce fait peut étonner pour les échelles de compétences linguistiques car on s'attend à y trouver des spécifications des connaissances linguistiques. Pourtant, l'absence de ces spécifications au sein des descripteurs est normale car leur fonction ne consiste pas à énumérer « les contenus linguistiques ou pragmatiques » lesquels font partie des compétences à chacun des niveaux communs de référence (Tagliante 2005 : 61). Ce rôle est rempli par les référentiels officiels qui définissent les savoirs grammaticaux,

lexicaux, phonologiques et (socio)culturels à maîtriser aux différents niveaux communs de référence (Tagliante 2005 : 61). Dans les *Instructions pour les langues vivantes pour le lycée et le collège*, ces quatre contenus sont inclus. Bien qu'il s'agisse de savoirs linguistiques et culturels, ceux-ci sont désignés en tant que compétences dans le référentiel, conformément au CECRL : compétence culturelle et lexicale, compétence grammaticale et compétence phonologique. [22]

L'absence de spécification des contenus linguistiques attendus dans la performance a pour effet que les descripteurs ne conviennent souvent pas à l'observation de la performance, bien qu'ils définissent ce que les apprenants sont capables de faire aux différents niveaux (Tagliante 2005 : 61). Leur caractère, souvent insuffisant pour l'observation de la performance les rend inaptes à évaluer celle-ci. Ce lien entre l'observation et l'évaluation de la performance s'explique par le fait qu'il est indispensable d'observer la performance avant de pouvoir évaluer cette dernière (Tagliante 2005 : 61). Par conséquent, il faut souvent combiner l'utilisation des descripteurs et les contenus spécifiés dans les référentiels lors de l'évaluation de la performance, comme indiqué dans les Instructions Officielles (Tagliante 2005 : 61).

Pour illustrer le manque de spécifications des contenus linguistiques, nous analyserons plusieurs descripteurs, inclus dans les échelles de compétences linguistiques. En ce qui concerne la compétence lexicale, le descripteur A1, situé en bas de l'échelle « Étendue de vocabulaire », définit que l'apprenant « possède un vaste répertoire de mots isolés et d'expressions relatifs à des situations concrètes particulières » (Conseil de l'Europe 2005 : 88). Puisque ce « vaste répertoire » n'est pas précisé par des exemples de « mots isolés et d'expressions » qui en font partie, le descripteur reste opaque. L'opacité des descripteurs est évidente aussi au sein de l'échelle « Maîtrise de vocabulaire ». C'est le cas de la définition du niveau de compétence A2 qui « Possède un répertoire restreint ayant trait à des besoins quotidiens concrets » (Conseil de l'Europe 2005 : 88). Comme dans le descripteur cité auparavant, le «répertoire restreint » n'est pas explicité par des exemples.

Quant à la compétence grammaticale, elle est considérée comme centrale pour le développement de la compétence communicative par le CECRL : « La compétence grammaticale, ou capacité d'organiser des phrases pour transmettre

un sens, est au centre même de la compétence communicative […] » (Conseil de l'Europe 2005 : 115). Néanmoins, cette compétence est traitée en termes très généraux par ce document (Westhoff 2007 : 676). En effet, il manque des précisions sur le lien entre les niveaux communs de référence et les items grammaticaux, dans le sens où le genre d'items à maîtriser aux différents niveaux de compétence n'est pas détaillé (Westhoff 2007 : 676). Un deuxième aspect renforce l'opacité de description de la compétence grammaticale dans ce document de référence. L'échelle des descripteurs « Correction grammaticale » qui établit bien le lien entre la correction grammaticale et les niveaux communs de référence n'indique pas clairement si la correction est fondée sur la connaissance des règles grammaticales et leur application ou bien sur la connaissance mémorisée des collocations et autres syntagmes langagiers (Westhoff 2007 : 676). L'exemple du descripteur du niveau A1 peut illustrer l'opacité de la définition de la compétence grammaticale : l'apprenant « a un contrôle limité de structures syntaxiques et de formes grammaticales simples appartenant à un répertoire mémorisé » (Conseil de l'Europe 2005 : 90). On se rend compte que ce descripteur manque de précision à propos des « structures syntaxiques et des formes grammaticales » à connaître, à l'exception de l'indication qu'elles doivent être « simples ». Ce qualificatif ne saurait compenser l'opacité causée par le manque de précisions grammaticales. L'opacité de ce descripteur est encore renforcée par l'expression floue « un contrôle limité » qui ne renseigne pas dans quelle mesure les formes et les structures sont à maîtriser.

Malgré le caractère très général et relativement imprécis de la description de la compétence grammaticale par le CECRL, les descripteurs fournissent certaines indications utiles sur les connaissances grammaticales attendues aux différents niveaux communs de référence (Westhoff 2007 : 677). Aux niveaux inférieurs, la correction formelle repose principalement sur la maîtrise des formes lexicales. Pour cette raison, le développement du répertoire lexical est accentué par les descripteurs des niveaux inférieurs au B2 (Westhoff 2007 : 677). Cette perspective se manifeste par le descripteur évoquant un « répertoire mémorisé » au niveau A1 : « A un contrôle limité de structures syntaxiques et de formes grammaticales simples appartenant à un répertoire mémorisé » (Conseil de l'Europe 2005 : 90). La notion de répertoire lexical réapparait dans le descripteur

du niveau B1. Il ne s'agit cependant plus « d'un répertoire mémorisé » mais « d'un répertoire de tournures et d'expressions fréquemment utilisées et associées à des situations plutôt prévisibles » (Conseil de l'Europe 2005 : 90). La communication à l'aide d'un répertoire lexical, même si ce dernier n'est pas mémorisé, n'exige pas la compétence d'utiliser consciemment les règles grammaticales dans le but de communiquer. L'usage conscient des règles grammaticales apparaît seulement dans les descripteurs définissant la compétence à partir du niveau B2 (Westhoff 2007 : 667). Ainsi, aux niveaux B2 et B2+, il est question respectivement d'un « assez bon contrôle grammatical » et d'un « bon contrôle grammatical » (Conseil de l'Europe 2005 : 90). Les niveaux C1 et C2 parlent d'un « haut degré de correction grammaticale » et « d'un haut niveau » de cette compétence (Conseil de l'Europe 2005 : 90).

2.4.2.2 Les descripteurs des niveaux avancés

Les descripteurs des niveaux avancés du CECRL sont aussi concernés par les critiques portant sur des enjeux particuliers. La maîtrise des compétences, définies par ces descripteurs ne peut être que difficilement évaluée par des locutions verbales comme « a conscience de », « apprécie », « tient compte de », « est conscient de » (Tagliante 2005 : 41). On le voit par exemple dans le descripteur qui définit la compétence sociolinguistique au niveau C2 : « Apprécie complètement les implications sociolinguistiques et socioculturelles de la langue utilisée par les locuteurs natifs et peut réagir en conséquence » (Conseil de l'Europe 2005 : 95).

Le deuxième enjeu délicat des descripteurs à ce niveau concerne leur domaine d'usage restrictif qui comprend seulement les apprenants et les usagers d'une langue étrangère ayant parfaitement réussi leur apprentissage et aucunement les locuteurs natifs. De ce fait, C2 est le niveau le plus élevé susceptible d'être ciblé par les établissements d'enseignement et de certification (Tagliante 2005 : 41). Il n'empêche qu'un niveau supérieur à C2 puisse être visé par les institutions qui s'adressent au public ayant besoin d'utiliser la langue à des fins particulières. (Tagliante 2005 : 41).

2.4.3 Dimensions qualitative et quantitative des descripteurs

La notion de compétence langagière présentée à l'aide des descripteurs repose sur deux piliers étroitement liés : la qualité et la quantité. La notion de quantité relative à la compétence langagière implique un certain nombre de domaines, de fonctions, de situations, de sujets et de rôles maîtrisés par l'usager de la langue (De Jong 2004). Quant à la notion de qualité, elle a un double sens. Premièrement, elle renvoie au degré d'effectivité d'usage de la langue qui se constitue par la précision d'expression et de compréhension du sens transmis. Deuxièmement, la notion de qualité fait référence au degré d'efficacité d'usage langagier. Cette dernière se manifeste par la communication avec le moindre effort possible (De Jong 2004). Ces notions peuvent être illustrées par l'exemple du descripteur B2 de l'échelle « Production orale générale » :

> Peut faire une description et une présentation détaillées sur une gamme étendue des sujets relatifs à son domaine d'intérêt en développant et justifiant les idées par des points secondaires et des exemples pertinents (Conseil de l'Europe 2005 : 49).

Dans ce descripteur, les éléments de quantité sont exprimés par : « une gamme étendue de sujets relatifs à son domaine d'intérêt » (Conseil de l'Europe 2005 : 49). C'est le cas parce que l'extrait cité se focalise sur ce que l'apprenant peut faire (Hulstijn 2007 : 2). En revanche, les éléments de qualité sont signalés par : « une description et une présentation détaillées […] en développant et justifiant les idées par des points secondaires et des exemples pertinents » (CECRL 2005 : 49). Il faut noter que la notion de qualité contient un seul sens dans ce descripteur, à savoir l'effectivité, car il n'y est pas question de l'efficacité.

La combinaison des notions de qualité et de quantité ne concerne pas seulement les descripteurs qui servent à expliciter les niveaux communs de référence dans les différentes activités de communication langagière. [23] Les descripteurs explicitant les différents niveaux au sein des compétences communicatives langagières sont construits selon le même principe (Hulstijn 2007 : 2), par exemple pour le descripteur du niveau B2 de l'échelle « Étendue linguistique générale » :

> Possède une gamme assez étendue de langue pour pouvoir faire des descriptions claires, exprimer son point de vue et développer une argumentation sans chercher ses

mots de manière évidente et en utilisant des phrases complexes (Conseil de l'Europe 2005 : 87).

On peut ici reconnaitre la combinaison des dimensions quantitative et qualitative. La dimension quantitative est ainsi exprimée par : « […] une gamme assez étendue de langue », « pour pouvoir faire des descriptions », « exprimer son point de vue » et « développer une argumentation ». Tandis que le premier syntagme cité renvoie au répertoire langagier maîtrisé, les trois extraits qui le suivent se réfèrent aux fonctions langagières attendues à ce niveau de compétence.

En ce qui concerne la dimension qualitative, celle-ci est exprimée par les syntagmes suivants : « descriptions claires », « en utilisant des phrases complexes », « sans chercher ses mots de manière évidente » (Conseil de l'Europe 2005 : 87). Les deux premières citations font référence à la première notion contenue dans la dimension qualitative, à savoir l'effectivité, qui sert des fins de précision d'expression dans la manière d'utiliser la langue. En revanche, « sans chercher ses mots de manière évidente » renvoie à la notion d'efficacité, car elle contient l'idée de communiquer avec le moindre effort possible.

Par rapport à la question de l'équilibre entre les dimensions qualitative et quantitative de la compétence langagière, Hulstijn propose de s'imaginer trois cas de figure, en précisant qu'ils ne relèvent pas nécessairement de la réalité (Hulstijn 2007 : 6). Le premier cas de figure concerne les apprenants qui sont situés à un bas niveau quantitatif, en raison du peu de tâches communicatives qu'ils sont en mesure d'effectuer. En revanche, leur performance langagière est marquée par une qualité linguistique élevée, en raison du haut niveau d'effectivité et d'efficacité (Hulstijn 2007 : 3). Le deuxième cas de figure renvoie aux apprenants et usagers d'une langue dont la performance fait preuve de haute quantité, mais d'une qualité linguistique basse. Malgré un grand nombre de tâches communicatives effectuées, leur qualité est pauvre à cause d'un bas niveau d'effectivité et d'efficacité (Hulstijn 2007 : 3). Troisièmement, il y a les apprenants et les usagers d'une langue dont le niveau de quantité égale celui de la qualité de leur performance (Hulstijn 2007 : 3). Aux trois niveaux inférieurs du CECRL, lors des activités de production orale et de compréhension orale, le

niveau de quantité et celui de qualité des apprenants est typiquement déséquilibré, le niveau de qualité étant inférieur à celui de quantité (Hulstijn 2007 : 7). Ce déséquilibre s'explique par le manque habituel de pratique de la langue étrangère lors de l'apprentissage. Les descripteurs du CECRL reflètent d'ailleurs cet état de fait (Hulstijn 2007 : 7). Cependant, ce document manque de données empiriques qui montrent quel niveau minimal de connaissances et de compétences est nécessaire pour assurer les fonctions adéquates de la performance (Hulstijn 2007 : 7).

En outre, la combinaison de la dimension quantitative et qualitative dans les descripteurs met en évidence que, dans la compétence langagière, les éléments de quantité sont toujours associés à ceux de qualité. Ce constat fait écho à l'enjeu fondamental qui fait toujours l'objet de recherches, à savoir la relation entre les formes et les fonctions dans une langue (Hulstijn 2007 : 4).

2.4.4 Formulation des descripteurs

L'opacité des descripteurs dénoncée par certains chercheurs comme Byrnes (2007 : 643), est liée à leur formulation. Celle-ci ne correspond pas systématiquement aux lignes directrices existantes pour l'élaboration des descripteurs. Ces lignes directrices résultent de la théorie et de l'expérience de l'étalonnage en évaluation des langues, ainsi que des opinions des enseignants dans les différents projets de recherche, y compris dans le projet du *Fonds national suisse de la recherche scientifique* (Conseil de l'Europe 2005 : 148).

Premièrement, les descripteurs doivent être systématiquement formulés de manière positive (Conseil de l'Europe 2005 : 148). La formulation positive au moyen de *can-do statements* est respectée par tous les descripteurs, même par ceux qui définissent les compétences aux niveaux inférieurs (North 2007 : 657). Certains descripteurs formulés de cette manière incluent les conditions dans lesquelles la compétence peut se manifester, par exemple, le descripteur A2 de la sous-échelle « S'adresser à un auditoire» : « Peut répondre aux questions qui suivent si elles sont simples et directes et à condition de pouvoir faire répéter et se faire aider pour formuler une réponse » (Conseil de l'Europe 2005 : 50).

Deuxièmement, les descripteurs doivent être précis. La précision se montre par la description concrète des caractéristiques de performance et des

tâches à réaliser (Conseil de l'Europe 2005 : 149). Les formulations floues sont à éviter. Cette ligne directrice n'est pourtant pas respectée par tous les descripteurs. Plusieurs sont ainsi formulés de manière imprécise. Les descripteurs qui incluent le mot « gamme » sont flous car le sens exprimé n'est pas clair. Or, cette désignation est incluse dans les définitions de nombreux descripteurs, appartenant aux différentes échelles de compétences, par exemple, le descripteur B1 au sein de l'échelle « Étendue linguistique générale » et le descripteur B2 de l'échelle « Étendue de vocabulaire » (Conseil de l'Europe 2005 : 88).

Troisièmement, les descripteurs doivent être clairs (Conseil de l'Europe 2005 : 149). La clarté doit être respectée tant au niveau du contenu que de la formulation langagière. La clarté du contenu tient à la présence de liens logiques explicites cependant que la clarté de la formulation se manifeste par le choix de phrases simples. Malgré la mise en œuvre de cette ligne directrice, l'usage d'une langue compliquée dans les descripteurs a été dénoncé (North 2007 : 657).

Le critère suivant concernant la formulation des descripteurs est la brièveté. En proposant un nombre limité de traits isolés, ces descripteurs procurent deux avantages considérables. Premièrement, ils définissent des tâches typiques qui sont à accomplir par les apprenants aux différents niveaux de compétence. Deuxièmement, il est primordial pour les enseignants de se référer à des descripteurs brefs lors des opérations d'évaluation (Conseil de l'Europe 2005 : 149).

La dernière ligne directrice concerne leur indépendance, qui est essentielle (Conseil de l'Europe 2005 : 150). En effet, les descripteurs dans le CECRL satisfont à ce critère car chaque descripteur présente un critère distinct qui est défini de manière indépendante des autres descripteurs (North 2007 : 657). Cette caractéristique est très importante pour assurer que les descripteurs expriment bien un nouvel aspect de compétence et peuvent ainsi servir en tant qu'objectifs sans dépendre d'autres descripteurs (Conseil de l'Europe 2005 : 150).

Au-delà de ces lignes directrices, évoquées explicitement dans le Cadre européen, les descripteurs satisfont à deux autres critères qui ne sont pas explicitement nommés mais observables à partir de leur formulation. Premièrement, ils sont globaux dans la mesure où ils ne comportent pas de listes

de phénomènes grammaticaux et lexicaux, de connaissances culturelles et de capacités cognitives qui sont à maîtriser aux différents niveaux de compétence (Lallement 2007 : 28). Ces indications se trouvent dans les programmes d'enseignement et dans d'autres référentiels pédagogiques (Lallement 2007 : 28). En ce qui concerne le système d'enseignement français, ces listes se trouvent dans les *Instructions Officielles* publiées par le Ministère de l'Éducation Nationale. Deuxièmement, les descripteurs conviennent à tous les niveaux d'enseignement et d'apprentissage des langues, qui commence à l'école élémentaire et qui se poursuit tout au long de l'enseignement secondaire (Lallement 2007 : 28). Ils sont adaptés aux différentes étapes d'apprentissage parce qu'ils présentent des objectifs réalistes à poursuivre (Lallement 2007 : 28).

Il faut noter que la formulation des descripteurs ne se fonde pas sur la recherche en acquisition des langues, mais sur l'examen des performances par les enseignants. Ce sont ces performances qui ont été analysées pour identifier les concepts et les formulations utilisées (North 2007 : 657). Ainsi, les descripteurs sont formulés afin de correspondre aux perceptions de la compétence langagière par les professeurs. La raison de la non-prise en compte des résultats de la recherche en acquisition de la langue seconde est que cette dernière n'était pas en mesure de fournir des définitions de compétence à l'époque du développement du CECRL (North 2007 : 657).

2.5 Usage du CECRL

2.5.1 Influence du CECRL sur la politique éducative

Le grand impact du CECRL sur la politique et la pratique éducatives est dû à son caractère supranational et au long travail de conception, de négociation et d'implémentation fourni par le Conseil de l'Europe avant la mise en œuvre de ce document (Byrnes 2007 : 644). L'adoption du CECRL a été promue par une autre institution politique, la Commission Européenne. Cette organisation accorde les mêmes valeurs sociales, culturelles et éducatives à l'apprentissage des langues que le Conseil de l'Europe, mais elle y reconnait également une valeur économique (Little 2007 : 647). La compétence plurilingue est en effet considérée comme un besoin dans la poursuite d'objectifs économiques (Little 2007 : 647). La Commission Européenne a entrepris plusieurs démarches

destinées à informer les citoyens européens de l'existence du CECRL et ainsi permettre de promouvoir son acceptation et d'élargir son influence sur la politique et les pratiques éducatives en langues vivantes étrangères. Premièrement, cette organisation a financé des projets ciblant la mise en œuvre du CECRL dans les différents domaines dont DIALANG constitue un exemple (Little 2007 : 647). Deuxièmement, la Commission Européenne a inclus la Grille pour l'auto-évaluation dans le passeport européen. Ce dernier se compose d'un certain nombre de documents censés permettre aux citoyens européens de rédiger leurs expériences et qualifications à un format standardisé (Little 2007 : 647). Troisièmement, cette institution a décidé de développer l'Indicateur européen de la compétence langagière dont la fonction est de relier la performance des apprenants aux niveaux du CECRL. La décision d'élaborer ce document semble particulièrement prometteuse car elle oblige les États européens à tenir compte des niveaux communs de référence du CECRL lors du développement de leur politique d'éducation linguistique (Little 2007 : 647). La politique et la pratique menées par la Commission Européenne se sont révélées efficaces car le CECRL a reçu une grande attention de la part des enseignants européens. Sa traduction dans plus de 20 langues témoigne de la réussite de la promotion de ce document (Alderson 2004 : 20).

En conséquence de ces démarches entreprises par la Commission Européenne, les gouvernements de certains États membres de l'Union Européenne ont adopté des lois accordant à ce document de référence une grande influence sur la politique éducative de leurs pays. La France est un excellent exemple de la grande influence du CECRL sur la politique éducative d'un pays. En effet, la politique française en matière d'enseignement des langues vivantes est liée au CECRL depuis 2007. Depuis lors, ce document constitue le socle de l'enseignement et de l'apprentissage des langues tout au long de l'éducation du premier et du second degré. La reconnaissance formelle de ce référentiel en tant que fondement de l'éducation nationale en langues vivantes a donné lieu à des spécifications sur la question des niveaux de compétence du CECRL à atteindre aux différents niveaux de scolarité.[24] Au-delà des spécifications arrêtées et publiées par le Ministère de l'Éducation nationale, ce dernier souscrit également aux valeurs exprimées par le CECRL[25]. On le voit par exemple à la possibilité pour les étudiants des « classes européennes » de

passer des examens administrés par les autorités éducatives d'autres pays de l'Union Européenne [26].

Les décisions et les mesures prises par le Ministère de l'Éducation nationale sont des indicateurs importants de l'influence toujours croissante du CECRL. L'impact de plus en plus fort de ce document s'explique, bien sûr, par les valeurs idéologiques qu'il représente, mais aussi par la réalité du marché du travail caractérisé par une intégration croissante. Cette dernière implique la migration professionnelle entre les pays au sein de l'Union Européenne rendant indispensables la reconnaissance réciproque des qualifications en langues et, par conséquent, la coopération entre les établissements d'enseignement à travers les différents pays (Byrnes 2007 : 644). Les mesures saisies par les autorités éducatives françaises expriment également la confiance en la fiabilité du CECRL ainsi que les bénéfices à atteindre par l'adossement de la politique éducative nationale à ce document (Byrnes 2007 : 644).

2.5.2 Les enjeux sensibles de l'usage du CECRL

L'influence croissante du CECRL pose la question des rapports de force entre ce document supranational et les politiques éducatives nationales. En effet, les traditions éducatives sont fondamentales pour les identités culturelles et politiques de beaucoup de pays au sein de l'Union Européenne (Byrnes 2007 : 647). Les États membres décident de façon souveraine si et à quel degré le CECRL va aider à déterminer leur politique éducative ainsi que l'enseignement, l'apprentissage et l'évaluation en langues (Little 2007 : 647). Le Conseil de l'Europe ne peut pas intervenir directement dans la politique et la pratique éducatives des États membres de l'Union Européenne, et cette institution ne remet pas non plus en question le rôle déterminant des pays européens dans ce domaine (Van Ek 2001 : 1).

Le Conseil de l'Europe souligne sa fonction d'aide et de soutien aux États individuels qui définissent leurs politiques linguistiques eux-mêmes : « La Division aide les États membres à revoir leurs politiques linguistiques éducatives en vue de promouvoir la diversité linguistique et le plurilinguisme » (Conseil de l'Europe 2009 : 1). Le passage cité montre que l'aide proposée aux États membres a pour but de leur permettre d'adopter des mesures efficaces afin de

développer le plurilinguisme de leurs ressortissants (Van Ek 2001 : 1). Il n'est donc pas question pour le Conseil de l'Europe d'imposer sa politique linguistique aux États individuels (Van Ek 2001 : 1).

Du fait de la question de la souveraineté des États membres de l'Union Européenne en matière d'éducation, cruciale pour nombre d'entre eux, l'influence croissante du CECRL dans ce domaine constitue un enjeu très sensible. Il est d'autant plus sensible que la pression des responsables dans la Division des Politiques Linguistiques d'intervenir davantage dans les politiques éducatives nationales ne cesse d'augmenter (Byrnes 2007 : 644). Cette question du pouvoir du CECRL peut seulement être résolue par les responsables de la politique éducative au sein du Conseil de l'Europe (Byrnes 2007 : 644).

La question des rapports de pouvoir entre le CECRL et les États individuels n'est pas le seul enjeu sensible lié à l'usage de ce document. L'emploi du CECRL a donné lieu à des effets indésirables qui doivent être analysés de façon critique par les responsables dans le but de corriger les effets pervers décelés. Un défaut courant est l'usage inadapté du référentiel, en rupture avec les intentions originales des concepteurs. Bien que ce document ne soit pas prévu pour tous les publics de manière uniforme, il est utilisé par exemple dans le cas des migrants. Or ce public n'était pas visé par le CECRL à l'époque de sa conception (Krumm 2007 : 667). L'application de ce référentiel à ce public non ciblé a pour effet d'invalider les intentions originales du document de référence (Krumm 2007 : 667).

L'inadaptation du CECRL à évaluer les compétences langagières des migrants s'explique aussi bien par son usage actuel dans plusieurs États membres de l'Union Européenne que par son contenu (Krumm 2007 : 667). Son utilisation contreproductive consiste en son rôle dans la prise de décision concernant le statut des migrants. Ainsi, dans plusieurs pays de l'Union Européenne, le droit de résidence ou la naturalisation sont liés aux obligations de suivre un cours de langues et de passer un examen afin d'attester du niveau de compétence demandé par les autorités administratives (Krumm 2007 : 668). Cette manière d'utiliser le CECRL va à l'encontre des fonctions attribuées à l'apprentissage des langues par le Conseil de l'Europe parce que dans ce cas, l'apprentissage des langues ne remplit pas sa fonction de promotion de la compréhension mutuelle et de l'intégration sociale, mais agit en tant qu'outil de

ségrégation (Krumm 2007 : 668). Un tel rôle assigné à l'apprentissage des langues constitue un risque et peut mener à la démotivation du groupe d'apprenants (Krumm 2007 : 668). Il est évident que l'usage inadapté du CECRL n'est pas seulement un enjeu scientifique ou linguistique, mais qu'il peut créer un grand nombre de problèmes d'ordre politique, social et humain.

A part cet emploi inapproprié du CECRL, le contenu du document lui-même n'est pas adapté aux migrants qui, dans leur majorité, se distinguent nettement des populations autochtones par leur socialisation, leurs habitudes et attitudes culturelles ainsi que par les réalités et les besoins ressentis dans leur nouveau cadre de vie (Krumm 2007 : 668). Or, un grand nombre de descripteurs dans les échelles du CECRL sont très éloignés des contextes sociaux et culturels propres aux migrants qui restent mal connus pour cette partie de la population. Ainsi, les migrants doivent en priorité utiliser la langue de leur nouveau pays de résidence dans des contextes vocationnels et administratifs. Les besoins auxquels cette partie de la population est confrontée ne font actuellement pas partie des échelles de compétences du CECRL. Nous présenterons deux exemples pour comprendre dans quelle mesure certains descripteurs sont inadaptés aux contextes sociaux familiers des migrants. Le premier descripteur définit la compétence au niveau A1 de la sous-échelle « S'adresser à un auditoire », comme un paramètre de la production orale : « Peut lire un texte très bref et répété, par exemple, pour présenter un conférencier, proposer un toast » (Conseil de l'Europe 2005 : 50). Le contexte évoqué par ce descripteur est très éloigné des contextes sociaux et culturels des migrants dans leur grande majorité. Il faut remarquer que même les descripteurs qui évoquent un contexte social plus général que ce dernier sont souvent également détachés des besoins et des capacités des migrants (Krumm 2007 : 668). Ils ont, par exemple, rarement l'occasion de poser des questions informelles aux membres de la population autochtone dans leur nouvelle langue ce qui représente une compétence définie par le descripteur A1 dans l'échelle de compétences « Interaction orale générale » (Conseil de l'Europe 2005 : 50). Ce fait révèle la nécessité d'adapter les descripteurs aux besoins et aux capacités spécifiques des migrants au lieu d'évaluer leur niveau de compétence général (Krumm 2007 : 668). Il est indispensable de le faire non seulement pour reconnaitre les différences multiples des populations concernées et leur accorder avec équité

des droits civils, mais également au nom de l'égalité dans la société (Krumm 2007 : 668).[27] La différence essentielle des migrants qui doit être prise en compte est leur compétence plurilingue. Elle peut seulement être testée en cas d'évaluation de leurs compétences dans plusieurs langues maîtrisées, contrairement aux procédures d'évaluation monolingues courantes à l'heure actuelle (Krumm 2007 : 669).

L'évaluation d'un niveau de compétences général, contreproductive pour les raisons évoquées, se manifeste également par le fait que dans la plupart des tests destinés aux migrants, un même niveau doit être atteint dans toutes les compétences évaluées. L'attente véhiculée par ce genre de tests empêche, premièrement, d'intégrer des personnes qui pourraient l'être, tout en attestant de niveaux variés dans les différentes compétences langagières (Krumm 2007 : 668). Deuxièmement, ce procédé contredit le principe de la « compétence partielle » évoqué dans le CECRL (Krumm 2007 : 668). Cette compétence partielle est un fait accepté dans l'apprentissage des langues et qui, de plus, remplit des fonctions attribuées par rapport à un objectif qu'on se fixe :

> […] il ne s'agit pas de se satisfaire, par principe ou par réalisme, de la mise en place d'une maîtrise limitée ou sectorisée d'une langue étrangère par un apprenant, mais bien de poser que cette maîtrise, imparfaite à un moment donné, fait partie d'une compétence plurilingue qu'elle enrichit. Il s'agit aussi de préciser que cette compétence dite « partielle », inscrite dans une compétence plurielle, est en même temps une compétence fonctionnelle par rapport à un objectif délimité que l'on se donne. (Conseil de l'Europe 2005 : 105).

Pour faire cesser ces dérives dans l'usage du CECRL, il est urgent d'adapter ce document au contexte de migration et de développer son potentiel d'évaluation des compétences plurilingues et partielles. Le Conseil de l'Europe a établi un groupe de travail en 2006 qui se penche sur les politiques linguistiques pour l'intégration des adultes issus d'immigration (Krumm 2007 : 669). En se focalisant sur cet enjeu sensible, le Conseil de l'Europe manifeste ainsi sa volonté de transformer le CECRL, jusqu'alors instrument privilégié et uniforme, en un outil adapté à des profils linguistiques très variés au sein de toutes les populations de l'Union Européenne (Krumm 2007 : 669).

2.5.2.1 La question de la validité

Le problème de la validité est un autre enjeu sensible qui doit être résolu par le CECRL. En effet, la question de la validité est pertinente pour tout cadre commun

de référence, utilisé comme un document générique inspirant des applications locales spécifiques (North 2007 : 658). Les échelles de compétences du CECRL doivent, d'une part, être détachées du contexte afin de pouvoir fournir des paramètres généralisables à différents contextes spécifiques, mais, d'autre part, les descripteurs au sein des échelles doivent être transférables dans tous les contextes pertinents ainsi qu'appropriés aux fonctions remplies par ces contextes. En termes pratiques, cette condition implique, en premier lieu, que le schéma descriptif du CECRL et les descripteurs se réfèrent à la répartition de la compétence en catégories, entreprise par les théories de la compétence langagière. Or, les théories qui existent ne sont pas adéquates pour fournir une base à cela (North 2007: 658). En deuxième lieu, le schéma descriptif et les descripteurs doivent convenir aux contextes des apprenants cibles malgré l'impossibilité de les prédire avec certitude. En troisième lieu, ils doivent être faciles à comprendre par les praticiens dans les domaines d'enseignement et d'évaluation langagière (North 2007: 658). Les descripteurs des échelles illustratives du CECRL essaient de satisfaire à ces critères, mais selon North, ils remplissent les conditions évoquées seulement en combinaison avec les descripteurs adaptés aux contextes spécifiques inclus dans la banque d'items développée pour le *Portfolio Européen des Langues* (North 2007 : 658).

Le deuxième enjeu de validité, central pour le CECRL, est l'hypothèse du niveau de difficulté identique des éléments langagiers situés à un même niveau sur l'échelle de compétences. Cette hypothèse est liée à l'enjeu de la validité de l'échelle globale de compétences. Or, cette qualité n'a pas encore été démontrée de façon empirique en raison du manque de résultats provenant de la recherche en acquisition de la langue seconde (North 2007 : 658). En revanche, la plupart des descripteurs au sein des échelles illustratives sont empiriquement validés. Cela implique que les descripteurs concernés ont le même niveau de difficulté, quels que soient les langues cibles, les communautés linguistiques et les secteurs éducatifs (North 2007 : 658).

2.5.3 Le ratio de l'influence du CECRL sur l'enseignement et l'apprentissage des langues et sur l'évaluation des compétences

A l'heure actuelle, les niveaux communs de référence sont bien ancrés dans les systèmes d'éducation européens car ceux-ci doivent fonder leurs programmes et leurs procédures d'évaluation des compétences sur des normes communes. De plus, les institutions d'éducation européennes sont obligées de rendre compte des acquis des étudiants aux agences responsables (Brown 2010 : 87). L'agence chargée des évaluations standardisées des élèves en Europe est l'Agence exécutive Éducation, audiovisuel et culture, qui travaille pour la Commission Européenne. [28]

Malgré l'influence des niveaux communs de référence du CECRL sur l'éducation langagière en général, ce document exerce une plus grande influence sur l'évaluation en langues que sur l'enseignement et l'apprentissage des langues vivantes (Westhoff 2007 :676). Les processus impliqués dans l'alignement des tests en langues avec les niveaux communs de référence du CECRL ont plus attiré l'attention des experts que d'autres aspects inclus dans ce document (Conseil de l'Europe 2011 : 6). La priorité accordée à cette composante du Cadre européen commun de référence se manifeste par la conception et la publication d'un certain nombre d'outils destinés aux responsables de l'évaluation en langues. Parmi ces documents figurent *Le Manuel pour relier les examens de langues au Cadre européen commun de référence pour les langues (CECR),* un *Supplément de Référence* (technique) au M*anuel pour Relier les examens de langues au Cadre européen commun de référence pour les langues (CECR),* un *Manuel pour l'élaboration et la passation de tests et d'examens de langues*, des illustrations des niveaux européens de compétences en langues, des grilles d'analyse de contenu pour la production orale et écrite, des actes d'un colloque sur la recherche concernant la standardisation des niveaux de référence (DNR) pour les langues nationales et régionales (Conseil de l'Europe 2011 : 6). Au-delà de l'influence du CECRL sur l'évaluation, la publication de ces documents démontre son importance durable pour les activités de la recherche. En effet, le Cadre européen commun de référence est devenu le modèle fondamental pour la recherche sur les tests en langues étrangères (Alderson & Banerjee 2002 : 81). Malgré les critiques qui ont

pu être formulées et dont nous nous sommes fait l'écho, ce document constitue une base incontestable pour la recherche sur la conception et l'évaluation des tests (Alderson & Banerjee 2002 : 81).

L'impact croissant du Cadre européen commun de référence dans le domaine de l'évaluation en langues se manifeste également par le fait qu'un grand nombre de tests commerciaux disponibles sur le marché avant la conception du CECRL ont été associés aux niveaux communs de référence après la publication de ce document (Little 2007 : 648). Cet adossement rapide des tests au CECRL est critiqué par certains chercheurs qui dénoncent un procédé prématuré dans beaucoup de cas (Little 2007 : 648). La critique s'explique aussi par la difficulté souvent présente d'établir des liens valables et fiables entre les tests langagiers disponibles et les niveaux communs de référence, malgré l'intention annoncée d'adosser les dispositifs d'évaluation au CECRL (Alderson 2004 : 21). Selon les concepteurs de la grille d'analyse d'items de compréhension écrite et orale, qui est l'instrument destiné à cet usage, il est dans l'intérêt des équipes d'auteurs de pallier les difficultés de relier leurs dispositifs d'évaluation aux niveaux de compétences du CECRL. Les liens gagnent en effet à devenir transparents pour tous les utilisateurs (Alderson 2004 : 21). Ce besoin d'établir des liens valables et fiables explique les appels lancés au Conseil de l'Europe pour qu'il aide les concepteurs de tests et d'examens dans leurs efforts et pour qu'il cautionne les tests ouvertement liés au *Cadre européen commun de référence* (Conseil de l'Europe 2009 :4). Cet enjeu a fait l'objet d'un séminaire organisé par les autorités finlandaises à Helsinki en 2002. À l'issue de ce séminaire, la Division des Politiques linguistiques du Conseil de l'Europe a décidé de développer un *Manuel pour relier les examens de langues au Cadre européen commun de référence pour les langues* (Conseil de l'Europe 2009 :4).[29] Cet ouvrage a été conçu dans l'intention d'aider les concepteurs d'examens et de tests à relier leurs instruments à ceux du CECRL. Il constitue, à ce titre, un nouvel outil, en complément des documents élaborés auparavant qui fournissent déjà des points de référence et des objectifs communs en tant que base pour la structuration transparente et cohérente de l'enseignement, de l'apprentissage et de l'évaluation en langues (Conseil de l'Europe 2009 :4).

On peut identifier plusieurs raisons à l'influence du CECRL sur les pratiques pédagogiques actuelles ; influence qui reste inférieure à son impact sur l'évaluation des compétences en langues. La première cause est l'opacité déjà évoquée d'un certain nombre de descripteurs. Ce manque de clarté et d'explicite a pour effet un moindre impact du document sur l'enseignement des langues (Westhoff 2007 : 676). Par ailleurs, les pratiques d'enseignement des langues vivantes qu'on observe actuellement ne sont pas forcément compatibles avec les échelles de compétences du CECRL (Westhoff 2007 : 677). Les usages d'enseignement suivent le principe pédagogique proposé par la plupart des manuels de langues européens qui, en simplifiant, se ramène à la trilogie « Présentation-Pratique-Production » (Thornbury 1999 : 128-129). Cette approche se base sur l'application de règles grammaticales permettant de produire du discours en créant de nouveaux énoncés bien formés. Elle contredit le principe du développement concentrique de la compétence en langues vivantes qui se trouve à la base des échelles de compétences du CECRL et qui est confirmée par la recherche sur l'acquisition du langage (Westhoff 2007 : 676). Selon le principe à la base du CECRL, la priorité des apprenants d'une langue vivante est d'apprendre autant d'unités et d'expressions lexicales que possible le plus vite possible (Westhoff 2007 : 677). La communication par le biais de l'application des règles grammaticales peut avoir lieu seulement après la mémorisation d'un répertoire lexical assez large car il est dès lors plus efficace d'appliquer les règles grammaticales permettant de produire du discours (Westhoff 2007 : 677). D'après les recherches effectuées lors du développement du CECRL, la transition entre ces deux manières de production langagière se produit entre les niveaux B1 et B2 (Westhoff 2007 : 678).

Pour élargir l'influence du CECRL sur l'enseignement et l'apprentissage des langues vivantes, certaines suggestions énoncées sont censées changer les pratiques pédagogiques courantes et les fonder sur les échelles de compétences du CECRL. Premièrement, il ne faut pas se figurer que les items grammaticaux individuels sont constitutifs des différents niveaux de compétences et les présenter ainsi aux apprenants. En effet, le lien entre les items spécifiques de grammaire et les niveaux communs de référence du CECRL n'est pas confirmé par la recherche sur l'acquisition en langue seconde, bien que la plupart des manuels d'apprentissage des langues vivantes l'affirment (Westhoff 2007 : 678).

Deuxièmement, il semble que la présentation linéaire des items grammaticaux individuels empêche les apprenants d'avancer vers les niveaux supérieurs de la compétence langagière parce qu'une telle méthode pédagogique les empêche d'acquérir un répertoire lexical de base (Westhoff 2007 : 678). [30]

Aux niveaux inférieurs de compétence langagière, la correction grammaticale peut seulement être atteinte par une pratique de communication intense en utilisant les collocations, les phrases et les diverses formules langagières (Westhoff 2007 : 678). L'information métalinguistique systématique et explicite remplit ses fonctions d'amélioration de la compétence communicative des apprenants à partir du niveau B2 seulement. Le caractère superflu de ce genre d'informations n'empêche pas que des règles grammaticales simples puissent être utiles à partir du niveau A2 par leur potentiel de prévention des erreurs grammaticales fréquentes. Elles sont utiles à condition d'être fonctionnelles dans le contexte et de ne pas déranger la communication en cours (Westhoff 2007 : 679).

Ce changement de perspective dans l'enseignement doit être accompagné de deux autres changements. En premier lieu, il est nécessaire de renouveler les programmes de manière systématique (Westhoff 2007 : 678). Le CECRL peut ainsi influencer l'élaboration des instructions officielles de deux manières (Little 2007 : 649). La première consiste à relier les objectifs d'apprentissage souhaités aux niveaux communs de référence. Un des exemples est l'objectif de faire atteindre le niveau B1 en langue seconde aux élèves à la fin du collège en France. La deuxième méthode pour employer le CECRL lors de la conception des curricula est d'utiliser le schéma descriptif de ce document pour analyser les besoins des apprenants en matière de compétences langagières et pour préciser leur répertoire communicatif cible en termes pédagogiques (Little 2007 : 649).

En deuxième lieu, il est indispensable de modifier le focus dans les manuels d'apprentissage des langues. Ces derniers doivent accentuer, notamment aux bas niveaux de compétences, les tâches communicatives portant un sens au lieu de proposer des exercices se focalisant sur les enjeux grammaticaux spécifiques. Il est préférable de proposer des tâches communicatives sous formats et dans le contexte associables avec les descripteurs du CECRL (Westhoff 2007 : 679). Dans ce contexte, il faut noter

qu'une série de manuels dédiés aux niveaux de compétence différents a été publiée en France (Beacco, de Ferrari, & Lhote 2006).

2.5.3.1 Reconnaissance mutuelle de qualifications langagières

Comme évoqué au début de ce chapitre, l'un des trois buts du *Cadre européen commun de référence* était de promouvoir la reconnaissance réciproque des qualifications en langues (Conseil de l'Europe 2005 : 12). Les niveaux communs de référence dans ce document ont été élaborés afin de rendre possible la comparaison objective des qualifications en langues entre les différents pays (Brown 2010 : 87). Il est évident que celle-ci constitue une étape indispensable avant de pouvoir atteindre ce but fixé par les concepteurs du CECRL.

Malgré cet objectif déjà envisagé au *Symposium intergouvernemental de Rüschlikon*, dix ans avant la publication du CECRL en 2001, la reconnaissance mutuelle de qualifications langagières par toutes les parties concernées reste une question complexe (Conseil de l'Europe 2009 : 3). Ceci est le cas en dépit d'une grande influence sur l'évaluation des compétences exercée par le CECRL. Il existe de nombreuses causes à ces difficultés de reconnaissance réciproque des qualifications en langues qui varient, en outre, selon les pays européens. Les facteurs qui peuvent empêcher une reconnaissance mutuelle des qualifications proviennent des politiques nationales, des traditions et des cultures d'évaluation d'un pays autant que des intérêts légitimes des organismes spécialisés dans les tests de langues (Conseil de l'Europe 2009 : 3).

En Europe, les systèmes d'évaluation ont des traditions très diverses. On trouve d'abord des systèmes dans lesquels les concepteurs d'examens opèrent selon un mode classique annualisé. Les épreuves sont préparées par une commission de spécialistes et notées en fonction des connaissances attendues. Dans de nombreux cas, l'examen ou le test débouchant sur une qualification reconnue est préparé par l'enseignant ou le personnel de l'école, plutôt que par une commission externe. Parfois encore, il est fait appel à un expert extérieur pour le contrôle (Conseil de l'Europe 2009 : 3). Par ailleurs, on s'aperçoit que de nombreux examens se concentrent sur la mise en œuvre de tâches spécifiées, avec des critères écrits, un barème et une formation des examinateurs permettant d'assurer une cohérence. Ils incluent ou excluent, selon le cas, une forme de pré-test ou de validation empirique. On trouve enfin des systèmes

extrêmement centralisés qui utilisent essentiellement des tâches à réponse fermée pour mesurer les capacités de réception (Conseil de l'Europe 2009 : 3-4). Ces tâches sont extraites de banques d'items. S'y ajoutent parfois des tâches de production, habituellement écrites, afin de mesurer la compétence et de délivrer les certifications (Conseil de l'Europe 2009 : 4).

Les inégalités en matière de ressources et de compétences constituent également une raison aux difficultés de reconnaissance mutuelle des qualifications langagières entre pays européens. Ainsi, les ressources humaines et matérielles disponibles varient entre les institutions, ce qui a un impact sur les compétences desdites institutions. Les ressources humaines et matérielles suffisantes dont disposent certains établissements offrent la possibilité de développer et de mettre en application les procédures de haute qualité en matière de formation des évaluateurs ainsi que dans le domaine du contrôle de la qualité des dispositifs et des processus d'évaluation (Conseil de l'Europe 2009 : 3). Bien que la disponibilité des ressources et des compétences s'influencent mutuellement, ces paramètres restent indépendants. La mise à disposition de ressources suffisantes n'assure pas la connaissance des techniques nécessaires à l'évaluation, basée sur les standards et la validation adéquate des tests (Conseil de l'Europe 2009 : 4). Malgré l'existence de nombreux facteurs susceptibles de restreindre la reconnaissance mutuelle des qualifications, il y va de l'intérêt de chacun que l'on applique des procédures adéquates en matière d'évaluation (Conseil de l'Europe 2009 : 3-4).

2.5.4 Devoirs futurs du CECRL

Le développement d'une influence plus forte, actuellement manquante, du CECRL sur l'enseignement et l'apprentissage des langues est liée à trois autres défis à relever. Le besoin le plus évident est d'utiliser le CECRL de façon à produire un impact significatif et durable sur les objectifs d'apprentissage des langues. Ceci présuppose l'emploi du référentiel dans le but de relier l'élaboration des programmes, les pratiques pédagogiques et les pratiques d'évaluation (Little 2007 : 652). Pour parvenir à ce but, il ne suffit pas que le CECRL soit à l'origine de certaines directives passées par les autorités ministérielles, précisant, par exemple, les niveaux de compétence à atteindre aux différentes étapes du

parcours scolaire. Le CECRL est également censé encourager les enseignants et les formateurs de langues à réfléchir à leurs pratiques pédagogiques actuelles. La réflexion doit notamment porter sur l'analyse des besoins concrets d'apprentissage, la détermination des objectifs et l'évaluation des progrès (North 2007 : 659).

Le deuxième défi qui se présente consiste à adapter les niveaux communs de référence aux besoins des apprenants en langues pendant la période de leur scolarisation (Little 2007 : 652). Little dénonce le fait que la plupart des descripteurs dans le CECRL ne sont pas adaptés aux besoins communicatifs des jeunes apprenants, ni par les activités communicatives décrites ni par les domaines dans lesquels ils se situent (Little 2007 : 652). Alors que les descripteurs des niveaux A1 à B1 sont jugés susceptibles d'être adaptés aux besoins spécifiques des jeunes utilisateurs, les descripteurs aux trois niveaux supérieurs sont considérés comme totalement inadéquats. Ceci est dû au haut degré de maturité cognitive, de réussite éducative et d'expérience professionnelle qui est présupposé à ces niveaux (Little 2007 : 652). Pour adapter les échelles de compétences aux besoins, aux expériences et aux objectifs de ce public, un projet a été lancé par la *Division des Politiques linguistiques du Conseil de l'Europe* visant à développer le *Cadre Européen commun de référence pour les langues* au sein de l'éducation scolaire. Le document qui est en cours d'élaboration présentera certains points communs avec le CECRL, mais se distinguera néanmoins de celui-ci sur plusieurs aspects (North 2007 : 658). En ce qui concerne les convergences, plusieurs catégories du schéma descriptif du CECRL sont pertinentes pour l'éducation scolaire et feront donc partie du nouveau *Cadre européen commun de référence*. Les catégories descriptives faisant partie des activités d'interaction et de production orale et écrite sont d'une pertinence particulière pour le Cadre commun de référence scolaire.

Dans le même temps, il y a au moins trois divergences essentielles entre les deux Cadres de référence. Premièrement, les descripteurs du CECRL définissent les résultats d'apprentissage se manifestant par un comportement particulier. En revanche, le Cadre européen commun destiné aux élèves devrait s'intéresser à la description des compétences émergentes et des conditions éducatives qui leur sont favorables. Deuxièmement, ce nouveau Cadre commun

de référence doit tenir compte du transfert des compétences acquises dans la langue de scolarité sur l'apprentissage des langues étrangères (North 2007 : 658). Troisièmement, le Cadre commun de référence scolaire doit situer l'acquisition des compétences langagières au sein du développement cognitif et social général des élèves (North 2007 : 65).

Le troisième défi est de concilier la tension qui résulte de la définition des compétences en langue seconde à l'aide des échelles de compétences tandis que la politique éducative du Conseil de l'Europe est focalisée sur le plurilinguisme. Or, le plurilinguisme est ancré dans la langue native de l'individu, désignée L1 (Little 2007 : 652). La tension soulignée par Little existe effectivement, mais elle a une autre explication que celle évoquée. Le plurilinguisme est une notion qui vise à « sortir de la dichotomie d'apparence équilibrée qu'instaure le couple habituel L1/L2 en insistant sur un plurilinguisme dont le bilinguisme n'est qu'un cas particulier » (Conseil de l'Europe 2005 : 129). Selon cette définition, la dichotomie entre L1 et L2 établie par les définitions de la compétence en langue seconde dans le *Cadre européen commun* constitue effectivement un problème, qui s'explique non pas par l'ancrage du plurilinguisme dans la L1, mais par le fait que la dichotomie entre L1 et L2 n'est pas incluse dans la notion de compétence plurilingue. Cette compétence ne sépare pas les compétences communicatives langagières selon les langues maîtrisées, mais contient « l'ensemble du répertoire langagier à disposition » d'un individu (Conseil de l'Europe 2005 : 129).

2.5.5 Application du CECRL I: les Portfolios européens des Langues

Le *Cadre européen commun de référence pour les langues* et le *Portfolio européen en Langues* (PEL) ont été liés dès leur conception (North 2007 : 656). Premièrement, la décision de créer les portfolios a été prise en 1991, c'est-à-dire la même année que celle du CECRL (Tagliante 2005 : 75). Le deuxième lien entre ces deux documents est la définition des compétences acquises en fonction des niveaux du Cadre européen commun, ce qui rend les portfolios comparables pour toutes les langues et transférables à tous les systèmes d'enseignement européens (Tagliante 2005 : 75).

En termes d'évaluation, ce document se distingue par sa focalisation explicite sur le développement de la compétence en langue seconde en vue de former l'identité plurilingue et pluriculturelle des apprenants (Tagliante 2005 : 76). Cette approche a des conséquences sur les pratiques pédagogiques, les méthodes d'évaluation et l'élaboration du curriculum (Byrnes 2007 : 642). En ce qui concerne les méthodes d'évaluation, l'auto-évaluation a une fonction primordiale dans le concept des portfolios européens des langues car celle-ci est considérée comme indispensable pour l'usage efficace de la langue (Little 2007 : 649). La définition des compétences est effectuée par les apprenants eux-mêmes dans les portfolios moyennant le processus d'auto-évaluation. L'importance accordée à l'auto-évaluation dans ce document se manifeste par le fait qu'elle apparaît sous forme d'évaluation formative et sommative des compétences (Little 2007 : 649). La biographie langagière sert l'évaluation formative, tandis que le passeport remplit la fonction d'évaluation sommative. La variation des descripteurs employés dans le PEL par rapport à ceux utilisés dans le CECRL s'explique par l'usage unique de descripteurs approuvés par le comité de validation [31] (Little 2007 : 650). Le procédé de sélection des descripteurs appropriés a exigé plusieurs tâches. La première était de décider, en s'appuyant sur les enquêtes des professeurs et des étudiants, quelles échelles de compétences contenues dans le CECRL et quels niveaux communs de référence seraient à inclure dans le PEL. La deuxième tâche consistait à développer et à éditer les descripteurs en maintenant les compétences de base suggérées dans le CECRL, mais à les modifier en fonction d'un contenu et d'un langage plus appropriés à la tranche d'âge respective des utilisateurs du PEL.[32] (Hasselgreen 2003 : 17). L'analyse qualitative des descripteurs ainsi adaptés a été effectuée sur la base de l'auto-évaluation pour certaines compétences et sur la base de l'étalonnage des descripteurs selon les niveaux pour d'autres compétences (North 2007 : 658).

Le PEL poursuit des objectifs qui vont au-delà du développement de la compétence en langue seconde et de la saisie des progrès à l'aide de supports écrits. Premièrement, il offre la possibilité aux apprenants de collecter leurs expériences d'apprentissage et d'usage des langues étrangères sous une forme structurée [33]. Deuxièmement, ce document remplit une fonction pédagogique grâce à son effet motivant. Il parvient à motiver les apprenants en reconnaissant

les efforts fournis pour développer leurs compétences linguistiques et culturelles ainsi qu'en présentant les résultats d'auto-évaluation de façon valorisante (Tagliante 2005 : 76). Troisièmement, ce document remplit une fonction d'information grâce au bilan de compétences fourni. Les informations qu'il contient sont utiles aux apprenants pour la poursuite d'études et pour la recherche d'un emploi (Tagliante 2005 : 75). Enfin, ce document a pour but de rendre le processus d'apprentissage des langues plus transparent aux apprenants et d'encourager le développement de leur autonomie (Little 2007 : 649). Dans cette perspective, ce n'est pas un hasard si le PEL se focalise sur l'auto-évaluation car cette approche pédagogique sert à développer l'autonomie des apprenants (Little 2007 : 649). En étant habitués à évaluer leurs compétences ainsi qu'à planifier et à surveiller leur apprentissage, les apprenants deviennent plus autonomes. On peut assurer la validité de l'auto-évaluation en demandant aux apprenants de prouver leurs déclarations par rapport à la maîtrise des compétences (Little 2007 : 649). Néanmoins, le rôle primordial de l'auto-évaluation dans ce document suscite un certain scepticisme chez les professeurs, notamment de la part de ceux qui ne sont pas familiarisés avec les approches pédagogiques susceptibles de développer l'autonomie des apprenants (Little 2007 : 651).

En ce qui concerne la structure des portfolios européens des langues, ils se composent toujours de trois documents qui recouvrent chacun une fonction précise: un passeport, une biographie langagière et un dossier (Tagliante 2005 : 76). Le passeport contient la grille d'auto-évaluation du CECRL qui permet à l'apprenant d'autoévaluer son niveau de compétences de façon sommative, pour chaque compétence langagière et dans chaque langue qu'on apprend ou maîtrise (Tagliante 2005 : 76). Le passeport met également en lumière les expériences d'apprentissage et d'usage des langues étrangères, en évoquant les cours auxquels l'apprenant a participé, les qualifications obtenues et les périodes de résidence dans les communautés dans lesquelles ces langues sont parlées (Little 2007 : 650). La biographie des langues sert à l'apprenant à déterminer ses objectifs d'apprentissage et à surveiller lui-même son progrès individuel. La surveillance de son avancement implique l'indication de ses compétences linguistiques et culturelles (Tagliante 2005 : 76). L'apprenant est également

encouragé par ce document à réfléchir aux différentes dimensions d'usage et d'apprentissage des langues, y compris à la formation de l'identité plurilingue et pluriculturelle. Dans le dossier, l'apprenant rassemble les preuves de ses compétences en langues qui mettent en évidence les compétences considérées comme acquises, en l'occurrence les échantillons de ses travaux personnels ainsi que ses certificats et diplômes obtenus (Little 2007 : 650).

Il faut noter qu'une version standard du PEL n'a pas été élaborée. En 1991, le Conseil de l'Europe a invité les États membres de l'Union Européenne à développer et à piloter leurs propres portfolios européens de langues (Little 2007 : 651). Préalablement à l'élaboration de ces documents, les principes et les lignes directrices pour leur conception et leur validation ont été établis et rassemblés dans la *Résolution sur le Portfolio européen des Langues*, adoptés par les ministres de l'Éducation des États membres de l'Union Européenne.[34] Ces principes concernent les caractéristiques propres aux portfolios européens, par exemple, les trois parties qui les composent, ainsi que leurs objectifs.[35] Les portfolios pilotes ont passé une phase d'expérimentation de deux ans, avant d'être généralisés en 2000, l'année européenne des langues (Tagliante 2005 : 75). La décision de créer les portfolios européens des langues fut un grand succès. Jusqu'à l'année 2010, un grand nombre de modèles furent élaborés par un large éventail d'organismes. Dès lors que les portfolios créés respectaient les principes et les lignes directrices du Conseil de l'Europe, ils étaient accrédités par le Comité de validation du Portfolio européen des langues [36] (Tagliante 2005 : 75). Ainsi, 118 modèles ont pu être validés par ce comité jusqu'à l'année 2010. La nécessité d'accréditer les portfolios résulte de la qualité variable des modèles développés ainsi que des différences au niveau de leur conformité aux principes et lignes directrices établis par le Conseil de l'Europe [37] (Little 2007 : 651).

Les modèles accrédités peuvent être répartis en différentes catégories selon plusieurs paramètres. Le premier concerne les apprenants auxquels un modèle donné est destiné. Le deuxième paramètre couvre les secteurs de l'éducation dans lesquels l'usage de ce modèle est envisagé. Les secteurs pour lesquels les Portfolios européens de langues ont été élaborés s'étendent de l'enseignement primaire à l'enseignement supérieur jusqu'à l'éducation pour les adultes et les

groupes ayant des besoins particuliers. Le troisième paramètre qui différencie les portfolios est le format sous lequel ils ont été édités, à savoir en format papier ou en format électronique. Le dernier paramètre de catégorisation des modèles est l'organisme par lequel ils ont été édités. Certains portfolios européens ont été publiés par des ministères, d'autres par des éditeurs du secteur commercial et d'autres encore par des organisations internationales non-gouvernementales (Little 2007 : 651).

Pour démontrer le caractère concret des Portfolios européens de langues, nous explorerons à titre d'exemple le modèle français, appelé *Portfolio Européen des Langues*, destiné aux jeunes à partir de 15 ans, et surtout libre d'accès.[38] Les descripteurs inclus dans ce modèle proviennent de la *Banque de descripteurs pour l'auto-évaluation créée pour le Portfolio européen des langues,* mise à la disposition des auteurs des PEL gratuitement [39]. Les descripteurs sont regroupés en listes décrivant les niveaux de compétences dans chacune des cinq activités de communication langagière distinguées par le CECRL.[40] Il faut noter que les niveaux de compétences, déterminés à l'aide des descripteurs inclus dans les portfolios ne doivent pas être confondus avec les niveaux de compétences réels, car l'auto-évaluation, pratiquée de façon exclusive dans tous les PEL, ne suffit pas à cet usage. L'évaluation du niveau réel de compétences reste la fonction des enseignants et des formateurs.[41] L'objectif de chaque portfolio européen des langues est de proposer aux apprenants un support structuré pour prendre conscience de leur méthode d'apprentissage, y réfléchir et pouvoir l'ajuster ou l'adapter à leurs besoins particuliers. [42]

En conformité avec tous les modèles de portfolios européens des langues, les descripteurs utilisés dans le modèle présenté sont adaptés aux utilisateurs envisagés sur le plan de leur contenu et du langage employé (Hasselgreen 2003 : 17). Il est notamment intéressant d'examiner la manière dont les descripteurs utilisés dans ce portfolio ont été modifiés par rapport à ceux employés dans le CECRL.[43] Les descripteurs inclus dans ce modèle se distinguent par trois caractéristiques particulières. Premièrement, ils sont accompagnés par un exemple qui sert à faciliter leur compréhension, étant donné que les descripteurs ne sont pas spécifiques à un contexte et sont

formulés de façon générale.[44] L'un des descripteurs définissant les compétences réceptives à l'oral au niveau A1 peut illustrer cette qualité :

> I can understand information and simple instructions.
>
> e. g. when I am told where to find something or someone or when I am asked to come, to open my book, to go to the board, to wait, etc. [45]

Cet exemple montre que le descripteur est adapté au milieu scolaire, évoquant les situations fréquentes au quotidien des élèves.

Deuxièmement, les listes de descripteurs proposent à deux niveaux, A2 et B1, l'étalonnage des compétences en trois niveaux intermédiaires : A2.1, A2.2, A2.3, B1.1, B1.2 et B1.3. L'inclusion des définitions de compétences à ces niveaux constitue la différence principale entre les descripteurs de ce portfolio et ceux contenus dans la grille d'auto-évaluation du CECRL. L'arborescence plus fine permet la prise en compte de tout progrès des apprenants même minime.[46] Les trois descripteurs de la compréhension de l'oral au niveau A2 peuvent illustrer ce fait. Pour des raisons de clarté, les exemples qui accompagnent chaque descripteur ne sont pas cités:

> A2 I can understand very common expressions and vocabulary closely concerning myself (e.g. my family, purchases, close environment, work). I can grasp the essential meaning of simple, clear messages
>
> A2-1 When the speaker uniquely uses almost only words and expressions that I ought to know…
>
> - I can understand if he is asking a question, if he is stating something or if he is asking to do something.
>
> - I can understand when he is introducing himself, is speaking about his family and his likes and dislikes.
>
> - I can understand when he asks me what I like.
>
> - I can recognise words and expressions I know in a narrative or a dialogue.
>
> A2-2 When the speaker uses simple sentences to talk about everyday issues because he knows he is talking to someone who is learning the language...
>
> - I can understand when he introduces another person.
>
> - I can understand simple instructions about how to get from one place to another, on foot or by public transport.
>
> - I can understand the general subject of a discussion I hear.

- I can understand the main theme of a short narrative which contains connected sentences.

A2-3 When the speaker is addressing a wider audience on issues which I am familiar with, but using sentences which are short and simple...

- I can understand the key information of a short message.

- I can understand a narrative if it is about facts in the present, past or future.

- I can understand the logical composition of a narrative.

- I can follow the TV news headings or televised documentaries presented quite slowly and clearly in standard language, even if I do not understand all the details. [47]

Il est frappant que les descripteurs aux trois niveaux intermédiaires se réfèrent à la même capacité réceptive, attendue au niveau A2, à savoir être capable de comprendre des messages simples et clairs qui concernent l'apprenant de près. Néanmoins, la complexité des structures langagières varie en fonction des trois niveaux intermédiaires. Au niveau A2.1, les messages simples se limitent uniquement aux mots et expressions connues. Ceci implique qu'il ne soit pas demandé à l'apprenant d'inférer le sens d'un message à partir d'éléments familiers. En outre, la longueur des messages à comprendre au niveau inférieur ne dépasse pas une phrase simple seule. L'apprenant est incapable de comprendre le sens des structures langagières plus larges, mais seulement d'y reconnaître les mots et les expressions connues. Cette caractéristique du niveau A2.1 constitue la différence principale par rapport au niveau suivant.

Au niveau A2.2, la compréhension du sujet principal d'un discours cohérent, constitué à partir de phrases simples, est attendue. Ce discours peut être une discussion ou une narration. Comme au niveau A2.2, au niveau intermédiaire supérieur la compréhension d'un message composé de plusieurs phrases brèves et simples est également demandée à l'apprenant, mais celle-ci ne se limite toutefois pas au sujet principal. Au niveau A2.3, la capacité de comprendre les informations principales des messages provenant de différentes sources est attendue, y compris celles d'émissions de radio articulées lentement et clairement. Un autre trait distinctif à ce niveau est le fait que le destinataire du discours n'est plus l'unique interlocuteur, mais que le discours est adressé à un

large public dont l'interlocuteur fait partie. Pour conclure l'analyse de ces trois descripteurs intermédiaires, il faut constater que la compréhension du sens principal de messages simples et clairs exigée par le niveau A2 est différenciée en fonction des niveaux de compétence intermédiaires. Les différences portent sur la qualité de compréhension du message écouté, la complexité des structures langagières utilisées, sur la source du discours ainsi que sur son destinataire. La différenciation des messages en fonction de ces quatre paramètres révèle un large éventail de capacités réceptives au sein d'un même niveau de compétences. Il faut préciser que c'est le cas pour tous les niveaux de compétences de ce PEL et pas seulement pour le niveau A2 ou les descripteurs des activités de réception, choisis à titre d'exemple. Une telle organisation des descripteurs se distingue de celle qui a lieu dans le CECRL. Dans ce dernier, il n'y a pas de descripteurs supérieurs qui servent de référence aux autres descripteurs. Ceci est valable aussi pour les descripteurs des niveaux avancés, A2+, B1+ et B2+.

Une caractéristique commune à tous les modèles de Portfolios européens de langues est l'emploi des mêmes listes de descripteurs pour toutes les langues apprises ou maîtrisées par les utilisateurs dans lesquelles ils veulent évaluer leurs compétences.[48] Cette qualité constitue un trait commun avec les échelles de compétences et la grille pour l'auto-évaluation du CECRL qui sont également conçues pour toutes les langues. En indiquant ses compétences dans les différentes langues dans les cases juxtaposées, la comparaison des savoir-faire dans ces langues est encouragée. Ceci incite l'utilisateur de ce modèle à prendre conscience de la complémentarité des compétences dans ses différentes langues pour mener à bien des situations de communication diverses et variées. Cette prise de conscience favorise le développement de la compétence plurilingue des apprenants [49].

Bien que la décision prise par le Conseil de l'Europe d'élaborer et d'installer les Portfolios européens de langues ait été un succès, il est impossible de savoir dans quelle mesure cet outil d'apprentissage est effectivement utilisé en pratique et non pas seulement distribué (Little 2007 : 652). Selon les études empiriques menées, l'usage de cet instrument peut avoir un impact positif sur l'enseignement et l'apprentissage des langues (Sisamakis 2006 :108). Néanmoins, des indices

suggèrent qu'il est souvent considéré comme un complément facultatif aussi bien par les apprenants que par les enseignants, pour qui son usage implique un travail supplémentaire. Ce point de vue s'explique par le détachement de la plupart des modèles des programmes en vigueur dans les systèmes éducatifs (Little 2007 : 652). Afin de faire évoluer les postures, il semble nécessaire de fonder les portfolios proposés aux apprenants sur des programmes d'enseignement valables. Ceci nécessite l'adaptation des directives et des méthodes d'enseignement à la perspective actionnelle du CECRL, en donnant une place à l'auto-évaluation, en raison du rôle central de cette forme d'évaluation pour l'usage efficace des Portfolios européens de langues (Little 2007 : 652).

2.5.6 Application du CECRL II: la conception de la Grille d'analyse du contenu des items de compréhension écrite et orale

Après la publication du CECRL, le projet néerlandais, nommé « Dutch CEF Construct Project », visa à déterminer si ce document était un instrument adéquat pour la construction des tests de langues (Alderson 2004 : 4). Il faut noter que le CECRL ne saurait être considéré comme utile et pertinent que si les échelles de compétences et les informations qu'il fournit donnent suffisamment d'indications pour l'élaboration des items aux différents niveaux de compétences (Alderson 2004 : ii). Dans le cas contraire, les chercheurs devaient déterminer quel type d'instrument serait susceptible d'aider à construire des tests adossés au CECRL. Ils cherchèrent donc à déterminer les éléments nécessaires pour développer un tel instrument (Alderson 2004 : ii). Le but ultime de ce projet de recherche était de créer un outil, fondé autant que possible sur le CECRL, qui décrirait les compétences de compréhension écrite et orale à la base des items et des tests sur les six niveaux communs de référence (Alderson 2004 : 1).

La première fonction attribuée à cet instrument est de guider les concepteurs des tests lors de la construction de nouveaux items et lors de l'analyse des items existants à tous les niveaux communs de référence du CECRL. Il faut que l'instrument développé contienne des critères linguistiques, psycholinguistiques, sociolinguistiques ainsi que pragmatiques pertinents afin de sélectionner les items aux différents niveaux communs de référence et afin de les

construire et de les réviser (Alderson 2004 : 3). Au-delà de ces fonctions pour les items dans les tests individuels, ce dispositif est prévu comme une aide à la conception de la banque d'items fondée sur le CECRL et à la sélection des items appropriés pour cette banque (Alderson 2004 : 1). Concernant ce dernier objectif, les chercheurs espèrent construire une banque d'items qui permettrait de relier les tests et les examens nationaux aux niveaux communs de référence du CECRL. En effet, la capacité de relier les dispositifs d'évaluation aux niveaux communs de référence est considérée comme essentielle, de sorte qu'elle figure au début de la description du projet de recherche consacré à l'élaboration de la grille.[50] La deuxième fonction prévue pour la banque d'items est d'illustrer les niveaux communs de référence par les items calibrés contenus dans cette banque. Cette fonction constitue également un besoin urgent (Alderson 2004 : 1).

Il faut rappeler que dans ce projet, les descripteurs définissant les compétences réceptives à l'oral et à l'écrit ont suscité un intérêt en raison des projets de recherche menés auparavant, y compris DIALANG, qui ont constaté que les données empiriques utilisées pour justifier les échelles de compétences étaient moins fortes pour les compétences réceptives à l'écrit et à l'oral que pour les compétences productives (Alderson 2004 : 3). Les échelles illustratives de compétences étalonnant ces dernières ont été jugées adéquates pour l'évaluation de la performance écrite et orale. Puisque ce n'est pas le cas pour les compétences réceptives, ce projet s'est focalisé là-dessus (Alderson 2004 : 3). Pour atteindre les objectifs évoqués, l'équipe du projet a décidé d'effectuer les quatre étapes d'analyse suivantes.

Premièrement, il a été envisagé de développer une grille d'analyse du contenu des items et des tests de compréhension écrite et orale en anglais, en français et en allemand. À la suite du développement de cet instrument, il a été prévu d'examiner un nombre d'items et de tests qui prétendent être explicitement adossés aux différents niveaux communs de référence du CECRL. Lors d'une troisième étape, il a été décidé d'examiner les points communs et les différences entre les spécifications des tests examinés (Alderson 2004 : 4). Enfin, on s'est attaché à analyser la manière dont les tests « opérationnalisent » le développement de la compétence langagière dans les items (Alderson 2004 :4). Ces étapes d'analyse avaient pour but de développer un instrument qui serait à

la fois un complément pratique du CECRL et un cadre théorique plus spécifié (Alderson 2004 :4). La méthode appliquée par l'équipe du projet était inductive car les résultats obtenus à chaque étape ont été utilisés pour la réflexion sur les données recueillies, pour la planification des étapes suivantes, ainsi que pour la révision des outils théoriques (Alderson 2004 :5). Pour développer de nouveaux instruments théoriques, il a d'abord été décidé d'adhérer aux formulations originales utilisées dans le CECRL et de n'apporter des adaptations qu'ensuite, en fonction des besoins (Alderson 2004 : 5). Afin de clarifier les activités menées dans le cadre du projet évoqué, nous décrivons ci-dessous ces étapes plus en détail.

2.5.6.1 Etape 1 : Analyse du CECRL

Il était évident pour les experts que le CECRL était le document central nécessaire à examiner afin de concevoir l'instrument envisagé. Pour cette raison, l'équipe du projet a analysé ce document en vue d'identifier ses lacunes éventuelles (Alderson 2004 : 6). Ainsi, les échelles de compétences générales et illustratives pour la compréhension orale et écrite ont été compilées. A cet usage, les descripteurs constitutifs de ces échelles ont été regroupés par niveaux et non par activités de communication langagière, comme c'est le cas dans le CECRL. Deuxièmement, les extraits du *Cadre* européen commun entretenant un quelconque lien avec les compétences réceptives à l'oral et à l'écrit ont été compilés. Enfin, les passages issus du *Manuel préliminaire pour relier les examens de langues* au *Cadre européen commun de référence pour les langues (CECRL),* jugés pertinents pour les tests de compréhension orale et écrite, ont également été conservés (Alderson 2004 : 6).

Lors de la deuxième phase de cette étape, les informations pertinentes pour la construction des tests contenues dans les échelles, ainsi que le texte associé ont été repérés. Les données extraites ont ensuite été réparties en trois catégories. La première catégorie regroupait les verbes utilisés dans le *Cadre* européen commun de référence et caractérisaient la nature de la compréhension, par exemple comprendre, reconnaître, localiser, déduire [51]. La deuxième et la troisième catégorie se focalisaient respectivement sur les sujets et les sources des textes qui sont censés être compris à un certain niveau commun de référence (Alderson 2004 : 6). Le cadre d'analyse a été conçu sur la

base de toutes ces informations, organisées en trois colonnes. Ensuite, l'instrument a été révisé et appliqué à un échantillon de tests de compréhension écrite et orale (Alderson 2004 : 6).

L'élaboration et la révision du cadre d'analyse ont été accompagnées par des discussions sur le rôle du pilotage et du calibrage d'un item ainsi que sur la difficulté de traiter les données empiriques (Alderson 2004 : 13). Ces discussions ont émergé suite au problème conceptuel fondamental rencontré dès le début du projet, à savoir la signification du classement d'un item à un niveau de compétences particulier. Cet enjeu est étroitement lié à la question de la validité des niveaux communs de référence eux-mêmes (Alderson 2004 : 13). Il est essentiel de répondre à ces deux questions en raison de leur importance pour la construction des tests. Afin de valider tout instrument d'évaluation, il faut démontrer qu'il peut attribuer à un candidat un niveau de compétences qui correspond à ses compétences réelles (Alderson 2004 : 13). Le lien entre la validation des tests et la validation des niveaux de compétences, qui sont des concepts, pose un problème circulaire. La validation de ces deux éléments, des tests aussi bien que des concepts théoriques, est seulement possible en cas de validation préalable de l'autre élément (Alderson 2004 : 13). [52]

2.5.6.2 Etape 2 : Mise en application du cadre d'analyse

Le cadre d'analyse révisé lors de l'étape précédente a été à nouveau modifié lors de cette phase du projet en ajoutant les éléments qui n'étaient pas inclus dans le *Cadre* européen commun, mais jugés essentiels pour caractériser les tests de compréhension écrite et orale (Alderson 2004 :6). Ensuite, la deuxième version de ce nouvel instrument, désigné la « grille », a été développée. Celle-ci, surnommée « grille 2 », contenait les éléments considérés comme nécessaires pour caractériser les items de compréhension écrite et orale contenus dans une banque (Alderson 2004 :6). La grille révisée a été appliquée à la catégorisation d'un nombre d'items de compréhension écrite et orale inclus dans DIALANG. Le but de cette analyse était de tester la grille révisée, d'identifier les difficultés de son usage et les points faibles éventuels des catégories qu'elle contenait (Alderson 2004 :14). A la lumière de l'analyse des résultats, la grille a été modifiée à nouveau et transformée en un instrument numérisé, appelé « grille

3 ». (Alderson 2004 : 14). Il faut noter que plusieurs dimensions dans la grille ne se sont pas avérées discriminatoires entre les différents niveaux communs de référence. En outre, aucune dimension de la grille n'a fait preuve d'une association significative avec les niveaux communs de référence (Alderson 2004 : 14).

2.5.6.3 Etape 3: Mise en application élargie de la grille révisée

À la troisième étape du projet, la grille 3 a été employée pour l'analyse d'un nombre d'items et de tâches relevant de sources variées.[53] Citons à titre d'exemple, en plus de DIALANG, les examens néerlandais de fin d'études HAVO 2000 et MAVO 1999, les Certificats Nationaux Finlandais pour l'Anglais et le Certificat Français de l'Enseignement Supérieur en Langues Étrangères (Alderson 2004 :7). En outre, les spécifications et les lignes directrices utilisées pour la construction des tests ainsi que la conception des items par certaines autorités européennes en matière d'évaluation en langues, par exemple DIALANG et l'Association des Évaluateurs en Langues en Europe, ont été examinées en détail lors de cette étape (Alderson 2004 :7).

Il faut noter que le taux d'accord entre les analystes était supérieur à celui constaté en phase précédente et a atteint 75% dans plusieurs dimensions (Alderson 2004 : 16). Malgré ce taux d'accord élevé, l'accord entre les spécialistes n'était que modéré sur le contenu de certains paramètres dans certains items, à savoir sur les opérations identifiées ainsi que sur les compétences testées[54] (Alderson 2004 : 17). La difficulté de se mettre d'accord sur l'opération nécessaire pour répondre à un item à un niveau de compétences donné est compatible avec les résultats de la recherche menée sur les activités de lecture en langue étrangère (Alderson 2000 : 298). Ce constat démontre qu'il est impossible de distinguer les niveaux de compétences en fonction des opérations demandées parce que ces dernières ne sont pas les indicateurs d'un niveau de compétence particulier (Alderson 2004 : 17). En ce qui concerne d'autres dimensions de la grille, elles n'ont pas montré d'association significative avec les niveaux communs de référence dans cette phase du projet non plus, à part la dimension de vocabulaire (Alderson 2004 : 17).

Les résultats clés du projet mené sont les grilles d'analyse des items de compréhension orale et écrite qui permettent aux concepteurs des tests

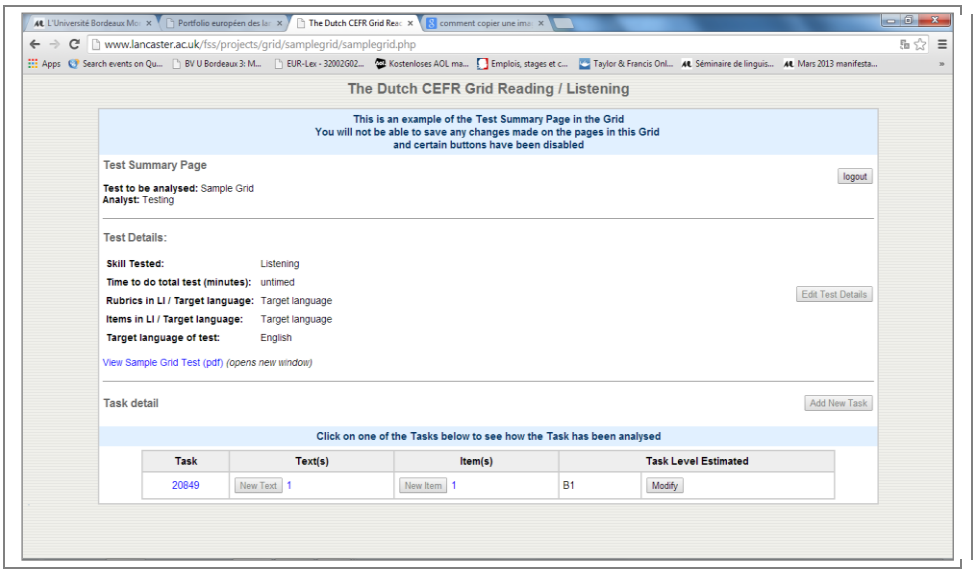
d'analyser les tâches et les items évaluant ces deux compétences en vue de relier les tests au CECRL.[55] En outre, les grilles fournissent les lignes directrices sur la manière de construire les items et les tests aux différents niveaux du CECRL (Alderson 2004 :1). De ce fait, ces instruments remplissent les objectifs visés par l'équipe du projet. Afin d'atteindre ces buts, les grilles ont été révisées cinq fois. Il faut noter que les items ont été analysés indépendamment par des experts afin de vérifier la faisabilité de l'utilisation de la grille et de repérer les problèmes spécifiques de son emploi (Alderson 2004 :4).

Il ne faut pas oublier que l'analyse du contenu d'un test ou d'un item ne suffit pas à valider leur niveau de compétence présumé et par voie de conséquence leur adossement au CECRL. Le processus de relier les tests au CECRL se compose de deux éléments également importants. Le premier consiste à caractériser le contenu des items et des tâches**.** Ce processus est facilité en complétant la grille.[56] La deuxième composante aussi importante pour relier les items et les tâches au CECRL est le fait d'effectuer des études sur des échantillons représentatifs du public visé et d'analyser les résultats empiriques du test.[57] Le lien présumé d'un test au CECRL peut seulement être exploré et, éventuellement, validé si les résultats de l'analyse du contenu ont été combinés avec les standards établis sur la base des résultats empiriques du test.[58] Pour concrétiser ces explications précédentes nous allons présenter la conception de la grille numérisée dans le sous-chapitre suivant.

2.5.6.4 La conception de la grille numérisée

Pour faciliter le travail de compréhension de la grille numérisée, celle-ci n'est pas présentée seule sur le site, mais sa conception et son usage sont illustrés par deux dispositifs. Ceux-ci, accessibles depuis la page d'accueil, ont des fonctions pédagogiques différentes.[59] Le premier dispositif est une grille échantillon qui montre aux utilisateurs comment une tâche a été analysée. Il s'agit de la tâche évaluant la compréhension orale en anglais extraite du test DIALANG.[60] La grille se compose de deux pages dont la première, appelée « test summary page », sert à résumer le test.[61] Elle remplit cette fonction en raison de l'inclusion des paramètres concernant le test entier, à savoir les compétences évaluées, la durée de temps imposée, le nombre de rubriques, le nombre d'items et la langue

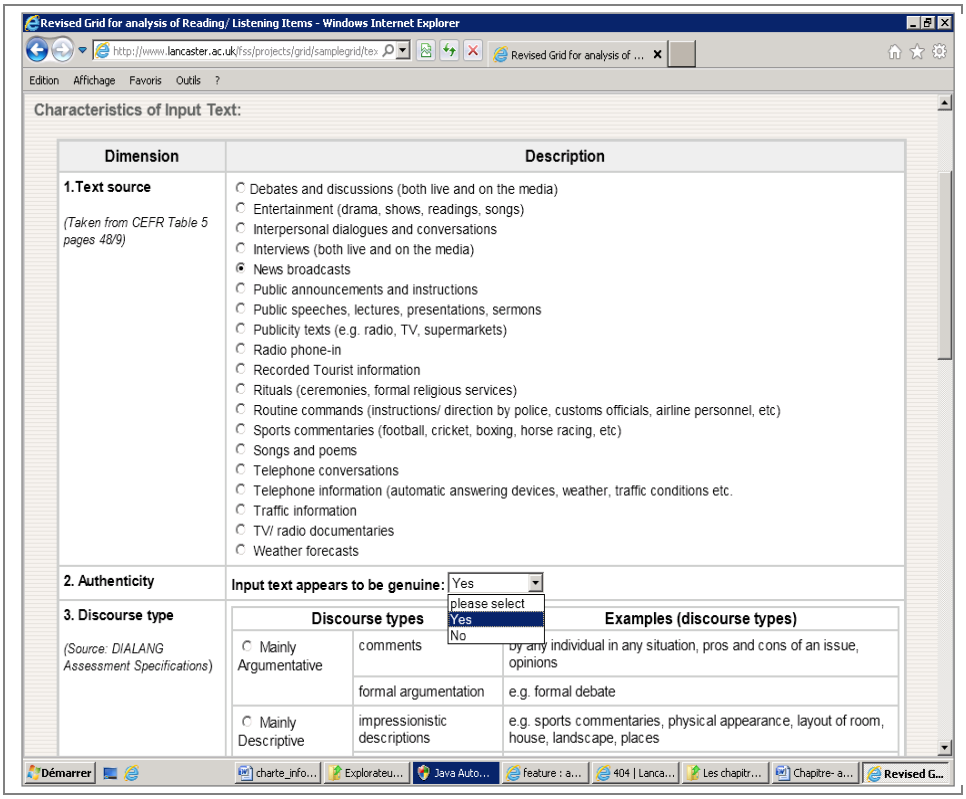
cible. La première page de la grille contenant les informations sur le test échantillon DIALANG illustre la disposition graphique des paramètres inclus et des caractéristiques repérées.



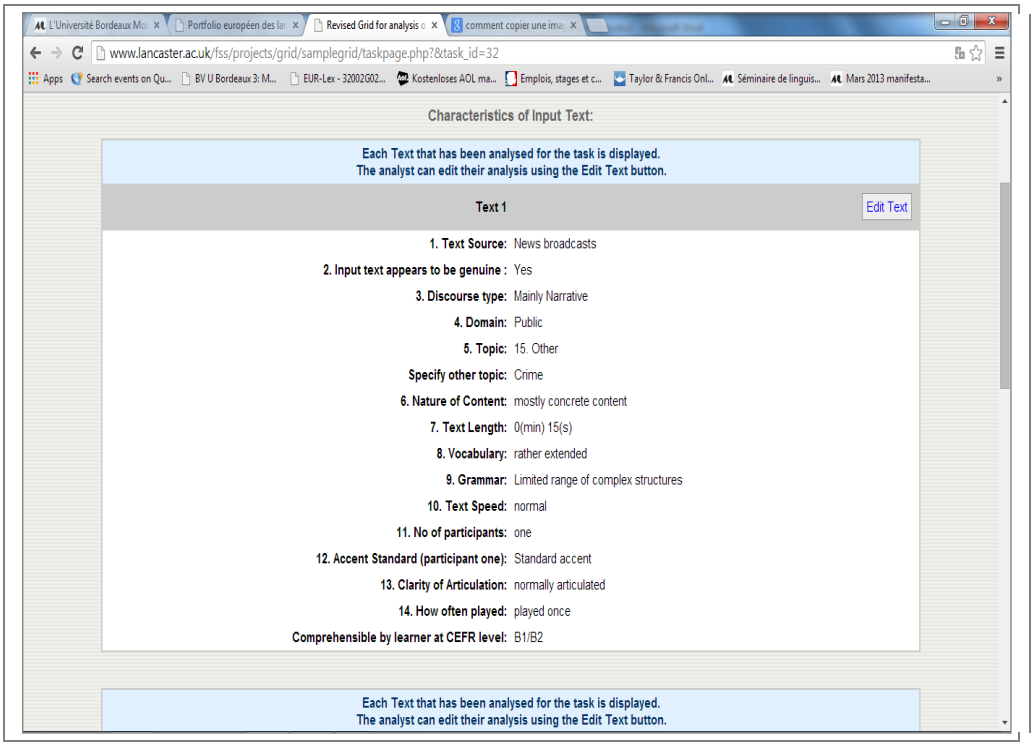
On voit qu'il manque sur cette page des informations concernant les rubriques et les items contenus dans le test. L'entrée « target language », à traduire par « langue cible », n'est évidemment pas adéquate. Le tableau en bas de la première page fournit des informations sur la composition de la tâche analysée sur la deuxième page de la grille. Ainsi, la tâche issue du DIALANG, présentée dans la grille échantillon, contient un texte et un item. [62]

La deuxième page de la grille a pour fonction d'analyser une tâche séparée. Pour ce faire, la tâche est décomposée en parties constitutives, à savoir un ou plusieurs textes et un ou plusieurs items analysés séparément. Cela résulte en une répartition de la grille en fonction du nombre des parties intégrales d'une tâche. Pour caractériser le contenu de la tâche donnée, il faut entrer les informations sur chaque texte et chaque item dans la grille en spécifiant leurs caractéristiques dans les dimensions indiquées parmi plusieurs options extraites du CECRL.[63] Le niveau de compétences estimé qui est la dernière dimension de la grille apparaît également dans une liste d'options. Ces options sont présentées selon les dimensions soit sous la forme d'un QCM soit sous la forme d'un menu déroulant. Les deux possibilités de disposition des options sont montrées ci-dessous. [64] Malgré leurs formes différentes, il s'agit d'un même format, qui est le questionnaire à choix multiples.

.



Puisque la tâche présentée dans la grille échantillon se compose de deux textes parlés et d'un item, la grille est constituée de trois parties [65]. Pour illustrer le résultat de l'analyse du contenu de chaque texte et de l'item séparément, les grilles concernées sont exposées successivement. La première grille présentée sert à caractériser le premier texte parlé inclus dans la tâche.

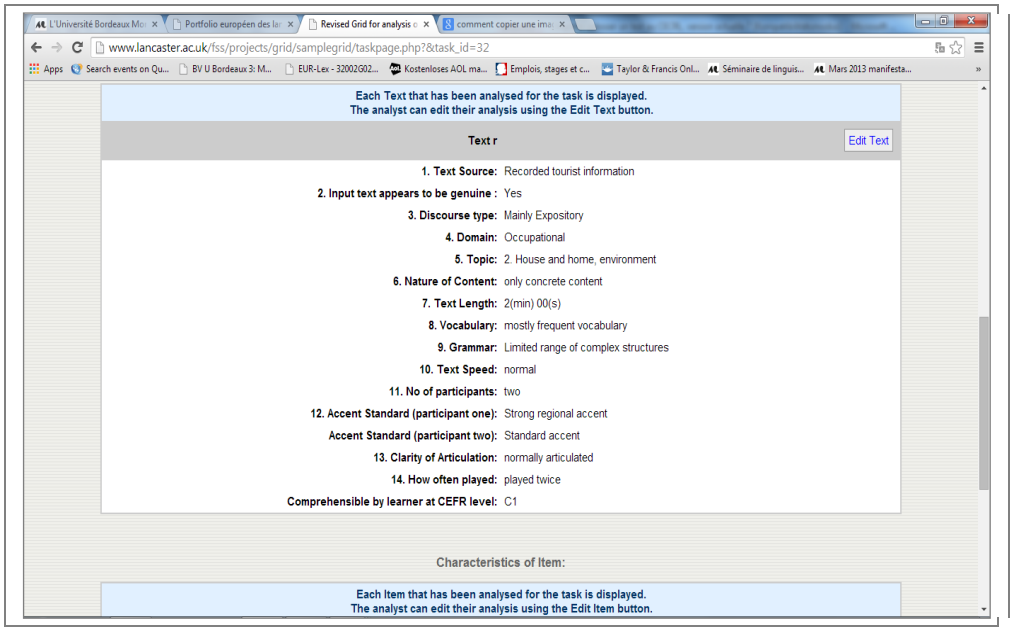


La grille destinée à cet usage contient 14 dimensions issues du CECRL, propres à l'analyse du contenu des textes. Ces dimensions sont complétées par l'estimation du niveau auquel le texte est compréhensible. Les grilles d'analyse des textes écrits ne sont pas identiques à celles destinées à l'analyse des textes parlés. Ces dernières contiennent cinq dimensions supplémentaires par rapport aux grilles d'analyse des textes écrits, ce qui est dû à la différence entre les deux types de textes. Il s'agit de la vitesse du texte articulé, du nombre de participants, de l'accent, de la clarté d'articulation ainsi que du nombre de fois où il est prononcé.

On peut répartir les paramètres évoqués en subjectifs d'une part et en objectifs d'autre part (Alderson 2004 : 14). Les dimensions subjectives sont plus nombreuses, à savoir la source du texte, le sujet, l'authenticité, la nature du contenu, la grammaire et le vocabulaire, la vitesse du texte articulé, l'accent et la clarté d'articulation. Les dimensions objectives concernent le domaine auquel le texte appartient, la durée du texte parlé, le nombre de participants et de fois prononcé, ainsi que le nombre de mots dans le texte écrit.

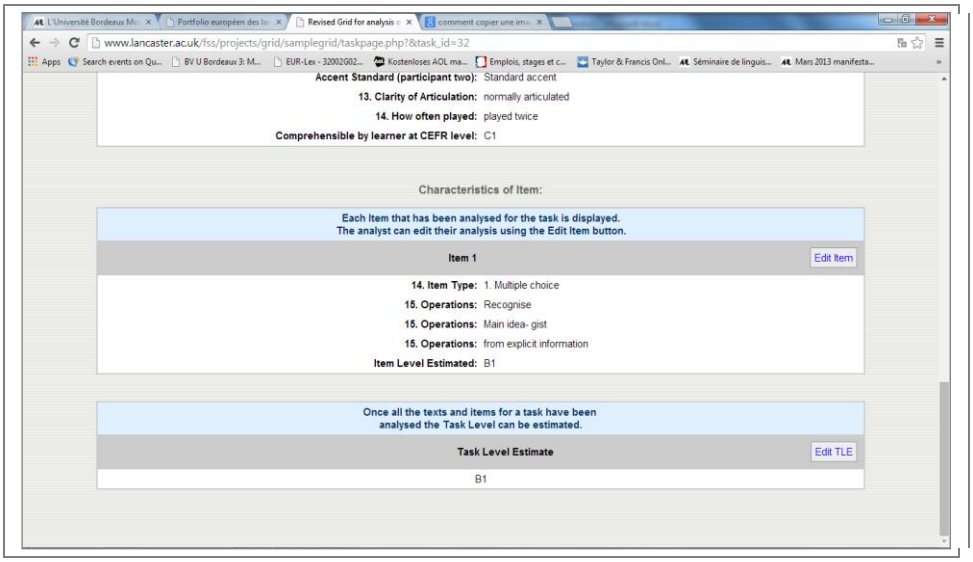
En ce qui concerne le deuxième texte constitutif de la tâche échantillon, il se distingue dans presque toutes ses caractéristiques du premier texte. Afin

d'illustrer ce constat, nous présentons ci-dessous la partie de la grille focalisée sur l'analyse de celui-ci.



On voit que les caractéristiques communes aux deux textes sont décrites par quatre dimensions uniquement, à savoir l'authenticité apparente du texte, la grammaire, la vitesse du texte articulé et la clarté d'articulation. En l'occurrence, les deux textes sont considérés comme authentiques et on considère que la grammaire contient un répertoire limité de structures complexes. Quant à la vitesse du texte articulé, elle est considérée comme normale dans les deux cas. Enfin, les textes sont articulés normalement. La différence entre les textes par rapport à la grande majorité des caractéristiques n'est pas étonnante, mais la divergence en nombre d'écoutes surprend, étant donné que les deux textes font partie de la même tâche. Il en est de même pour la différence entre les niveaux de compétence nécessaires à la compréhension de ces deux textes puisqu'ils sont destinés à évaluer le niveau de compétences d'un même apprenant qui doit comprendre les deux textes pour pouvoir répondre à l'item.

La grille destinée à l'analyse des items comporte seulement deux dimensions qui sont le type d'item et les opérations nécessaires afin d'y répondre. L'item inclus dans la tâche échantillon est au format QCM et demande l'opération « reconnaître l'idée principale à partir de l'opération explicite ». Ces deux dimensions sont complétées par l'estimation du niveau de compétence de l'item dont nous présentons la grille d'analyse ci-dessous.

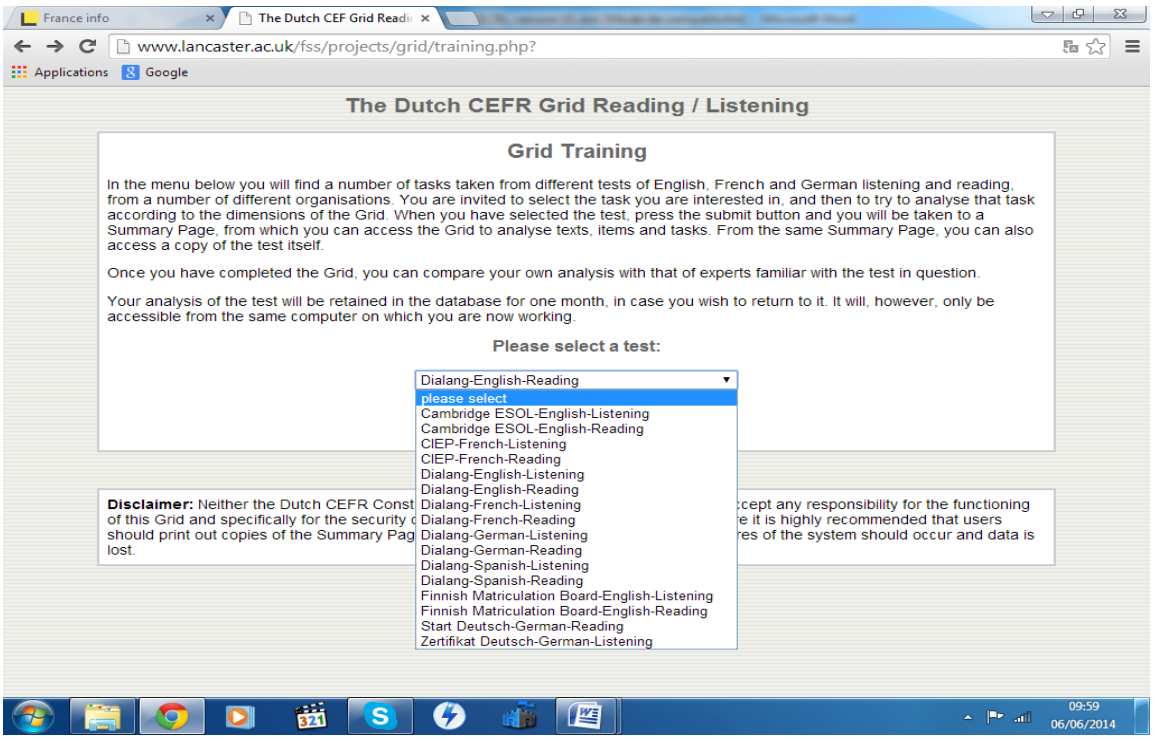


Il faut noter que le niveau de compétences estimé de l'item est inférieur aux niveaux de compétences jugés suffisant à la compréhension des deux textes, étant donné que même le texte le plus simple exige de l'apprenant un niveau intermédiaire entre les niveaux B1/B2. Cette divergence entre les niveaux nécessaires à la compréhension des textes et le niveau estimé de l'item implique que le niveau de compétences de l'apprenant est évalué par les textes et non par l'item lui-même. En cas de non maîtrise du niveau C1, exigé pour la compréhension de textes plus difficiles, il ne sera pas possible à l'apprenant de répondre correctement à l'item, même s'il y a correspondance entre le niveau de compétences de l'individu et le niveau estimé de l'item. Un tel item fait donc preuve d'un index de discrimination faible.[66] Il est frappant que le niveau estimé de la tâche coïncide avec le niveau estimé de l'item, ce qui signifie l'absence de prise en compte des niveaux de compétences attribués aux deux textes constitutifs de la tâche.

2.5.6.4.1 Le module d'entrainement

Plusieurs recommandations visant à améliorer l'efficacité de la grille présentée et à faciliter la compréhension et l'usage de celle-ci ont été faites suite à la mise en œuvre de cet instrument (Alderson 2004 : 19). Parmi les procédures recommandées figurent le développement d'un guide pour les utilisateurs contenant les conseils sur l'usage correct et abusif de la grille. Une autre recommandation concerne l'élaboration des lignes directrices par rapport à la compétence langagière à chaque niveau commun de référence, en termes de

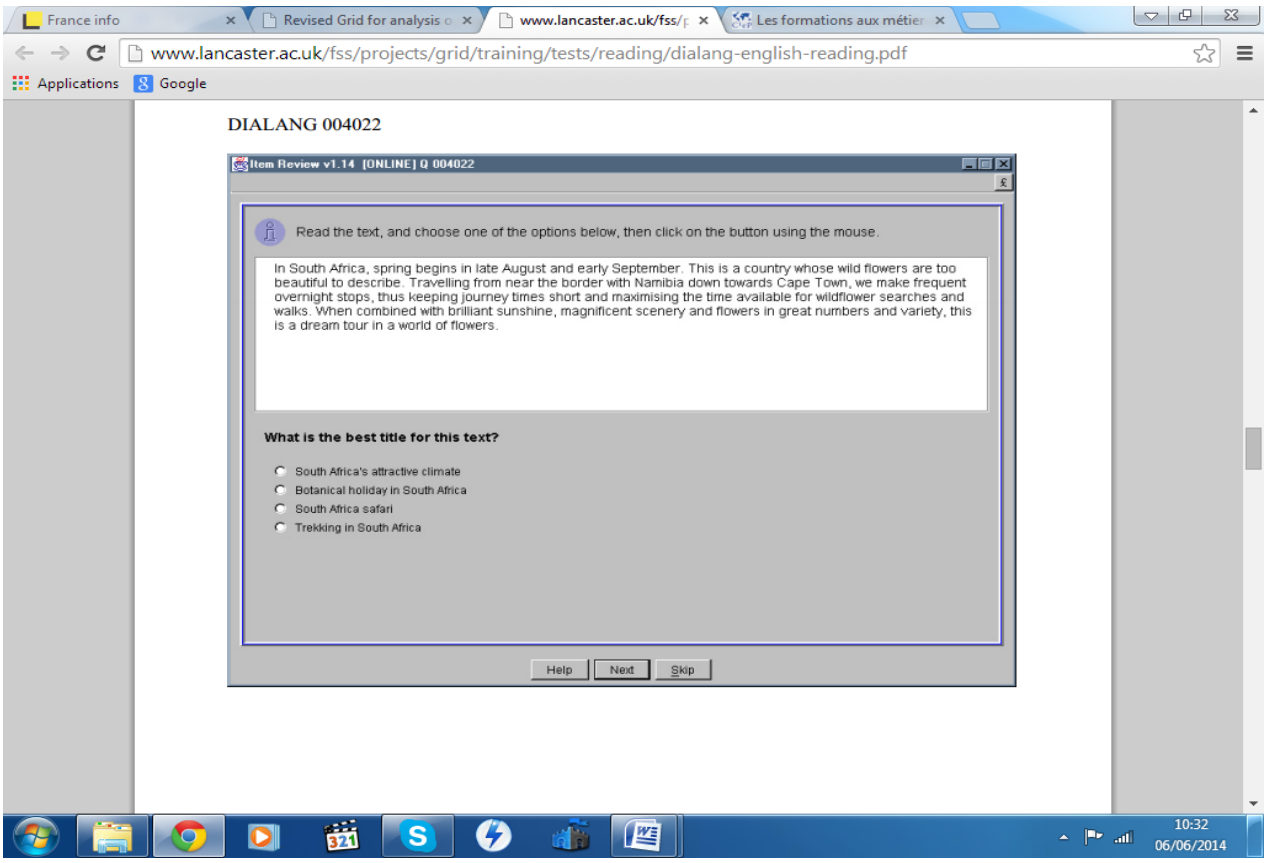
grammaire et de vocabulaire, et de préférence, également en termes d'aspects sociolinguistiques et pragmatiques de la compétence langagière. En outre, il est conseillé d'explorer l'interaction entre un ou plusieurs texte(s) et l'item constitutifs de la tâche (Alderson 2004 : 19). Enfin, il est recommandé d'ajouter un module d'entrainement afin de faciliter la compréhension du fonctionnement de la grille d'analyse. Cette suggestion a été mise en pratique et ainsi, le module d'entrainement est le dispositif censé expliquer l'usage de cet instrument aux utilisateurs, à côté de la grille échantillon. Le module d'entrainement permet tout d'abord d'analyser une ou plusieurs tâches en autonomie selon les dimensions de la grille puis de comparer son analyse avec celle d'experts tout à fait familiarisés avec le CECRL et la grille. En amont de l'analyse, l'utilisateur est invité à choisir une tâche selon ses propres besoins et ses objectifs d'apprentissage. Les tâches proposées relèvent d'un certain nombre de tests d'anglais, de français et d'allemand, conçus par diverses organisations situées dans quatre pays européens en l'occurrence, la Grande-Bretagne, la France, la Finlande et l'Allemagne:[67]

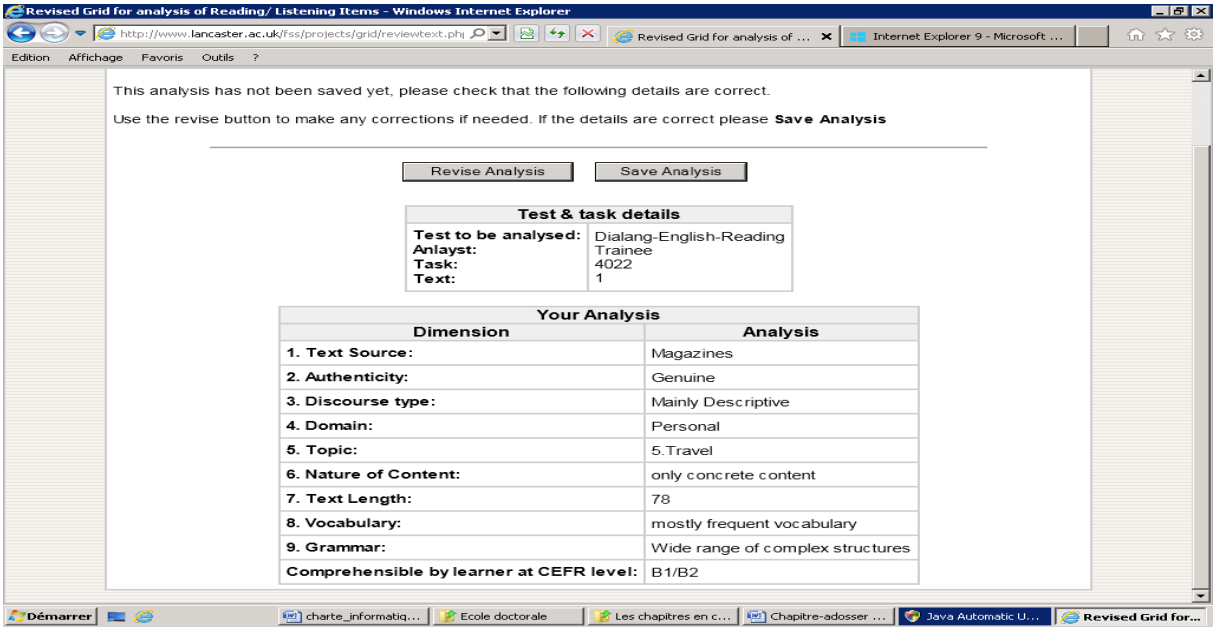


Pour illustrer le fonctionnement du module d'entrainement, la tâche suivante issue du DIALANG sera analysée. Cette tâche se compose d'un texte et

d'un item qui feront l'objet d'un examen à l'aide des grilles présentées ci-dessus. Notre analyse sera juxtaposée à celle effectuée par des experts pour révéler les dimensions divergentes de la tâche.

Le texte constitutif, présenté ci-dessous[68], a été choisi uniquement en raison de notre préférence personnelle pour le sujet traité qui est le voyage. La grille d'analyse du texte est placée ci-dessous [69]. Selon notre analyse, il s'agit d'un texte authentique issu d'un magazine. Le type de discours est descriptif et son contenu est de nature uniquement concrète. Quant aux caractéristiques linguistiques du texte, il comporte majoritairement un vocabulaire fréquent et un large répertoire de structures complexes. On l'estime compréhensible par des apprenants de niveau avancé, c'est-à-dire situés entre les niveaux B1 et B2.



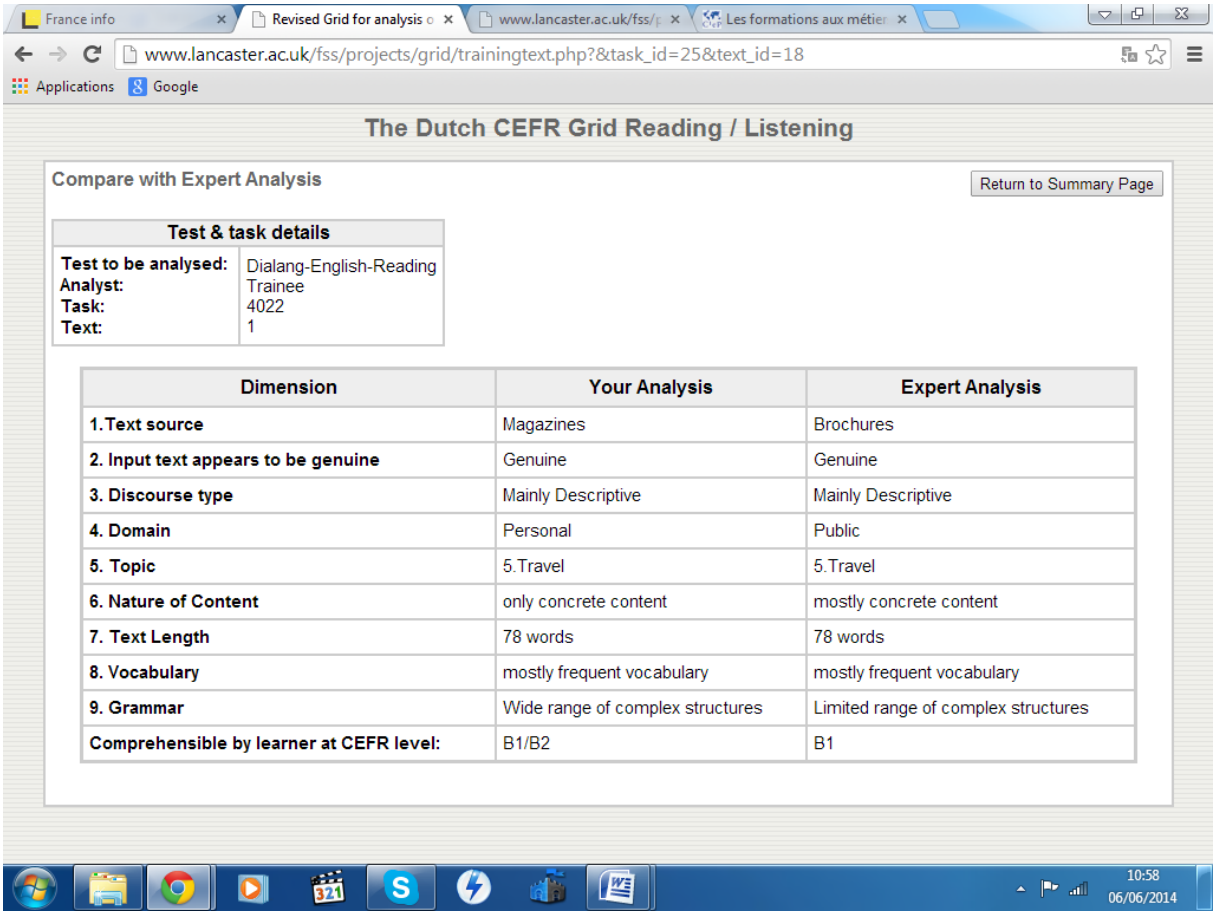


La grille ci-dessous représente la juxtaposition de notre analyse du texte, avec celle de l'analyse experte[70]. On constate des différences dans les quatre dimensions : la source du texte, le domaine, la nature du contenu et la grammaire. Concernant la source du texte, celle-ci est une brochure et non pas un magazine. Quant à la deuxième divergence, le texte provient du domaine public, contrairement à notre analyse de cette dimension. Pourtant, le classement de ce texte en domaine personnel nous paraît judicieux, compte tenu de la définition des domaines public et personnel dans le *Cadre européen commun de référence* (Conseil de l'Europe 2005 : 41); le voyage raconté par le narrateur fait en effet partie des « activités proprement individuelles » (Conseil de l'Europe 2005 : 41). Dans le cadre du voyage, le narrateur ne semble pas « engagé dans des transactions diverses pour des buts différents » (Conseil de l'Europe 2005 : 41). Néanmoins, le classement d'un texte dans différents domaines est possible dans certains cas étant donné qu'un « nombre de situations relèvent de plusieurs domaines » (Conseil de l'Europe 2005 : 41).

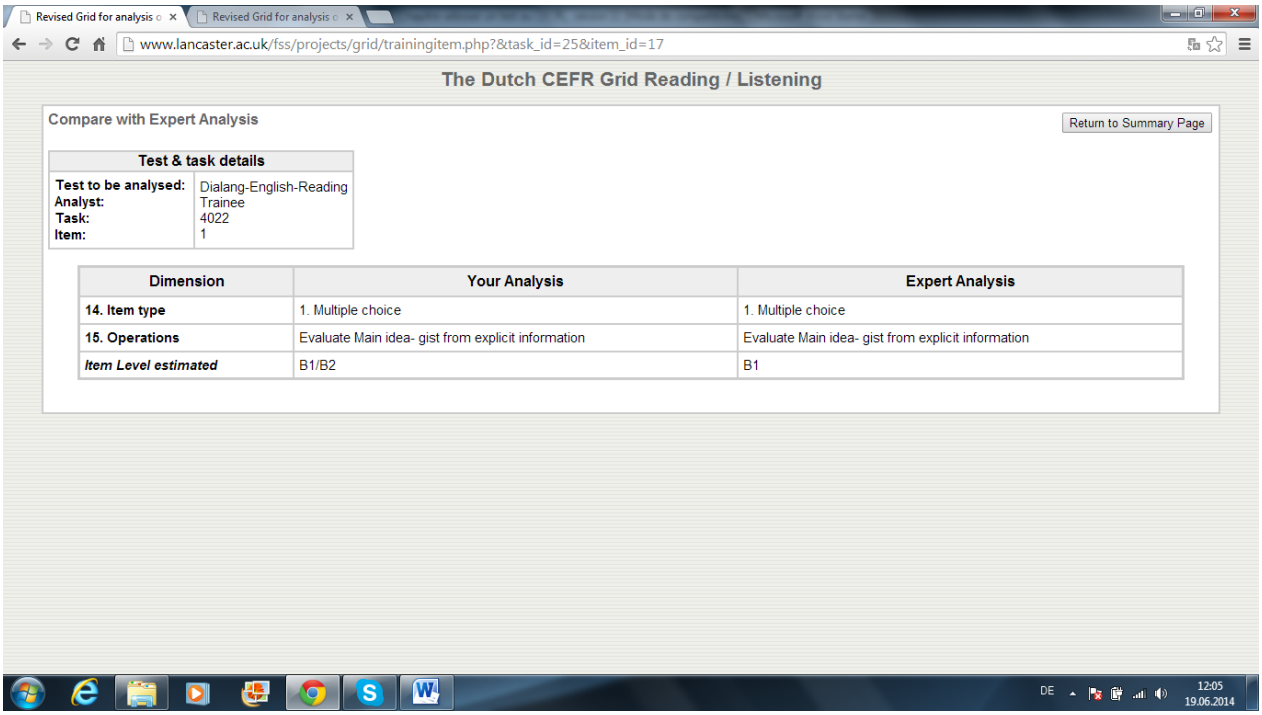
Contrairement encore à notre analyse, le contenu est considéré comme concret dans une large mesure mais pas uniquement. La dimension temporelle évoquée ne relève ainsi pas du concret, contrairement à la description des paysages. Le dernier paramètre faisant l'objet d'une analyse divergente, la grammaire, concerne la question de la complexité des structures. Selon notre analyse, le texte contient un large répertoire de structures complexes. Ce

jugement est dû à la présence de phrases contenant plusieurs subordonnées de types différents, en l'occurrence, des subordonnées relatives, circonstancielles de temps et de conséquence. La diversité des structures syntaxiques est accompagnée par une variété des formes verbales, à savoir le gérondif, le participe passé et l'indicatif. Néanmoins, les experts estiment que le répertoire grammatical est d'une complexité limitée car il n'atteint pas le maximum possible.

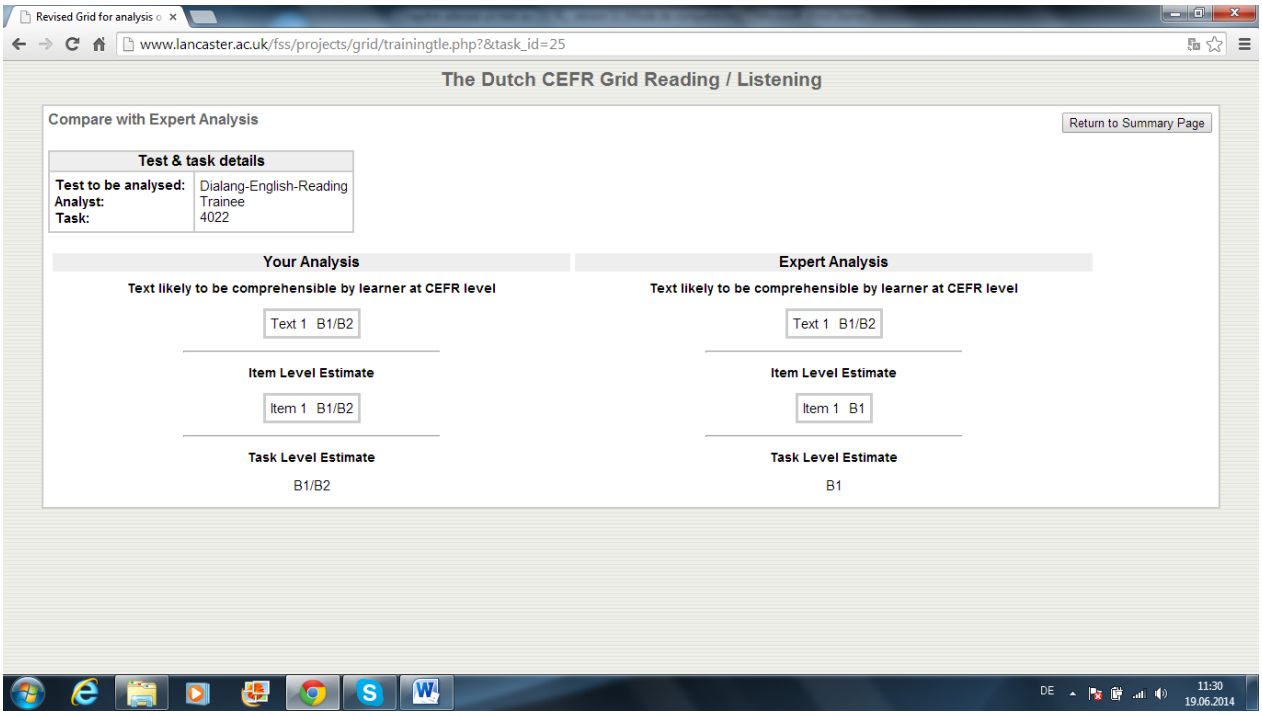
Le niveau estimé dans notre analyse ne coïncide pas avec celui effectué par les experts, même si la différence n'est pas considérable. Le niveau de compétence nécessaire à la compréhension de ce texte est estimé à celui des niveaux B1/B2, notamment en raison de la présence de plusieurs subordonnées et, par conséquent, de la complexité des phrases. Un bon contrôle grammatical est requis pour les comprendre, le même qu'exigé à partir du niveau B1+, et défini par le descripteur respectif au sein de l'échelle « Correction grammaticale » : « […] en règle générale, a un bon contrôle grammatical malgré de nettes influences de la langue maternelle » (Conseil de l'Europe 2005 : 90). Puisque le répertoire des structures complexes est perçu comme limité, les experts n'estiment pas que « le bon contrôle grammatical » soit nécessaire à leur compréhension. Ceci explique l'estimation du niveau de compétence au niveau B1 dans leur analyse. Le contenu concret et le vocabulaire majoritairement fréquent plaident en faveur du niveau B1 car ces caractéristiques réduisent la difficulté de compréhension induite par l'usage de phrases complexes. Le contenu concret du texte est en accord avec la compréhension attendue au niveau B1 : « Peut lire des textes factuels directs sur des sujets relatifs à son domaine et à ses intérêts avec un niveau satisfaisant de compréhension » (Conseil de l'Europe 2005 : 90).



Dans le cas de l'item, la seule divergence entre notre analyse et celle des experts concerne le niveau estimé. La grille contenant les deux analyses est présentée ci-dessous [71]. On voit que les deux analyses arrivent à la même estimation du niveau de compétences nécessaire pour répondre à l'item et comprendre le texte de la tâche. Cette estimation identique des niveaux de compétence s'explique par le lien très étroit de l'item avec le texte, de sorte qu'il n'est pas possible d'y répondre correctement en cas d'incompréhension du texte. Au-delà de la compréhension du texte, le stimulus de l'item demande à l'apprenant d'effectuer une opération cognitive d'évaluation. Bien que selon les résultats de la recherche portant sur le développement de la grille, les opérations demandées ne soient pas des indicateurs de niveaux de compétences d'un item, une opération d'évaluation est, évidemment, plus difficile à effectuer que celle qui consiste à reconnaitre les informations. Ce jugement n'est pas invalidé par la nécessité d'évaluer l'idée principale à partir de l'information explicite, ni par les détails recueillis à partir de l'information implicite.



La dernière phase de l'analyse consiste à estimer le niveau de la tâche. L'attention de l'apprenant est attirée sur la possibilité de passer à cette étape seulement après l'analyse de tous les textes et items. Cette phase demande d'entrer les niveaux de toutes les composantes de la tâche, ce qui sert à leur prise en compte lors de l'estimation du niveau de la tâche. La grille qui contient notre estimation des niveaux de compétence de toutes les composantes de la tâche, juxtaposée à l'estimation experte, est affichée ci-dessous. [72]



On voit que, dans notre analyse, la tâche est estimée au même niveau de compétences que celui jugé nécessaire à la compréhension du texte et à la réponse à l'item, à savoir au niveau B1/B2. Concernant l'analyse experte, on remarque une incongruité, car le niveau estimé nécessaire à la compréhension du texte est B1/B2 dans cette grille, alors que la grille d'analyse du texte présentée ci-dessus indique le niveau B1. L'estimation du niveau de la tâche est conforme aux estimations des niveaux de ses composantes, du texte et de l'item, discutées plus haut.[73] De ce fait, l'analyse de la tâche dans ce module d'entraînement se distingue de celle effectuée dans la grille échantillon. Dans cette dernière, le niveau estimé de la tâche coïncide uniquement avec celui de l'item.

2.5.6.5 Conclusion tirée après l'élaboration de la grille

La dernière étape du projet a révélé que les grilles élaborées se sont avérées utiles pour décrire et analyser non seulement des textes et des items, mais aussi des spécifications de tests (Alderson 2004 : 19). Malgré l'utilité de cet instrument, la difficulté de se mettre d'accord sur la signification et le mode d'utilisation d'un grand nombre de dimensions de la grille a été constatée, ce qui a révélé le besoin d'illustrer les grilles par des exemples afin d'assurer la compréhension identique des termes utilisés (Alderson 2004 : 18).

Le projet portant sur la conception des banques d'items était censé répondre au manque ressenti à la suite de l'élaboration de la grille d'analyse. Dès le début, la grille est considérée comme une grande aide pour mener à bien ce projet car elle permettrait une catégorisation adéquate des items en amont de leur usage dans les banques liées au CECRL.[74] Le projet, connu sous l'abréviation « EBAFLS » et financé par la Commission Européenne, a été mené en collaboration entre huit pays européens entre 2004 et 2007.[75] Il consistait en l'élaboration de banques d'items en vue de relier les tests et les examens nationaux au CECRL pour ensuite pouvoir les comparer objectivement. Il s'agit des items de compréhension écrite et orale issus des différents examens utilisés par un des pays participants. Les tests, analysés à l'aide de la grille présentée, sont rédigés en anglais, français ou en allemand et sont aux niveaux A2 à B2. L'usage de la grille n'est pas resté au stade du projet, mais cet instrument a effectivement été employé lors de l'élaboration des banques d'items. Avant d'être acceptés dans les banques, les items ont été classés selon les niveaux européens communs de référence à l'aide de la grille.[76] La classification des items et donc leur lien à un niveau commun de référence n'était pourtant pas suffisant pour les inclure dans une des banques.[77] Ils devaient aussi être typiques de la culture d'évaluation d'un pays donné sans pourtant mettre en jeu des connaissances spécifiques à la culture de ce pays. L'autre condition pour inclure les items dans les banques était la disponibilité des données empiriques relatives à ces items, ce qui présuppose leur utilisation préalable sur le terrain.[78] Ces critères, imposés à la sélection des items, étaient censés permettre la comparaison objective des tests, des certifications et des diplômes, et ainsi rendre l'évaluation en langues transparente, fiable et valide.[79]

La mise en application de la grille a démontré la complexité propre aux liens entre les tâches de compréhension écrite et orale et les niveaux communs de référence (Alderson 2004 : 21). En raison de la complexité constatée, on conseille d'appliquer la grille à une grande variété de textes et d'items, de comparer et d'analyser les résultats individuels de catégorisation ainsi que de comparer les niveaux de compétence estimés des items aux niveaux de difficulté sur la base des résultats empiriques (Alderson 2004 : 21). La complexité de relier les textes et les items aux niveaux communs de référence est aussi due à l'absence de liens entre les dimensions de la grille et les différents niveaux de

compétence. Pour cette raison, l'analyse du contenu des tâches doit être combinée avec l'investigation empirique de leur difficulté et avec les procédures de standardisation empiriques (Alderson 2004 : 21). [80]

2.6 Dialang et au-delà

Pour conclure ce chapitre, nous présenterons le système d'évaluation DIALANG dans ses différents aspects. Dû à son adossement au *Cadre européen commun de référence*, il a été décidé de situer cette partie à la suite de la discussion de ce référentiel.

2.6.1 Les origines et le développement de DIALANG

Le système d'évaluation en langues DIALANG a été édité en 2002. C'est un système d'évaluation développé sur la base des niveaux de compétence et des descripteurs du CECRL (Little 2007 : 649). En outre, les parties centrales de la définition de la compétence langagière incluses dans ce document officiel ont été mises en œuvre (Huhta et al. 2000 : 130). DIALANG a été conçu, monté et diffusé dans le cadre d'un projet subventionné par la Commission Européenne ainsi que par les vingt-cinq institutions participantes, notamment les universités, à travers l'Union Européenne (Alderson&Huhta 2005 : 301).

DIALANG a été développé en deux phases; la première de 1996 à 1999, et la seconde en 1999 (Alderson & Huhta 2005 : 306). Puisque ce système d'évaluation recouvre quatorze langues, autant d'équipes de développement ont été créées. Jusqu'à la fin de la Phase 1 de la conception du test, les banques d'items ont été conçues pour toutes les langues concernées, dont le nombre variait selon la langue de 525 à 3350, contenant plus de 2000 items en moyenne. En outre, les énoncés pour l'auto-évaluation ont été sélectionnés à partir du CECRL, modifiés, simplifiés dans certains cas, et traduits en 14 langues. Les tests de positionnement sur l'étendue du vocabulaire ont également été produits (Alderson & Huhta 2005 : 306).

2.6.1.1 Les problèmes rencontrés lors du processus de développement

Au cours du développement de DIALANG, plusieurs problèmes ont dû être résolus. Les raisons principales à cela sont la taille, la complexité ainsi que la nature novatrice de ce système d'évaluation. La grande complexité du développement de DIALANG est évidente, car il englobe des tests pour cinq domaines de compétences aux trois niveaux différents et ceci en quatorze langues (Alderson&Huhta 2005 : 308). DIALANG mérite son statut novateur pour plusieurs raisons; il s'agit en effet du premier test diagnostic directement adossé au CECRL (Alderson & Huhta 2005 : 308). Une autre qualité innovatrice repose sur le fait que ce test est automatisé, sécurisé et interactif (ibid : 309). L'interactivité se manifeste par le format adaptatif du test et par un large nombre de choix proposés aux candidats.[81] Les caractères innovants concernent également les méthodes utilisées et le bilan très élaboré. En revanche, les tâches elles-mêmes restent assez conservatrices (Huhta et al. 2002 : 132). C'est le cas même si les différents types de tâches sont contenus dans ce test, en l'occurrence, au format QCM et celles à réponse construite.

L'indépendance des échelles du CECRL par rapport à une langue particulière s'est avérée un inconvénient important lors de la conception de ce système d'évaluation, ce qui a conduit à l'ajout d'une base de données de la langue seconde au-dessous des descripteurs du CECRL (Byrnes 2007 :643). Les difficultés lors du processus de développement de DIALANG étaient également liées aux cultures d'évaluation différentes à travers l'Europe et au manque de pratiques qui portent sur le développement systématique des tests.

2.6.1.2 Le pilotage

Dès la Phase 1 de la conception du DIALANG, le plan de pilotage des différentes composantes du test a été produit, à savoir les items, les spécifications pour l'auto-évaluation ainsi que les items contenus dans le test de vocabulaire. Il a été décidé de les piloter sur Internet et d'utiliser les mêmes logiciels que ceux prévus pour l'administration ultérieure des tests opérationnels (Alderson & Huhta 2005 : 307). Pour permettre ce pilotage, 300 items, qui couvrent dans chaque langue les cinq domaines de compétences et les six niveaux distingués par le CECRL, ont été sélectionnés. Afin d'effectuer l'analyse initiale des items, la collecte de 100 réponses par item et la participation de 450

candidats ont été jugées nécessaires**.** Pour pouvoir effectuer l'analyse finale, il a été convenu de collecter 200 réponses par item (Alderson & Huhta 2005 : 307).

Au cours du pilotage, des problèmes ont émergé, comme par exemple la difficulté de trouver une quantité suffisante de candidats pour certaines langues. Ceci a empêché DIALANG de mettre en œuvre des tests empiriquement validés dans toutes les langues. Le développement des tests validés de manière empirique a été possible en anglais, français, allemand et espagnol (Alderson & Huhta 2005 : 309). Toutefois, les données empiriques ont été collectées pour les participants au pilotage de toutes les langues évaluées. Ces données mettent en évidence la variation de la population selon plusieurs catégories, à savoir, la langue native, l'âge, le niveau d'éducation, la durée d'apprentissage, le pourcentage d'usage de la langue cible et le niveau de compétences selon l'auto-évaluation. Malgré de larges variations, certaines caractéristiques des candidats sont plus fréquentes que d'autres, par exemple l'allemand en tant que langue native présente une tranche d'âge de 18 à 25 ans, un niveau d'éducation supérieur, et un usage de la langue cible une à deux fois par semaine (Alderson & Huhta 2005 : 310- 312).

2.6.1.2.1 Résultat du pilotage en anglais

Les participants au pilotage du test en anglais ne se distinguent pas considérablement de la population totale du pilotage en termes de caractéristiques personnelles et éducatives (Alderson & Huhta 2005 : 312). A la suite du pilotage, les items ont été calibrés en calculant leur niveau de difficulté à l'aide de la théorie IRT (Alderson & Huhta 2005 : 313).[82] Les résultats du calibrage montrent que la grande majorité des items ont pu être maintenus, et même la totalité dans le test de vocabulaire (Alderson & Huhta 2005 : 313). Ceci est dû tant à la qualité des items qu'au processus détaillé de la révision.

A la suite du pilotage et du calibrage des items, les standards devaient être fixés. Cette procédure implique l'attribution de niveaux de compétences aux scores. Puisqu'au moment de cette procédure, les données sur la performance des candidats à DIALANG ou à un test semblable n'existaient pas encore, la fixation des standards devait se fonder sur les jugements itératifs des experts concernant les items individuels (Alderson & Huhta 2005 : 315). Cette décision

reposait sur la question de savoir si les candidats à un niveau de compétence donné pouvaient répondre correctement à un item. Ces jugements, effectués par un groupe d'experts, ont été organisés en deux phases. En phase 1, six jugements, couvrant les six niveaux de compétences du CECRL, ont été énoncés sur chaque item. En phase 2, deux jugements ont été donnés, de même que la fiabilité de tous les juges impliqués a été calculée. À la suite de la fixation des standards, les seuils entre les différents niveaux de compétences ont été calculés à l'aide d'un programme informatisé. Ces seuils entre les niveaux sont utilisés pour l'estimation du niveau de compétences individuel des candidats (Alderson & Huhta 2005 : 316).

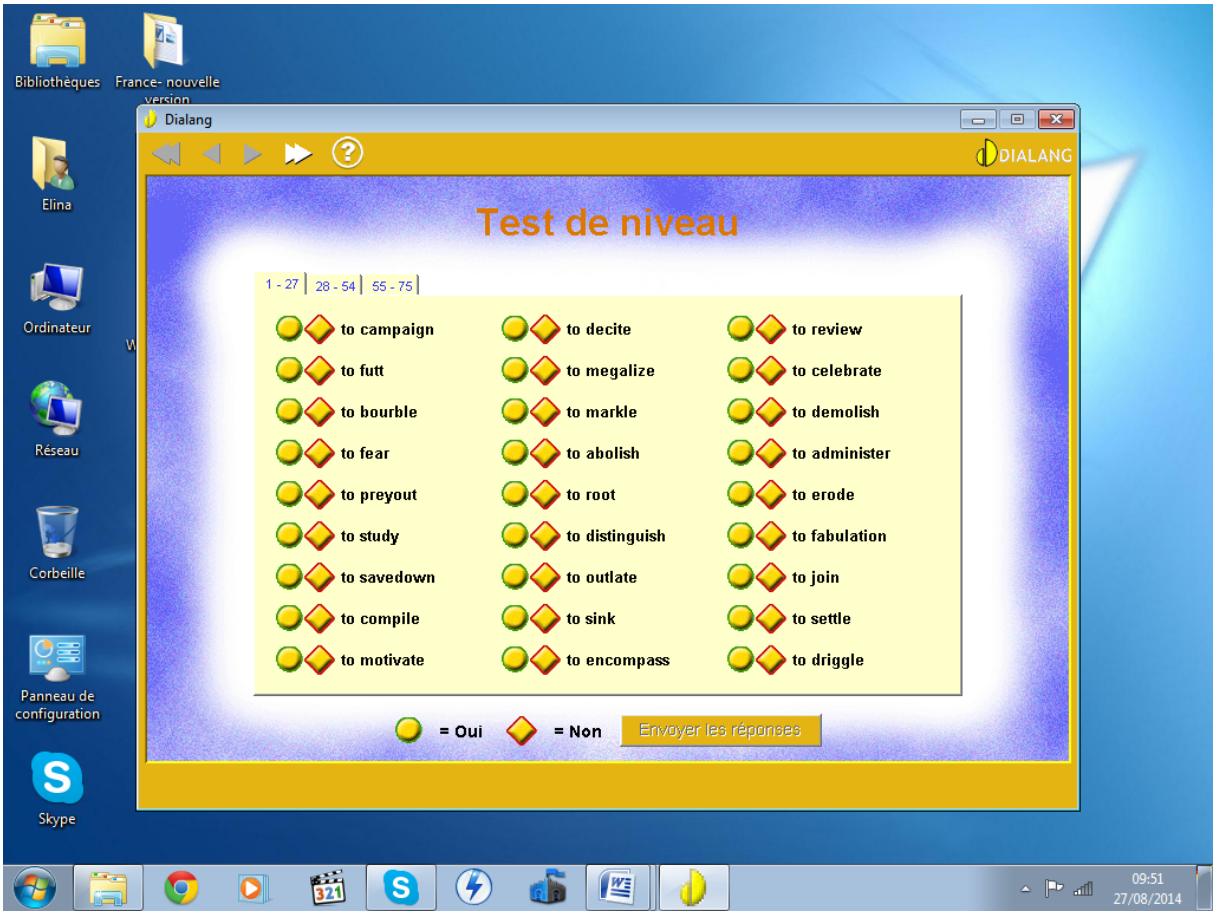
En plus des jugements théoriques de l'attribution des niveaux de compétences, les standards ont été fixés sur la base de la difficulté empirique des items. Ces deux procédures, les jugements théoriques, d'une part, et le calibrage, d'autre part, ont produit des niveaux très similaires dans le pilotage du test en anglais (Alderson & Huhta 2005 : 317). En outre, une corrélation étroite entre les compétences linguistiques et les activités de communication langagière a été repérée. Le premier cas de figure est la corrélation élevée entre la compétence grammaticale d'une part et la compréhension et l'expression écrites d'autre part. Le deuxième cas de figure est la corrélation entre le vocabulaire et, d'une part, la compréhension orale et, d'autre part, l'expression écrite. Ceci indique que les compétences grammaticales et lexicales jouent un rôle important dans les activités de communication langagière évoquées (Alderson & Huhta 2005 : 318). La corrélation est également observable entre l'auto-évaluation des candidats et les résultats au test. Elle est la plus élevée entre l'auto-évaluation et l'expression écrite, probablement en raison du feed-back reçu par les candidats lors de cette activité de communication (Alderson & Huhta 2005 : 317). Un retour fréquent peut mener à une auto-évaluation d'expression écrite plus juste (Alderson & Huhta 2005 : 318). Le pilotage du test en anglais a démontré que le travail de conception de ce système d'évaluation a été de très haute qualité, ce qui est primordial dans la perspective d'un usage à large échelle prévue pour ces tests (Alderson & Huhta 2005 : 320).

2.6.2 Type du test

DIALANG est un système de tests diagnostiques en quatorze langues européennes et en cinq domaines de compétences (Haahr & Hansen 2006 : 77): la compréhension de l'oral, de l'écrit, l'expression écrite, la grammaire et le vocabulaire. Le niveau global de compétence langagière n'est pas attribué par ce test, mais uniquement le niveau dans l'un des cinq domaines de compétences. Pourtant, il n'est pas affirmé que les domaines de compétences ne sont pas mutuellement liés (Huhta et al. 2002 : 135). Les langues proposées sont parlées en Europe occidentale (l'anglais, le français et l'allemand) en Europe du Sud (l'espagnol, l'italien, le portugais et le grec) ainsi qu'en Europe du Nord (le danois, le suédois, le néerlandais, le norvégien, l'islandais et l'irlandais). Le test est par ailleurs accessible gratuitement sur Internet. Ce test s'adresse notamment aux candidats adultes qui apprennent les langues soit en autonomie soit via des cours organisés. Bien qu'il se destine aux apprenants individuels, il peut être utilisé par les entreprises afin d'évaluer le niveau de compétences de leurs employés. Le public visé en priorité par la Commission Européenne est composé de jeunes adultes (Haahr 2004 : 80).

2.6.3 Conception de DIALANG

La conception de ce système d'évaluation prévoit neuf étapes qui ne sont pas toutes obligatoires dans le cadre de la démarche d'évaluation (Conseil de l'Europe 2005 : 162). La première étape concerne le choix de la langue utilisée pour les instructions, les spécifications pour l'auto-évaluation, pour les résultats du test et le feed-back. Après la description brève du test, qui constitue la deuxième étape, le candidat doit choisir la langue du test, parmi les quatorze proposées, ainsi que le domaine de compétences souhaité (Haahr 2004 : 81). Par la suite, un test de positionnement facultatif est présenté, dans lequel une liste de 75 mots est soumise aux candidats. Ces mots, dont cinquante sont réels et 25 inventés, sont placés dans un ordre aléatoire. Les candidats doivent décider pour chaque mot s'il existe ou pas dans la langue donnée, en répondant par « oui » ou « non » (Alderson & Huhta 2005 : 303). En cas de passation de ce test, le feed-back sur le niveau de performance est immédiatement communiqué aux candidats. Un extrait de ce test est présenté ci-dessous :



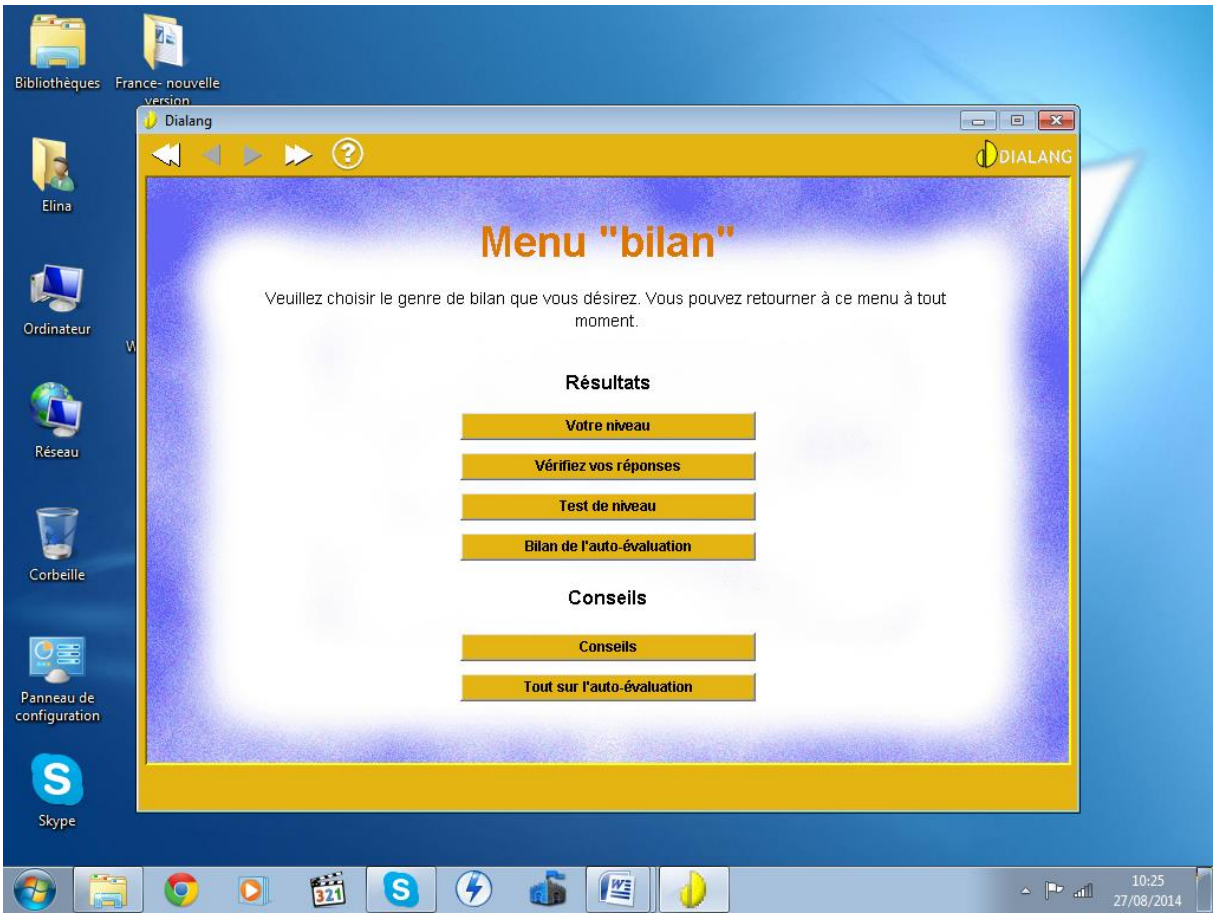
Le test de positionnement sur l'étendue du vocabulaire a été introduit dans DIALANG pour plusieurs raisons (Alderson & Huhta 2005 : 303-304), principalement parce qu'on doutait que l'auto-évaluation seule puisse pré-estimer le niveau de compétences d'un candidat de façon assez fiable pour présenter la version la plus appropriée du test au candidat (Alderson&Huhta 2005 : 303- 304). Pour compléter le résultat subjectif obtenu par l'auto-évaluation, il a été jugé nécessaire d'inclure un test objectif et court dans DIALANG. Le test de vocabulaire a été choisi parce qu'il satisfaisait à ces deux objectifs (Alderson & Huhta 2005 : 304). Ce type de test qui consiste à sélectionner « oui » ou « non » permet d'inclure un nombre élevé d'items, ce qui augmente la probabilité d'une pré-estimation précise du niveau de compétences des candidats. Ce test est considéré comme fiable, et est ainsi plus pondéré que l'auto-évaluation qui le suit (Alderson & Huhta 2005 : 304).

La phase suivante, également facultative, consiste en une auto-évaluation de la compétence langagière par les candidats dans le domaine choisi. Les apprenants doivent décider s'ils pensent pouvoir réaliser l'activité décrite dans

chaque énoncé. L'auto-évaluation est proposée seulement pour l'évaluation des compétences en activités de communication langagière, et non pour l'évaluation des compétences grammaticales ou lexicales. Ceci est dû à l'absence de descripteurs spécifiques à une langue particulière pour ces deux compétences linguistiques (Alderson & Huhta 2005 : 303). Pour chacune des trois activités de communication langagière, dix-huit spécifications qui relèvent du domaine de compétences choisi par l'apprenant sont proposées aux candidats. Il faut noter que les spécifications couvrent tout l'éventail des niveaux dans le domaine de compétences évalué. À la suite de l'auto-évaluation, les candidats ont la possibilité d'apprendre leur niveau de compétence soit immédiatement soit après le test dans le domaine de compétences choisi. Bien que cette étape soit facultative, elle est très importante car elle remplit plusieurs fonctions. La première de ces fonctions est d'ordre technique puisque le résultat à l'auto-évaluation détermine éventuellement, en plus du résultat au test de positionnement, la version du test qu'on va passer. En outre, le niveau de compétences autoévalué permet une comparaison avec le niveau évalué objectivement dans le test qui suit (Kaftandjieva & Takala 2002 : 106). L'auto-évaluation permet également de comparer la compétence langagière en différentes langues (Kaftandjieva & Takala 2002 : 106). A côté des fonctions techniques, l'auto-évaluation amène certains bénéfices sur le plan didactique (Haahr 2004 : 81). Devoir réfléchir à ses propres compétences au cours de l'auto-évaluation permet aux candidats d'apprendre plus efficacement (Haahr 2004 : 81).

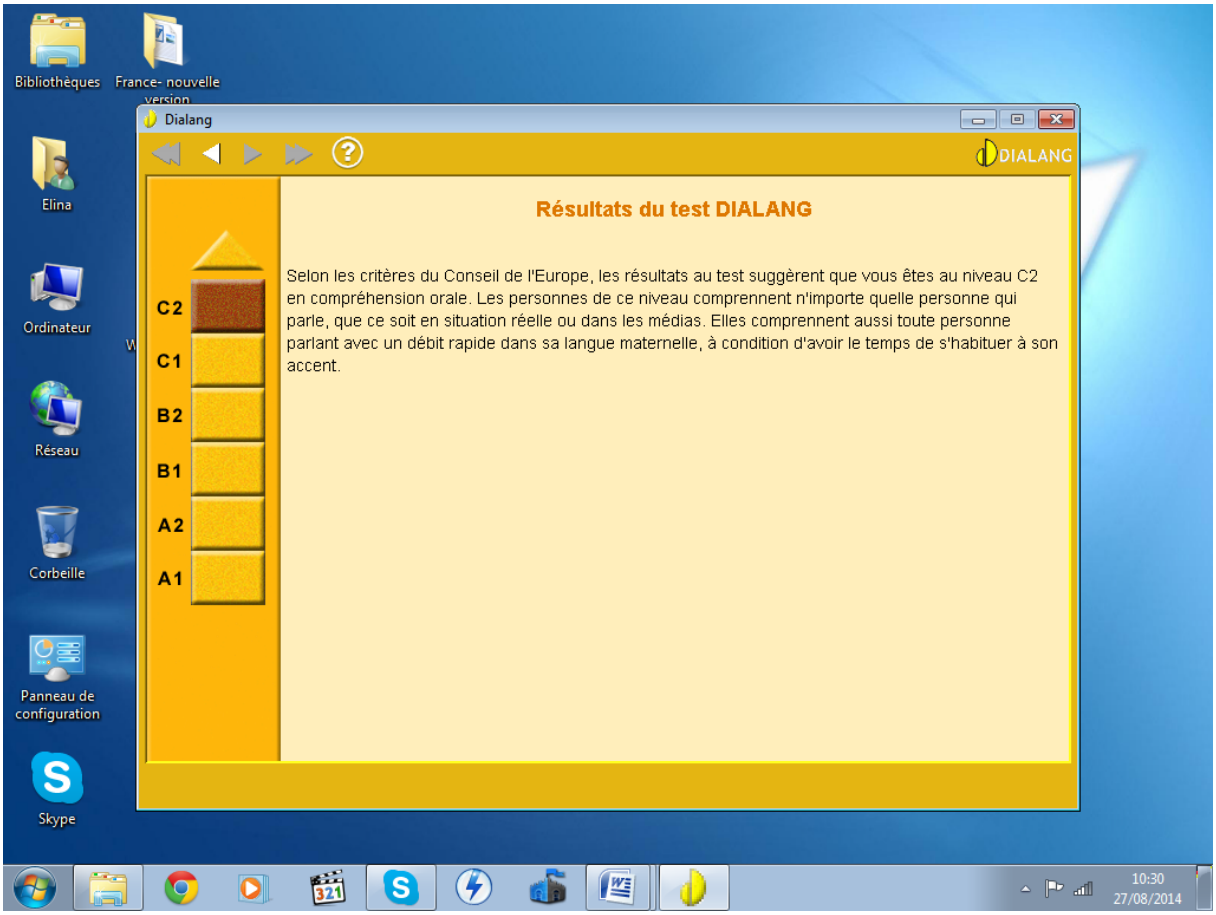
Par la suite, les candidats passent le test dans le domaine de compétences choisi et, en cas de passation de deux procédures de positionnement préalables, au niveau de compétences approprié. Il y a trente tâches à effectuer, quel que soit le domaine de compétences. Les items apparaissent en quatre formats, aussi bien à réponse fixe qu'à réponse construite : le QCM, les menus déroulants, les entrées de texte et les questions qui appellent une réponse construite (Alderson & Huhta 2005 : 304). Les candidats ne peuvent ni changer l'ordre des tâches, prévu par les concepteurs, ni leur réponse après être passés aux tâches suivantes. L'existence des trois versions de DIALANG qui recouvrent les différents niveaux de compétences aide

à atteindre deux objectifs poursuivis par ce système d'évaluation (Alderson & Huhta 2005 : 305). Le premier est de rendre la passation du test aussi pertinente que possible en présentant au candidat une version d'un degré de difficulté adéquat. Le deuxième objectif consiste à évaluer le niveau de compétences des candidats de façon précise sans que la procédure d'évaluation soit perçue comme trop fatigante par les candidats (Alderson & Huhta 2005 : 305). À la suite du test, un bilan complexe leur est proposé. Il englobe les résultats, composés de quatre parties, et des conseils, fournis en deux parties (Alderson 2005 : 34).



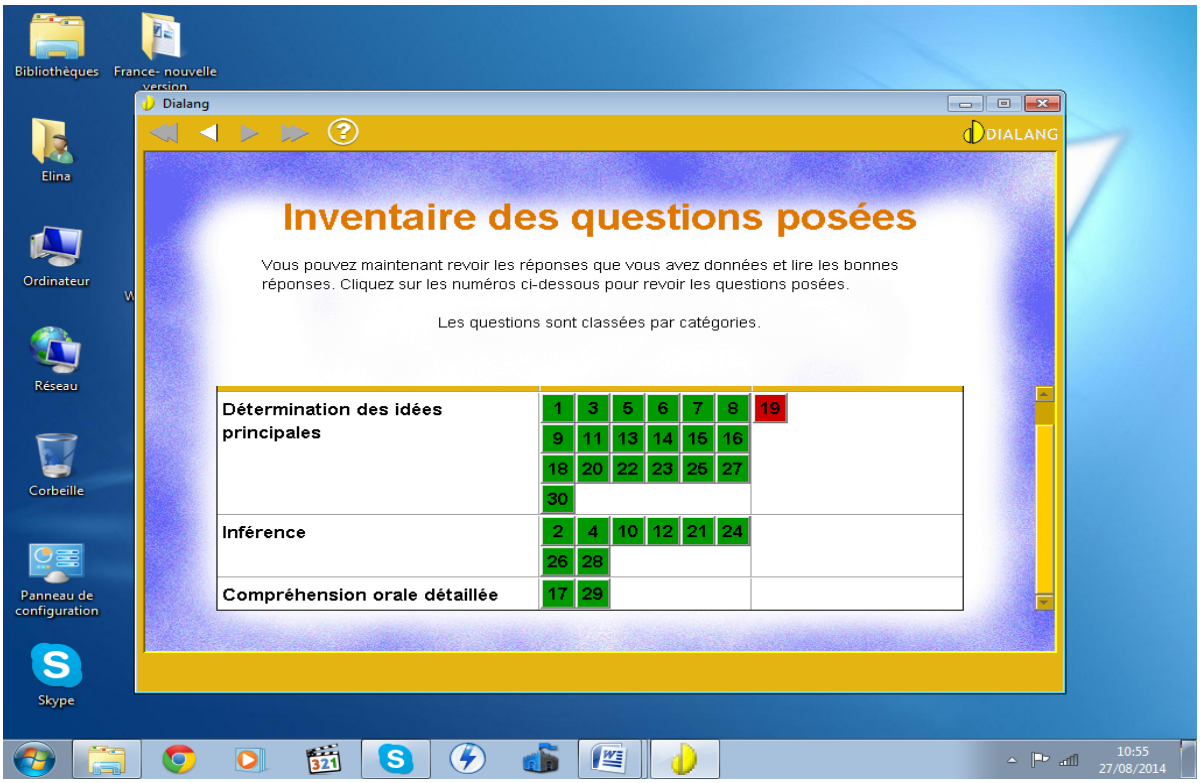
La première partie du bilan informe le candidat du niveau de compétences obtenu au termes des six niveaux couverts par le CECRL. En plus du niveau de compétences attribué, une brève description sur sa signification, issue du CECRL, est fournie (Alderson 2005 : 34). La description ne correspond pas exactement aux descripteurs des échelles de compétences du CECRL qui relèvent du domaine de compétence évalué dans le test, mais constitue un résumé et une reformulation des échelles pertinentes. Le score n'est pas fourni

aux candidats car selon certains chercheurs il ne serait pas utile dans un test diagnostic (Alderson 2005 : 34). [83]

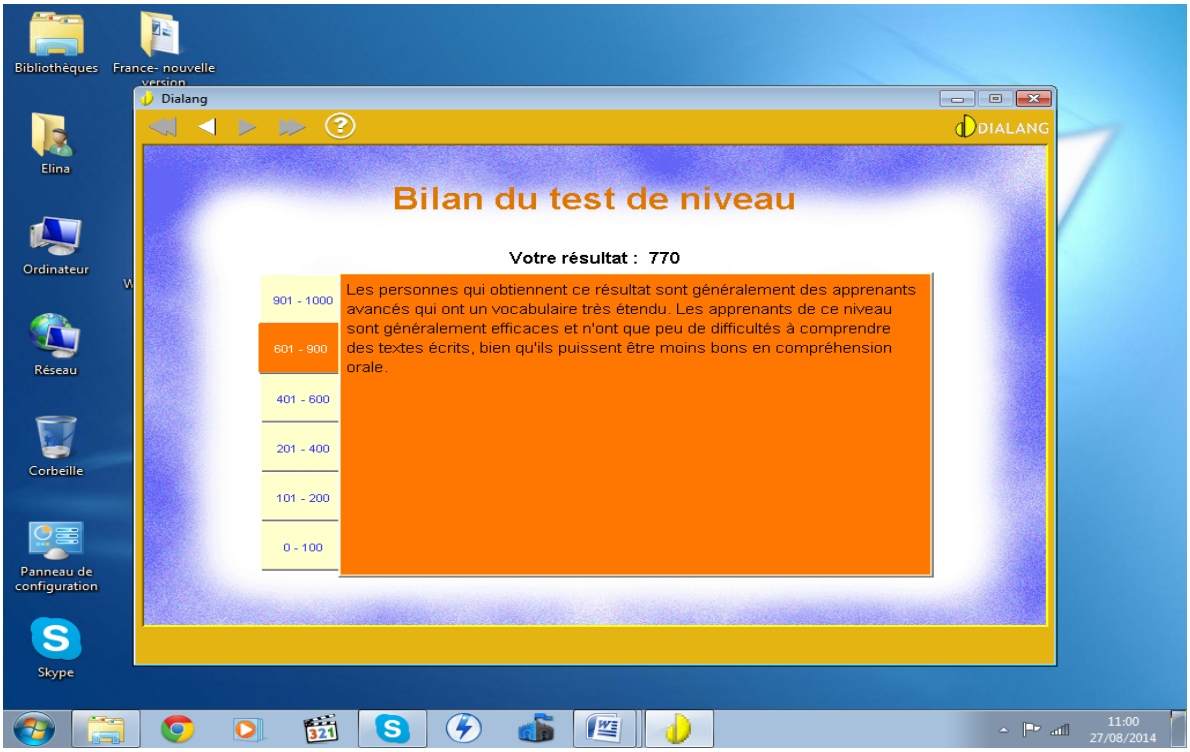


Puis, un retour plus détaillé sur la performance est donné au candidat, en soulignant les réponses correctes et incorrectes sous forme de tableau. Les items sont regroupés en fonctions des trois sous-compétences visées qui sont l'idée principale, le détail spécifique et l'inférence. Il a été décidé de renseigner les candidats sur ces points pour qu'il soient informés de leurs points faibles (Huhta 2002 : 478). En outre, ceci les rend conscients du fait que la compréhension se compose de facettes multiples. Bien que les candidats eux-mêmes jugent l'information sur les sous-compétences utile, les recherches explorant l'usage de cette information par les candidats dans la réalité manquent encore (Huhta 2002 : 478). Les résultats de la recherche, menée par Alderson (2005), indiquent que la performance aux sous-compétences évoquées ne varie pas en fonction des niveaux de compétences du CECRL (Alderson 2005 : 261). Les candidats aux différents niveaux de compétences ne se distinguent pas en matière de performance aux tâches qui relèvent de sous-compétences variées,

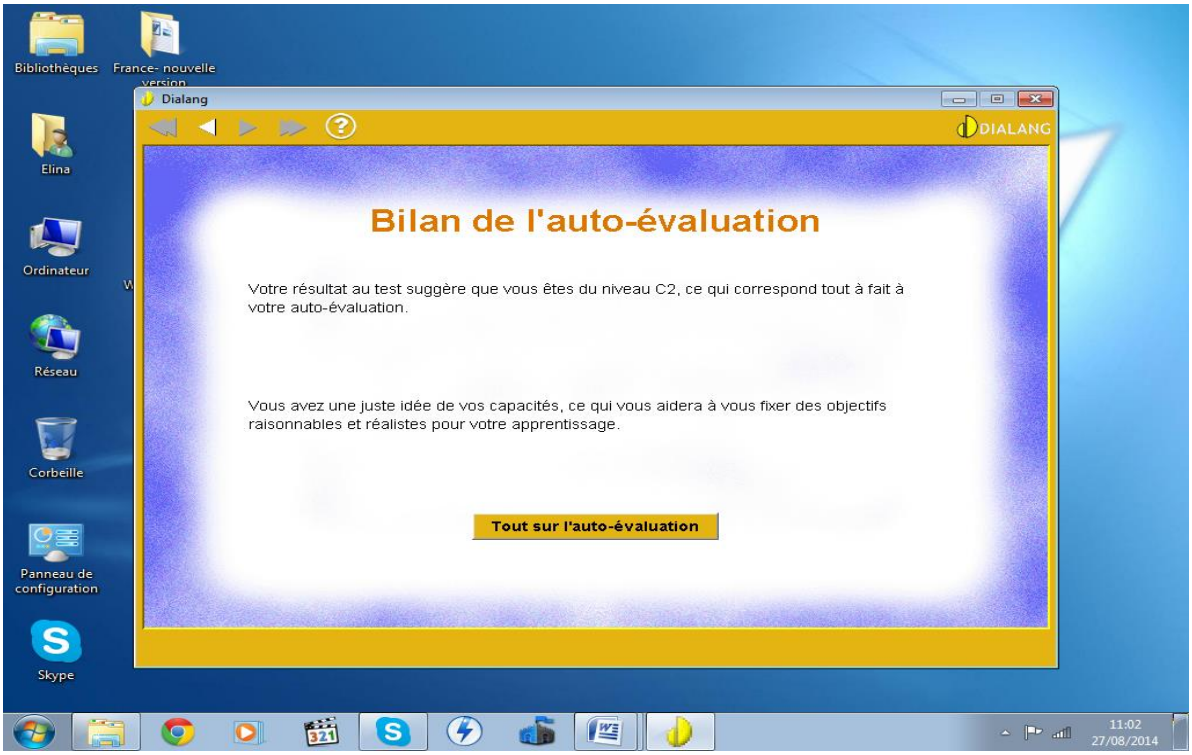
en l'occurrence de l'idée principale, du détail spécifique ou de l'inférence. Bien que l'inférence semble une sous-compétence plus difficile que les deux autres, sa maîtrise par les candidats au niveau A1 est la même qu'aux autres niveaux de compétences (Alderson 2005 : 261). Un exemple est exposé ci-dessous :



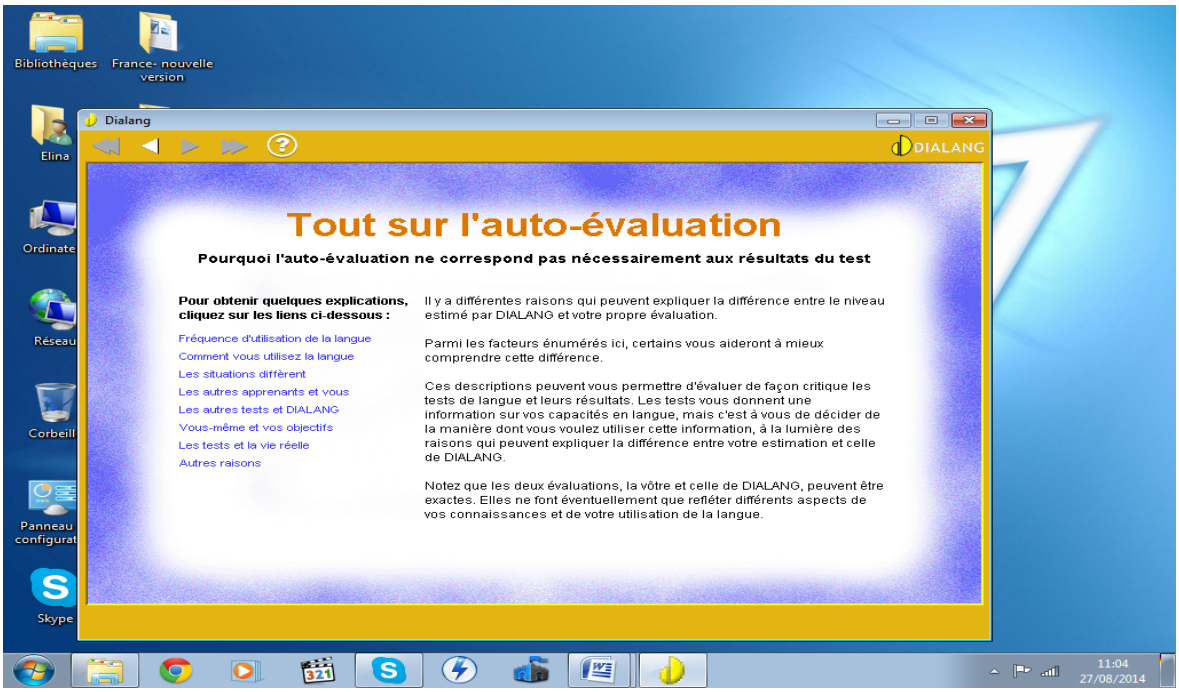
La troisième partie du feedback rapporte le score obtenu au test de vocabulaire sur une échelle de 1 à 1000, et donne une brève description de sa signification :



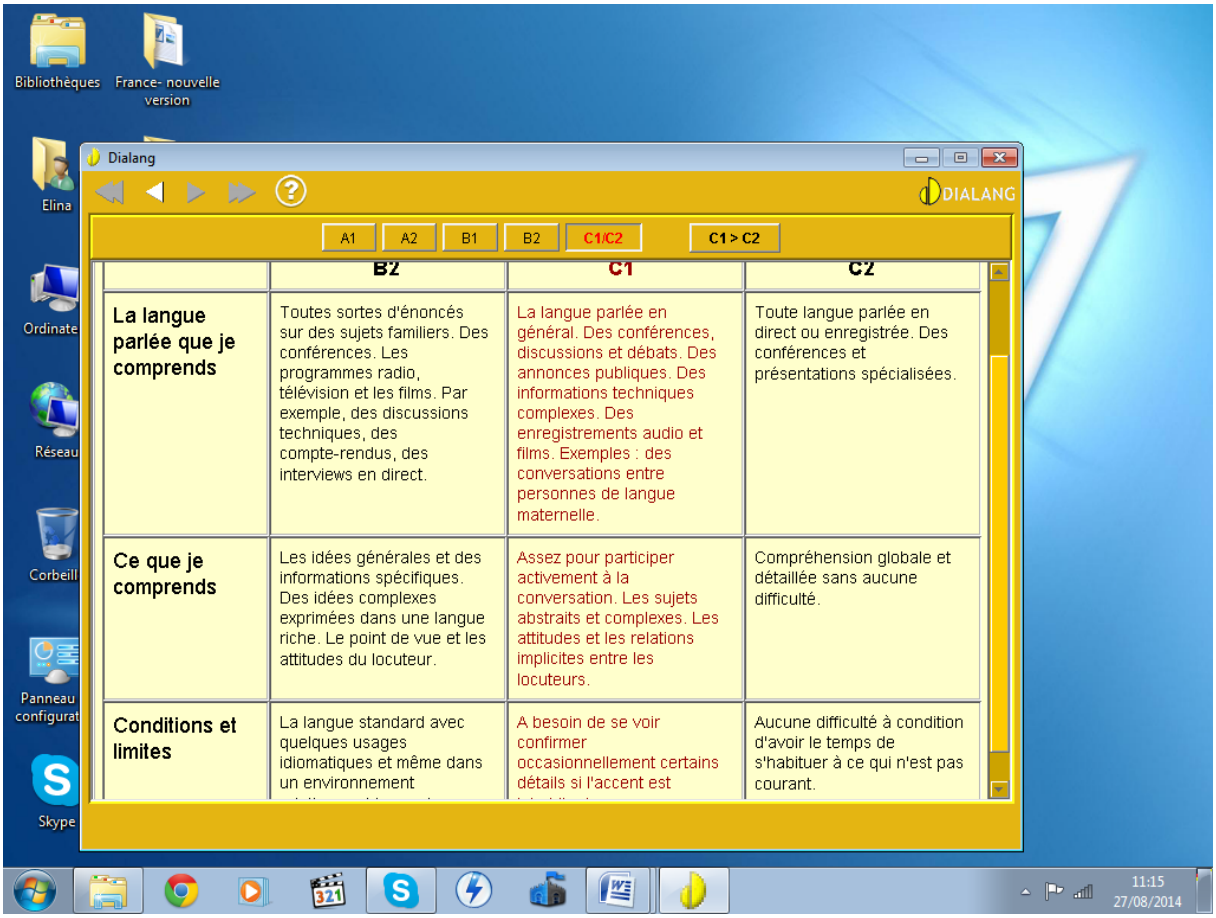
Dans la dernière partie, les candidats peuvent faire un retour sur leur résultat à l'auto-évaluation, qui est juxtaposé au niveau de compétences obtenu au test.



En cas d'écart entre ces deux résultats, les risques de sur ou de sous-évaluation ainsi que les raisons potentielles à ces dangers sont nommés. Cette partie du bilan offre aussi la possibilité de s'informer sur l'auto-évaluation, sur l'apprentissage et l'usage de la langue (Alderson & Huhta 2005 : 306).



La deuxième grande partie du bilan, appelée « Conseils », renseigne le candidat sur les différences entre son niveau de compétences dans le domaine évalué et les niveaux voisins. Les différences, présentées sous forme d'un tableau aux candidats, ne sont pas fondées sur un programme ou un cours particulier, mais s'appuient sur le CECRL (Haahr & Hansen 2006 : 77). Celles-ci sont regroupées en trois catégories: « Les types de textes que je comprends », « Ce que je comprends » ainsi que « Conditions et limites ». Conformément aux spécifications pour l'auto-évaluation, elles sont formulées à la première personne du singulier (Conseil de l'Europe 2005 : 170-172).



Ces échelles de démonstration, développées explicitement pour la partie « Conseils » de DIALANG, ont pour fonction d'inciter les apprenants qui se reconnaîtraient dans les descriptions à réfléchir sur l'apprentissage de la langue en général. Dans la deuxième partie des Conseils, intitulée «Tout sur l'auto-évaluation », un ensemble de suggestions est soumis aux candidats pour les aider à progresser de leur niveau de compétences actuel vers un niveau supérieur (Alderson & Huhta 2005 : 306). A première vue, cette partie semble aborder les raisons qui expliquent l'écart entre l'auto-évaluation et le résultat au test, par exemple la fréquence insuffisante d'usage de la langue ou une longue durée depuis son dernier usage. Toutefois, ces raisons, comme pratiquer régulièrement la langue à des fins de communication dans des contextes variés, font en même temps office de suggestions, destinées à améliorer le niveau de compétences.

2.6.4 Usage de DIALANG

DIALANG est le premier système d'évaluation à large échelle à des fins de diagnostic, c'est-à-dire, qui vise à déterminer les points forts et faibles dans la langue apprise (Haahr & Hansen 2006 : 46). À la différence d'autres tests à large échelle comme TOEIC, TOEFL ou IELTS, DIALANG ne délivre pas de certifications (Haahr 2004 : 83). Son but principal est donc d'informer les apprenants sur leur niveau de compétences acquis et leur apprentissage de la langue. Toutefois, il est indiqué dans les instructions qui précèdent le test que DIALANG peut également être utilisé pour répartir les étudiants dans différents cours de niveaux.[84] Ceci implique que DIALANG puisse remplacer un test de positionnement. La condition nécessaire pour l'usage approprié de DIALANG à des fins de positionnement est que le contenu des cours soit adossé aux niveaux de compétences et aux descripteurs du CECRL.

Un des buts majeurs de DIALANG est de permettre aux candidats de s'autoévaluer. Bien que l'autoévaluation soit facultative, cette étape joue un rôle très important au sein du test, pour de nombreuses raisons d'ordre technique, didactique et conceptuel. Par rapport aux raisons techniques, le résultat à l'autoévaluation détermine, éventuellement en association avec le résultat au test de positionnement, la version du test qu'on va passer (Haahr & Hansen 2006 : 46). En ce qui concerne les fonctions didactiques de cette étape, le niveau de compétences autoévalué permet une comparaison avec le niveau évalué objectivement dans le test suivant (Kaftandjieva & Takala 2002 :106). L'autoévaluation permet également de comparer la compétence langagière en différentes langues (Kaftandjieva & Takala 2002 : 106). La possibilité de comparaison entre les niveaux de compétences favorise la prise de conscience chez les apprenants, rendant ainsi l'apprentissage plus efficace.

L'autoévaluation contribue également à faire passer dans les esprits tous les concepts didactiques portés par le *Cadre européen commun de référence*. L'approche par l'évaluation, adoptée par DIALANG, est directement influencée par le concept d'apprentissage autodirigé, recommandé dans le CECRL (Huhta et al. 2002 : 130). L'autoévaluation est présentée comme un moyen qui aide les apprenants « à mieux gérer leur apprentissage » et donc, favorise

l'apprentissage autodirigé (Conseil de l'Europe 2005 : 145). En effet, cette procédure promeut la motivation et la prise de conscience chez les candidats, aidant ces derniers à reconnaître leurs points forts et leurs déficiences (Conseil de l'Europe 2005 : 145). De même que dans le CECRL, l'apprentissage autodirigé, l'autonomie et la motivation de l'apprenant constituent la base conceptuelle de DIALANG (Huhta et al. 2002 : 130).

2.6.5 Usage de l'auto-évaluation en général

Les points positifs, liés à l'autoévaluation dans DIALANG, n'impliquent pas que cette procédure possède seulement des avantages. Comme toute procédure d'évaluation, celle-ci présente des atouts et des inconvénients (Haahr&Hansen 2006 : 4). Le large éventail de compétences susceptibles d'être évaluées, ainsi que les coûts peu élevés nécessaires au développement et à la mise en œuvre des procédures d'auto-évaluation, constituent les points forts majeurs de celle-ci (Haahr&Hansen 2006 : 4). En revanche, la validité des résultats d'autoévaluation n'est pas toujours assurée.[85] Bien que la corrélation entre l'autoévaluation et les résultats au test soit observable non seulement lors du pilotage de DIALANG, mais également au cours d'autres procédures d'évaluation, celle-ci n'est pas parfaite (Hahr & Hansen 2006 : 4). On observe ainsi une tendance des candidats situés en bas sur l'échelle de compétence à surestimer leurs compétences lors de l'auto-évaluation. Pour cette raison, l'usage de l'autoévaluation comme seule méthode lors de l'évaluation des compétences chez les adultes est déconseillé (Hahr & Hansen (2006 : 4). En revanche, en combinaison avec l'évaluation directe, l'autoévaluation peut être utile parce qu'elle fournit une information sur l'auto-perception d'un candidat, qui peut mettre en évidence la motivation de celui-ci à s'engager dans les activités d'apprentissage (Haahr & Hansen 2006 : 5). L'autoévaluation représente un facteur corrélé positivement avec la fiabilité d'un test, à condition de respecter un certain nombre de principes (Bachman 1990 : 148). Les résultats de l'auto-évaluation correspondent davantage aux scores obtenus au test quand le contenu de l'auto-évaluation tient compte des besoins des candidats et des situations concrètes que si les phrases d'auto-évaluation sont formulées de façon abstraite (Bachman 1990 : 148). De plus, les questions d'autoévaluation qui

incitent les candidats à juger de la difficulté sur différents aspects de l'usage langagier semblent de meilleurs indicateurs de leurs compétences que les questions qui leur demandent de juger leur capacité à utiliser les différents aspects de la langue (Bachman 1990 : 148).

La procédure d'autoévaluation demande la conception et l'administration d'un questionnaire qui doit contenir des variables, liées au parcours personnel, éducatif ou professionnel, pertinentes pour les participants au test ainsi que des énoncés permettant de s'autoévaluer dans les domaines de compétences ciblées. De même que pour les tâches constitutives d'un test, le questionnaire élaboré à cet usage doit être piloté (Haahr & Hansen 2006 : 23). Conformément à ce principe, le questionnaire élaboré pour l'auto-évaluation dans POSILANG sera piloté sur le même échantillon de candidats que les tâches elles-mêmes.

2.6.6 Lien entre le CECRL et le système d'évaluation DIALANG

Le système d'évaluation DIALANG est fondé sur le CECRL dans une large mesure. Le cadre pour l'évaluation de DIALANG, l'échelle d'évaluation et les spécifications qui servent à informer les candidats de leurs résultats sont fondés directement sur les niveaux communs de référence du CECRL (Huhta et al. 2002 : 130). Ce document a été choisi en tant que base pour DIALANG pour deux raisons. La première est l'approche fonctionnelle du CECRL de la compétence langagière, dont l'implémentation a été visée par les équipes du projet DIALANG (Huhta et al. 2002 : 130). L'approche fonctionnelle considère les compétences langagières comme des moyens qui servent à accomplir les tâches en contexte social (Conseil de l'Europe 2005 : 15). La deuxième raison majeure dans le choix du *Cadre européen commun de référence* est la notoriété et l'acceptation large de ce document en Europe (Huhta et al. 2002 : 130).

2.6.6.1 L'usage du CECRL dans le cadre d'évaluation de DIALANG

Le cadre d'évaluation de DIALANG définit le cadre théorique qui est destiné à être opérationnalisé dans le test. Il est directement fondé sur les parties du CECRL, en l'occurrence sur ses concepts théoriques, ainsi que sur le répertoire des tâches communicatives, des thèmes, des activités de communication langagière, des types de textes et des fonctions langagières décrites dans le chapitre 4 du CECRL (Huhta et al. 2002 : 132). Le cadre d'évaluation de DIALANG est présenté ci-dessous:

Figure 1. Contents of the Dialang Assessment Specifications (DAF)

A. THE CONTEXT OF LANGUAGE USE

B. COMMUNICATIVE TASKS AND PURPOSES

C. COMMUNICATION THEMES / CONTENT

D. COMMUNICATIVE LANGUAGE ACTIVITIES

E. TEXTS

F. COMPETENCES

G. COMMUNICATIVE FUNCTIONS

H. STRATEGIES

I. SCALE OF PROFICIENCY

REFERENCES

Appendix 1 Specifications related to **communicative tasks and purposes** in the Council of Europe Threshold and Vantage level publications

Appendix 2 Specifications related to **themes and specific notions** in the Council of Europe Waystage, Threshold, and Vantage level publications

Appendix 3 Specifications related to **communicative activities** in the Council of Europe Waystage, Threshold, and Vantage level publications

Appendix 4 Specifications related to **texts** in the Council of Europe Waystage, Threshold, and Vantage level publications

Appendix 5 Specifications related to **functions** in the Council of Europe Waystage, Threshold, and Vantage level publications

(Huhta et al. 2002: 133).

On voit que le cadre d'évaluation de DIALANG s'organise en neuf sections, de A à I, qui n'est que brièvement décrit par le texte. Sa brièveté est liée à l'usage des définitions du CECRL dans les titres de ces sections qui ne sont pas expliquées, leur connaissance par les utilisateurs de ce document étant présupposée. Ceci rend crucial la nécessité d'une connaissance détaillée du CECRL, qui inclut également les concepts moins connus de ce document, comme par exemple le « texte » ou les « fonctions communicatives ». L'annexe au cadre d'évaluation DIALANG fournit des exemples (Huhta et al 2002 : 133). L'annexe 1 du cadre d'évaluation DIALANG, focalisée sur les tâches et les usages communicatifs, est présentée en annexe à ce chapitre. Pour une meilleure visibilité du progrès en compétence langagière entre deux niveaux voisins, le niveau seuil (*Threshold)* et le niveau avancé (*Vantage*) sont présentés ensemble (Huhta et al 2002 : 146). La mise en relief de certains mots et expressions sert également à une meilleure visibilité des termes centraux. Il ressort de ce document que les apprenants à ces deux niveaux de compétences sont capables de faire face aux situations quotidiennes de la vie courante (Huhta et al. 2002 : 146). Cependant, au niveau avancé (B2), ces situations peuvent être imprévisibles et même problématiques et requièrent un comportement flexible de la part de l'apprenant. L'usage de la langue est adapté à cette différence entre les situations maîtrisées à ces deux niveaux de compétences. Tandis qu'au niveau seuil (B1), l'usage de la langue est prévisible dans une large mesure, au niveau avancé (B2), le répertoire langagier élargi permet d'exprimer les intentions communicatives des apprenants de manière plus précise. Quant aux situations officielles, la différence entre les situations maîtrisées à ces deux niveaux de compétences ainsi que l'écart entre les tâches communicatives ne sont pas révélées dans ce document (Huhta et al. 2002 : 146). L'omission du niveau intermédiaire (*Waystage*) dans cette annexe s'explique par l'absence de spécifications des tâches et des usages communicatifs à ce niveau de compétences (Huhta et al. 2002 : 146). Les annexes ont été jugées utiles par les concepteurs de DIALANG, grâce à l'exemplification des concepts apportée, mais aussi grâce à la manière dont le texte est présenté ce qui permet de comparer les définitions des concepts clés du CECRL aux trois différents niveaux de compétences (Huhta et al. 2002 : 134).

En ce qui concerne la première section incluse dans le cadre d'évaluation de DIALANG, le contexte de l'usage langagier, l'approche actionnelle a été adoptée par DIALANG. Celle-ci a été choisie pour ce système d'évaluation en raison de la grande acceptation de l'approche actionnelle en Europe et de son adaptabilité au système multilingue. Malgré cette approche, il est possible de mettre en œuvre seulement une petite sélection de situations d'usage langagier dans DIALANG, puisque le contexte d'évaluation physique dans le test est l'ordinateur et que la procédure d'évaluation n'implique pas l'interaction avec d'autres personnes (Huhta et al. 2002 : 133).

Tandis que tous les concepts centraux du CECRL sont théoriquement adoptés par DIALANG, la vérification et la validation empirique de leur usage n'ont pas été possibles. Ceci concerne en premier lieu les compétences générales de l'individu ainsi que les trois composantes principales de la compétence communicative langagière, en l'occurrence les compétences linguistiques, sociolinguistiques et pragmatiques (Huhta et al. 2002 : 133). En deuxième lieu, ce constat concerne les stratégies utilisées par les apprenants lors des activités de communication langagière. Les concepteurs de DIALANG s'accordent à dire que les apprenants recourent à des stratégies dans le but d'accomplir les tâches communicatives. Toutefois, il n'a pas été possible de tenir compte de ces stratégies lors de la conception de DIALANG (Huhta et al. 2002 : 133-134).

Contrairement aux compétences et aux stratégies, les domaines contextuels utilisés dans DIALANG sont bien connus. Les domaines personnel et public sont le centre d'intérêt de ce système d'évaluation et couvrent ainsi la plupart des situations d'usage langagier. Le domaine occupationnel est aussi pertinent, même si les situations sont assez générales et non spécifiques à une profession quelconque, en raison de la nature généraliste de ce système d'évaluation. En revanche, le domaine éducatif est moins pertinent pour DIALANG, car ce système d'évaluation n'est pas lié à un seul contexte éducatif, mais convient plutôt à tous. Pour cette raison, il a été décidé de faire usage des opérations les plus génériques qui relèvent de ce domaine parmi celles énumérées dans le CECRL (Huhta et al. 2002 : 134).

2.6.6.2 Spécifications pour l'auto-évaluation

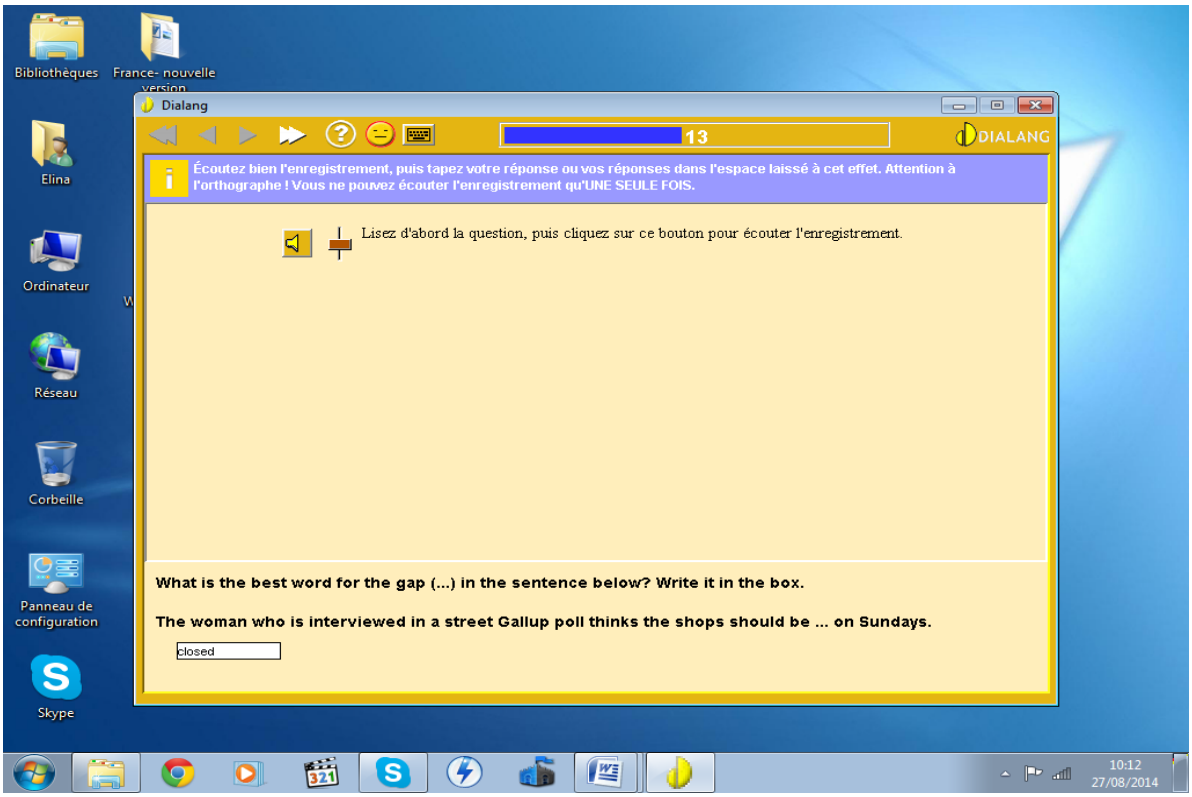
Les spécifications pour l'auto-évaluation s'appuient également, en grande majorité, sur les descripteurs du CECRL (Alderson & Banerjee 2002 : 81). Une partie seulement des spécifications a été reformulée. Dans ce cas, le verbe a été mis à la première personne du singulier, tandis qu'il se trouve à la troisième personne dans les descripteurs du CECRL. Cette reformulation des spécifications a été nécessaire en vue de permettre aux candidats de s'autoévaluer. Une autre partie des spécifications a été simplifiée afin de les rendre plus compréhensibles par les candidats au test. La structure de toutes les spécifications est la suivante : « Je peux … faire quelque chose ». Il y a un petit nombre de spécifications qui ont été rédigées exprès pour l'autoévaluation dans DIALANG et qui ne font donc pas partie des descripteurs du CECRL (Breakthrough 2001 : 10). Elles ont été ajoutées afin de satisfaire aux besoins spécifiques du système d'évaluation DIALANG, chaque fois que le contenu du CECRL a été jugé insuffisant (Conseil de l'Europe 2005 : 162). Les nouvelles spécifications se trouvent uniquement aux niveaux C1 et C2, comme par exemple : « Je n'ai pas besoin de faire vérifier ou relire ce que j'écris, sauf s'il s'agit d'un texte important » (Conseil de l'Europe 2005 : 166). La raison pour la présence des spécifications ajoutées seulement à ces deux niveaux est le nombre beaucoup plus élargi des tâches communicatives que les apprenants doivent maîtriser aux niveaux C1 et C2.

2.6.7 Format de DIALANG

La version courante de DIALANG propose un format adaptatif par rapport au niveau du test, ce qui implique que le système choisit le test de difficulté appropriée au niveau individuel de l'usager sur la base des réponses du candidat (Huhta et al. 2002 : 131). Ainsi, les tests dans chaque langue contiennent trois versions différentes – facile, intermédiaire et difficile – pour chacun des cinq domaines de compétence évalués. Le choix de la version présentée au candidat se fait sur la base du test de positionnement, qui évalue l'étendue du vocabulaire et de l'auto-évaluation préalables. En cas d'omission de ces deux étapes optionnelles, la version intermédiaire est soumise au candidat. La deuxième version de DIALANG, dont la mise en œuvre est en cours, sera adaptative par

rapport aux items individuels. Ceci signifie que le choix d'items proposés à un candidat se fondera sur ses réponses aux items préalables (Huhta et al. 2002 : 131). Puisque dans cette version du test les items seront tous adaptés au niveau individuel de chaque candidat, l'estimation très approximative du niveau de compétence d'un candidat suffira afin de commencer le test. Pour cette raison, l'auto-évaluation constituera la seule base pour ce premier placement des usagers et le test de vocabulaire deviendra superflu (Huhta et al. 2002 : 131).

Le format automatisé est si développé dans ce test qu'il permet d'évaluer le niveau en expression écrite non seulement par les tâches au format QCM, mais également à l'aide des tâches à réponse construite (Douglas & Hegelheimer 2008 : 122). Celle-ci se limite systématiquement à un mot, comme par exemple dans la tâche suivante :



L'usage du format automatisé a paru judicieux dès le début afin de fournir aux candidats le feedback diagnostique et les conseils concernant leurs compétences (Alderson & Huhta 2005 : 308). Il a paru évident aux équipes de développement que le recours aux tests sur support papier ne serait pas

compatible avec les objectifs du test pour des raisons organisationnelles. En outre, il a été décidé de se servir d'Internet en tant qu'environnement pour la conception du test, pour son pilotage et pour son administration opérationnelle (Huhta et al. 2002 : 131). Il est en effet plus facile de mettre à jour le contenu et les logiciels d'un test administré sur Internet que depuis un dispositif extérieur (Alderson & Huhta 2005 : 308). Le logiciel utilisé pour le développement de DIALANG est JAVA (Huhta et al. 2002 : 131)

2.6.8 Remarques conclusives

Le projet DIALANG est très apprécié par les chercheurs qui le jugent digne d''intérêt autant que d'admiration en raison principalement de sa nature innovatrice déjà soulignée (Douglas & Hegelheimer 2008 : 122-123). Dans ce système d'évaluation, une multitude de qualités nouvelles sont intégrées de façon unique et forment un ensemble cohérent (Aderson & Huhta 2005 : 309). Les caractéristiques innovatrices concernent aussi bien les fonctions multiples que le format de DIALANG, explicité ci-dessus. Les fonctions de ce dispositif, l'évaluation équivalente en quatorze langues ainsi que la proposition de feedback très complexe et pratique sont notamment valorisées (Douglas & Hegelheimer 2008 : 123). En revanche, les tâches elles-mêmes sont assez conservatrices (Huhta et al. 2002 : 132). Ceci est lié à la focalisation des tests actuellement en cours sur l'évaluation des compétences réceptives, en raison de l'évaluation automatisée des résultats. Pourtant, un grand nombre d'items expérimentaux a été créé par les équipes du développement de DIALANG qui prévoient de les incorporer dans des tests opérationnels. Ce genre d'items suppose la comparaison des réponses individuelles des candidats avec les réponses standardisées entrées dans le système d'évaluation (Huhta et al. 2002 : 134).

Une fonction pédagogique très importante est la responsabilisation des candidats. Bien que ce système d'évaluation contienne des tests objectifs, il rend les candidats responsables du processus d'évaluation (Alderson & Huhta 2005 : 302). La responsabilisation des usagers est mise en place grâce à un grand nombre de choix à faire au cours de la passation du test. Ces choix concernent la passation ou non du test de vocabulaire, l'auto-évaluation, la langue utilisée dans les instructions, la langue et le domaine des compétences évalués. La liberté de

décision des candidats s'étend également aux résultats et au feed-back proposés, car les candidats peuvent choisir les parties du feedback et du conseil qu'ils souhaitent consulter (Alderson & Huhta 2005 : 302). Enfin, la possibilité de passer autant de tests que souhaités dans les cinq domaines de compétences en quatorze langues fait également partie de la liberté proposée aux usagers de DIALANG.

En termes de format, il est apprécié que ce test soit automatisé tout en étant « low-stake », c'est-à-dire en ayant un impact faible sur l'avenir des candidats (Douglas & Hegelheimer 2008 : 122-123). DIALANG ne délivre en effet pas de certifications. Par conséquent, les résultats au test n'ont pas de répercussions importantes sur la vie professionnelle des candidats (Haahr 2004 : 83).

Notes:

[1] Les activités dans les champs de l'éducation et de la culture sont menées sous la direction du Conseil de la coopération culturelle (CDCC) qui est le Conseil au sein du Conseil de l'Europe. A l'heure actuelle, le Conseil de l'Europe comprend quarante-sept États signataires de la Convention culturelle européenne : les quarante et un Etats membres du Conseil de l'Europe, le Saint-Siège, le Bélarus, Monaco, la Bosnie-Herzégovine, l'Arménie et l'Azerbaïdjan. Le nombre si grand des membres du CDCC implique la diversification linguistique et culturelle du Conseil de l'Europe encore plus importante qu'à l'époque de la fondation de cette institution (Van Ek 2001 : ii).

[2] Conseil de l'Europe (2e édition, 2005). *Cadre Européen Commun de Référence pour les Langues. Apprendre, Enseigner, Evaluer*. Dans cette thèse, l'acronyme CECRL sera privilégié pour faciliter la lecture.

[3] Conseil de l'Europe. « Recommandation n° R (82) 18 du Comité des Ministres aux États Membres concernant les langues vivantes » (1982). Annexe A de Girard et Trim (1988).

[4] Conseil de l'Europe. « Recommandation n° R (82) 18 du Comité des Ministres aux États Membres concernant les langues vivantes » (1982). Annexe A de Girard et Trim (1988).

[5] A la différence de Tagliante (2005), par exemple, Bourguignon rejette le terme « approche actionnelle » en indiquant que ce dernier n'est pas évoqué dans le CECRL. Bourguignon indique à juste titre que le Cadre européen commun de référence parle de « perspective actionnelle » qui est virtuelle par définition et donc, nécessite d'être transformée en une démarche avant de pouvoir l'appliquer aux pratiques pédagogiques et évaluatives (Bourguignon 2010 : 33).

[6] La Division des Politiques linguistiques est chargée de la coopération internationale dans l'enseignement des langues depuis 1957 entre les états membres du Conseil de l'Europe. Les projets menés par cette division concernent la politique linguistique éducative, notamment le plurilinguisme, l'élaboration des standards de référence européens communs ainsi que les droits et les devoirs dans l'enseignement des langues (www.coe.int/lang/fr)

[7] http://www.coe.int/t/dg4/linguistic/dnr_fr.asp

[8]

http://www.coe.int/t/dg4/education/elp/elpreg/Source/Publications/ELP_StorySoFar_July2011_Final_FR.pdf

[9] Cette association, mieux connue sous le nom anglais « Association of Language Testers in Europe », est une association d'organismes chargés du développement des tests en langues et des certifications des compétences langagières. Elle englobe actuellement 34 membres, responsables de la conception des tests en langues et d'évaluation des compétences en 23 langues. Les objectifs de cette association consistent à promouvoir la société multilingue ainsi qu'à établir et à maintenir les standards d'évaluation langagière, www.alte.org .

[10] http://www.alte.org/attachments/files/alte_cando.pdf

[11] Les catégories dans la banque des descripteurs proviennent en partie des échelles de compétences consultées, et en partie des réflexions théoriques menées par le groupe chargé de la conception du CECRL au sein du Conseil de l'Europe (North 2000 : 182).

[12] Ces quatre secteurs sont le secteur d'enseignement secondaire inférieur et supérieur, la formation vocationnelle ainsi que la formation des adultes (North 2000 : 185).

[13] Ce sous-chapitre de la thèse s'appuie sur la nomenclature proposée par David Little, dans son analyse du schéma descriptif du CECRL (Little 2007 : 646). Il faut préciser que les catégories de l'activité langagière, à savoir, la réception, la production, l'interaction et la médiation, évoquées dans l'article de Little, sont désignées les activités de communication langagière dans le CECRL. Les termes « catégorie de l'activité langagière » et « activité de communication langagière » seront utilisés de manière interchangeable par la suite.

[14] Les cinq activités de communication langagière, contenues dans la grille pour l'auto-évaluation, sont la compréhension de l'oral, la compréhension de l'écrit, la production orale en continu, l'interaction orale, et la production écrite (Conseil de l'Europe 2005 : 26-27).

[15] La désignation « niveaux avancés » ne doit pas être confondue avec le « niveau avancé », car ce dernier renvoie au niveau B2 uniquement (Conseil de l'Europe 2005 : 24).

[16] Ainsi, la répartition de la compétence langagière en niveaux plus étroits est favorisée en situation d'apprentissage, tandis qu'en situation d'évaluation, l'échelle composée des niveaux conventionnels larges est privilégiée. (Conseil de l'Europe 2005 : 30).

[17] La raison pour laquelle il peut être utile d'énumérer les tâches, inférieures au niveau A1, est qu'elles peuvent servir d'objectifs pour les débutants (Conseil de l'Europe 2005 : 30).

[18] Il existe plusieurs autres échelles de compétences, à côté de celles, incluses dans le CECRL, par exemple, *Echelle de compétence langagière des Eurocentres* (1993).

[19] Le modèle de Rasch est un des modèles faisant partie de la théorie de probabilité. La fonction de ce modèle consiste à déterminer la difficulté d'un item isolé dans une banque d'items, mais il peut également être utilisé pour calculer la difficulté des descripteurs (Conseil de l'Europe 2005 : 151).

[20] Ce scepticisme par rapport à la linéarité d'acquisition de la compétence langagière est partagé par les adeptes des théories émergentiste et socioconstructiviste. Cependant, il s'explique par les raisons différentes. North & Schneider (1998) et Hulstijn (2007) dénoncent l'absence des données empiriques pour valider l'assertion de la linéarité d'acquisition des compétences langagières, qui est l'idée centrale dans le CECRL. En revanche, les adeptes des deux théories évoquées, étant adhérents à l'idée de « la non-linéarité des apprentissages », rejettent complètement cette idée centrale du CECRL en raison de leur positionnement théorique (Narcy-Combes 2007 : 9).

[21] Les activités de communication langagière, la production, la réception, d'interaction et la médiation, sont appelées les compétences fonctionnelles dans Hulstijn (2007 : 8) parce qu'elles servent à réaliser les tâches de communication (Conseil de l'Europe 2005 : 48). Selon l'approche actionnelle du CECRL, l'accomplissement des tâches de communication « passe par les activités langagières » (Conseil de l'Europe 2005 : 19).

[22] http://cache.media.eduscol.education.fr/file/LV/72/1/Programme_anglais_palier1_123721.pdf

[23] Le CECRL distingue quatre formes d'activité communicative: la production, la réception, l'interaction et la médiation (CECRL 2005 : 48). Chacune de ces formes peut se manifester à l'oral et à l'écrit, mais ces manifestations restent au sein de la même forme d'activité communicative.

[24] Ces spécifications ont été décrites dans les Instructions Officielles (IO), élaborées par le Ministère de l'Education Nationale et composées de deux paliers. Le palier 1, édité en juin 2006, a été applicable à la rentrée 2006 (CNDP 2006 : 1). Le palier 2, publié deux ans plus tard, est entré en vigueur à la rentrée 2008 (CNDP 2008 : 1). Les Instructions Officielles seront décrites dans le chapitre « Conception du nouveau test » de cette thèse.

[25] Malgré une très grande importance accordée à l'apprentissage des langues et au plurilinguisme dans le CECRL, ceux-ci ne sont pas considérés comme un but en soi, mais comme un moyen d'atteindre les objectifs sociaux, culturels et politiques, à savoir : « une plus grande mobilité, une communication internationale plus efficace qui respecte les identités et la diversité culturelle, un meilleur accès à l'information, une multiplication des échanges interpersonnels, l'amélioration des relations au travail et de la compréhension mutuelle » (Conseil de l'Europe 2005 : 11).

[26] Les classes européennes sont celles où une matière est enseignée en langue étrangère (Byrnes 2007 : 644).

[27] Les différences multiples impliquent la socialisation culturelle et linguistique, mais aussi leur plurilinguisme actuel (Byrnes 2007 : 668).

[28] http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/109FR.pdf

[29] Cet ouvrage sera évoqué dans le chapitre suivant, dans la sous-partie consacrée à la validation des items

[30] Cela s'explique par le fait que les connaissances lexicales et les connaissances grammaticales forment deux systèmes différents dans la mémoire (Westhoff 2007 : 677). La production du discours demande l'activation de ces deux systèmes des connaissances, mais aux bas niveaux de compétences, l'usage simultané des systèmes lexical et grammatical implique leur surcharge. En conséquence, le système de production du discours risque de collapser (Westhoff 2007 : 677).

[31] Les meilleurs descripteurs parmi ceux qui ont été approuvés ont été mis sur le site : http://www.coe.int/portfolio

[32] Il faut noter que les portfolios européens de langues ne s'adressent pas tous au public scolarisé. Il en existe plusieurs modèles dans chaque pays de l'Union Européenne qui couvrent toutes les tranches d'âge (Tagliante 2005 : 77).

[33] Le Portfolio européen des langues est présenté en détail et peut être consulté sur le site du Conseil de l'Europe : http://www.coe.int/t/dg4/education/elp/default_fr.asp

[34] La Résolution sur le Portfolio européen des langues (adoptée lors de la 20e session de la Conférence permanente des Ministres de l'Education du Conseil de l'Europe, Cracovie, Pologne, 15-17 octobre 2000) est accessible sur le site suivant : http://www.oapee.es/dctm/weboapee/iniciativas/portfolio/epel/resolutionelp2000fr.pdf?documentId=0901e72b8189cfc4

[35] Les organismes qui ont développé les portfolios européens de langues sont les ministères de l'éducation, les autorités régionales éducatives, les organisations internationales non gouvernementales du secteur de l'éducation, les projets européens, les réseaux d'établissements scolaires privés http://www.coe.int/t/dg4/education/elp/elpreg/Accredited_models/Accredited_ELP_2010_FR.asp

[36]

http://www.coe.int/t/dg4/education/elp/elpreg/Accredited_models/Accredited_ELP_2010_FR.asp

[37] La qualité des différents modèles élaborés varie fortement. Ainsi, les listes des descripteurs utilisées ne sont pas toujours adéquates. Le caractère inadéquat de certains modèles concerne même les listes des descripteurs utilisées pour l'auto-évaluation malgré le fait que le Conseil de l'Europe présente un grand nombre des descripteurs sur son site internet (Little 2007 : 651).

[38] Malgré un grand nombre des Portfolios européens des langues accrédités, il est rare de trouver un modèle qui est librement accessible. Dans la plupart des cas, les concepteurs fournissent une description du PEL qui peut être détaillée, mais n'autorisent pas d'accéder au modèle lui-même. C'est, par exemple, le cas pour le Portfolio développé par l'Université Montesquieu-Bordeaux IV qui a été accrédité par le Conseil de l'Europe en 2010.

Le portfolio choisi aux fins de démonstration est accessible gratuitement sur le site suivant:

http://www.crdp.ac-caen.fr/didier/portfolio/outils.htm

[39] *La Banque de descripteurs pour l'auto-évaluation créée pour le Portfolio européen des langues* a été conçue par la Division des Politiques linguistiques du Conseil de l'Europe dans le but de

mettre à la disposition des concepteurs des Portfolios européens des langues (PEL) des descripteurs constitutifs des PEL déjà existants. Puisque ces descripteurs sont déjà validés, leur usage est censé rendre l'élaboration des nouveaux PEL plus rapide et plus simple ainsi que faciliter la procédure de validation des nouveaux modèles. Ces descripteurs peuvent être utilisés tels quels lors du développement des nouveaux PEL. Les descripteurs inclus dans la Banque de descripteurs doivent satisfaire à deux critères. Premièrement, ils doivent être adossés ou susceptibles d'être adossés aux descripteurs et/ou aux catégories et aux niveaux communs de référence du CECRL. Deuxièmement, ils doivent être appropriés pour l'auto-évaluation, du point de vue de compréhensibilité de contenu et de formulation (Schneider& Lenz 2004 : 4).

[40] http://www.crdp.ac-caen.fr/didier/portfolio/visite.htm

[41] http://www.crdp.ac-caen.fr/didier/portfolio/pdfs/PortfolioEN-listening.pdf

[42] http://www.crdp.ac-caen.fr/didier/portfolio/visite.htm

[43] http://www.crdp.ac-caen.fr/didier/portfolio/visite.htm

[44] http://www.crdp.ac-caen.fr/didier/portfolio/pdfs/PortfolioEN-listening.pdf

[45] http://www.crdp.ac-caen.fr/didier/portfolio/visite.htm

[46] http://www.crdp.ac-caen.fr/didier/portfolio/visite.htm

[47] http://www.crdp.ac-caen.fr/didier/portfolio/visite.htm

[48] http://www.crdp.ac-caen.fr/didier/portfolio/visite.htm

[49] http://www.crdp.ac-caen.fr/didier/portfolio/pdfs/PortfolioEN-listening.pdf

[50] http://www.lancaster.ac.uk/fss/projects/grid/

[51] Les caractéristiques de compréhension sont exprimées par un nombre des verbes dans le CECRL, par exemple, comprendre, reconnaître, localiser, inférer etc. De telles caractéristiques sont désignées les « opérations » (Alderson 2004 : 6).

[52] Afin de résoudre le problème de circularité¸ un nombre des procédures successives a été proposé. Ces procédures sont évoquées dans la sous-partie consacrée à la validation des items du chapitre 3 : « Conception d'un nouveau test ». La première consiste à décrire les tâches et les items en utilisant les instruments de classifications élaborées, à savoir, cadres et les grilles d'analyse. A la suite de la description, il convient d'estimer le niveau d'un item à la base de ces instruments et des échelles de compétences contenues dans le CECRL. La troisième étape est de piloter les items en décrivant en détail les caractéristiques de l'échantillon des candidats. Ensuite, le calibrage des items à été suggéré qui implique le classement des items selon les différents niveaux de compétences. La procédure suivante est la standardisation afin de délimiter les différents niveaux de compétence (Alderson 2004 : 13). La dernière étape consiste à attribuer le niveau définitif aux items. Il faut noter que l'attribution de celui-ci est possible seulement en cas de coïncidence entre le niveau déterminé à la base des données empiriques et le niveau estimé à la base du contenu analysé. Pour le dire autrement, l'attribution du niveau définitif demande la coïncidence entre le niveau psychométrique et le niveau estimé d'un item. Les procédés évoqués seront expliqués en détails et appliqués au test de positionnement élaboré dans le cadre du projet de recherche à la base de cette thèse.

[53] Les tâches se composent des items et d'un ou des plusieurs textes (Alderson 2004 : 6).

[54] Comme évoqué ci-dessous, les opérations sont exprimées par les verbes qui caractérisent la nature de compréhension, comme « comprendre », « inférer », « reconnaitre »

[55] La grille est accessible sur le site suivant: http://www.lancaster.ac.uk/fss/projects/grid/

[56] http://www.lancaster.ac.uk/fss/projects/grid/samplegrid/samplegrid.php

[57] La nécessité de ne pas se limiter à l'estimation du niveau des items à l'aide de cette grille afin de relier le test au CECRL, mais de recueillir les données empiriques est soulignée par les concepteurs de cette grille, sur le site d'accueil : http://www.lancaster.ac.uk/fss/projects/grid/

[58] http://www.lancaster.ac.uk/fss/projects/grid/

La procédure qui consiste à établir les normes (ou standards) sera décrite dans le chapitre consacré à la conception du test POSILANG et appliqué aux items de ce test.

[59] Les deux composantes faisant partie du dispositif sont accessibles sur le site http://www.lancaster.ac.uk/fss/projects/grid/grid.php**.**

[60] http://www.lancaster.ac.uk/fss/projects/grid/samplegrid/samplegrid.php

[61] http://www.lancaster.ac.uk/fss/projects/grid/samplegrid/samplegrid.php

[62] L'information sur le nombre des textes ne correspond pas aux données fournies sur la deuxième page qui rend clair que cette tâche contient deux textes.

[63] http://www.lancaster.ac.uk/fss/projects/grid/grid.php.

[64]http://www.lancaster.ac.uk/fss/projects/grid/samplegrid/textpage.php?&mp=1&task_id=32&text_id=14

[65] http://www.lancaster.ac.uk/fss/projects/grid/samplegrid/taskpage.php?&task_id=32.

[66] Comme il sera expliqué dans le chapitre suivant, l'index de discrimination indique le pouvoir de discrimination d'un item, qui est sa capacité de distinguer entre les candidats aux différents niveaux de compétence (Alderson, Clapham & Wall 1995 : 81).

[67] http://www.lancaster.ac.uk/fss/projects/grid/training.php?

[68] http://www.lancaster.ac.uk/fss/projects/grid/training/tests/reading/dialang-english-reading.pdf.

[69] http://www.lancaster.ac.uk/fss/projects/grid/reviewtext.php?

[70] http://www.lancaster.ac.uk/fss/projects/grid/trainingtext.php?&task_id=25&text_id=18

[71] http://www.lancaster.ac.uk/fss/projects/grid/trainingitem.php?&task_id=25&item_id=17

[72] http://www.lancaster.ac.uk/fss/projects/grid/tle.php?&task_id=25

[73] L'estimation divergente du niveau nécessaire à la compréhension du texte dans cette grille par rapport à la grille d'analyse du texte n'est pas prise en compte car cette incongruïté a évidemment des raisons techniques.

[74] http://www.dipf.de/en/research/projects/european-bank-of-anchor-items-for-foreign-language-skills

[75] http://www.dipf.de/en/research/projects/european-bank-of-anchor-items-for-foreign-language-skills

[76] http://www.dipf.de/en/research/projects/european-bank-of-anchor-items-for-foreign-language-skills

[77] Le mot « banques » est utilisé délibérément au pluriel, car six banques ont été conçues dont chacune contient les items rédigés en une des trois langues évoquées et évaluant une de deux capacités réceptives.

[78] http://www.dipf.de/en/research/projects/european-bank-of-anchor-items-for-foreign-language-skills

[79] http://www.dipf.de/en/research/projects/european-bank-of-anchor-items-for-foreign-language-skills

[80] Ces étapes seront présentées et expliquées dans le chapitre suivant, « Conception d'un nouveau test », dans les sous-parties consacrées aux différentes étapes de la mise en œuvre d'un test.

[81] Le format adaptatif sera expliqué en détail au début du troisième chapitre. Ce format consiste à soumettre la version du test adaptée au candidat à son niveau individuel de compétences.

[82] La théorie IRT, l'acronyme pour « Item Response Theory », permet de calibrer le niveau de difficulté de l'item indépendamment de la compétence des candidats qui y ont répondu. En revanche, en cas de la calibration non fondée sur cette théorie, le niveau de difficulté est influencé également par les compétences des participants au test (Alderson&Huhta 2005 : 313).

[83] La discussion du test diagnostique qu'Alderson entreprend dans son ouvrage, *Diagnosing foreign language proficiency. The interface between language and assessment (2005),* met en évidence son point de vue que la fonction de ce type de test n'est pas de donner le score aux candidats, faute d'utilité de le faire (p. 12).

[84] http://www.lancaster.ac.uk/researchenterprise/dialang/dialang_reasons.htm

[85] Hahr & Hansen (2006 : 4) proposent d'attacher les valeurs externes aux échelles de réponse, ainsi que de formuler les items de telle façon que les effets de désirabilité sociale ne se produisent pas.

[86] http://www.lancaster.ac.uk/researchenterprise/dialang/dialang_reasons.htm.

[87] http://www.lancaster.ac.uk/researchenterprise/dialang/dialang_reasons.htm

Annexe – Chapitre 2

Appendix1. Sample page from the DIALANG Assessment Framework

DIALANG Assessment Framework Appendix 1

SPECIFICATIONS RELATED TO COMMUNICATIVE TASKS AND PURPOSES IN THE COUNCIL OF EUROPE'S THRESHOLD AND VANTAGE LEVEL PUBLICATIONS

August 1997 / November 1999

This appendix relates to the **section of communicative functions and purposes** (chapter 4.2,) in the Council of Europe Common Framework, draft 2, 1996. It lists the corresponding sections in the Threshold and Vantage documents. For ease of comparison, the same categories are presented side by side. The main differences between the specifications for the two levels are highlighted by using **bolding**. (N.B. Waystage does not specify communicative tasks and purposes in detail so this document includes the **specifications found in the Threshold and Vantage level documents only.)**

THRESHOLD LEVEL (1990)	**VANTAGE LEVEL (1996)**
1. Practical transactions	1. Practical transactions.
Learners will be able to cope with transactional situations of everyday life requiring **a largely predictable use of language.**	Learners will be able to cope with transactional situations in everyday life. At Vantage Level, learners will be able to deal more flexibly with **these situations** than at Threshold Level when they are **problematic or take an unexpected turn.** With enriched language resources (especially a wider vocabulary), learners will be able to express their needs and intentions **more precisely**, with less (though still some) need for compensatory strategies.
1.1 Contacts with officials	1.1. Contacts with officials
In all contacts with officials learners would be able to ask for repetition, clarification and explanation etc. of any information, questions or documents	N.B. In all contacts with officials learners would be able to ask for repetition, clarification and explanation etc. of any information, questions or

THRESHOLD LEVEL (1990)	VANTAGE LEVEL (1996)
not understood, and should be able to ask for the services of an interpreter and/or legal adviser in case of serious difficulty (chapter 12).	documents not understood, and should be able to ask for the services of an interpreter and/or legal adviser in case of serious difficulty.

3 De la conception à la validation d'un nouveau test de positionnement

Le moment est venu d'aborder le processus de conception et d'évaluation d'un nouveau test de positionnement en langue anglaise, nommé par acronymie POSILANG. Cependant, avant de présenter les différentes étapes de réalisation de ce dispositif particulier, il nous paraît important d'en préciser les objectifs et le cadre institutionnel. Nous décrivons ensuite le matériel et la méthode d'élaboration adoptés.

3.1 Objectifs

POSILANG a été conçu dans le cadre du projet « Didactique des langues : ressources numériques et hybridations » (2011-2014) financé par la Région Aquitaine et porté par l'Université Bordeaux Montaigne. L'objectif principal de ce projet était de concevoir un test de positionnement entièrement automatisé pour les établissements publics d'Enseignement Supérieur du Pôle de Recherche et d'Enseignement Supérieur (PRES) de Bordeaux et de la future *Maison des Langues et des Cultures*. Il était prévu d'utiliser le POSILANG pour évaluer les compétences des étudiants en langues vivantes étrangères, idéalement en anglais mais aussi en allemand, espagnol et italien. Ce test de positionnement, développé sur fonds publics, serait libre de droits, standardisé et adossé au *Cadre européen commun de référence pour les langues* (CECRL).[1] Des tensions entre les établissements partenaires, sans rapport avec le projet mais aux effets délétères, bloquèrent les financements à mi-parcours, figèrent la dynamique impulsée par la réalisation d'une *Maison des Langues* (aujourd'hui suspendue), si bien que la numérisation du dispositif et l'expérimentation simultanée sur plusieurs sites aquitains (Bordeaux, Bayonne, Pau) ne purent être menées à bien. Le projet put néanmoins être poursuivi, de façon volontariste, pour satisfaire les besoins de la présente étude doctorale et avec l'espoir d'une valorisation future sous l'égide de la nouvelle COMUE. Ainsi purent être amorcées les études préliminaires, la conception d'une maquette, la mise au point d'une version pilote, l'expérimentation auprès de deux groupes distincts d'étudiants (des primo-

entrants en L1 d'anglais et des étudiants en fin de parcours, inscrits en M2 sciences du langage).

Il est important de rappeler qu'un dispositif d'évaluation standardisé se distingue par son adhésion à certains niveaux de performance définis et maintenus à travers ses différentes phases. Parce qu'ils sont constants et s'appliquent aussi bien à des tests distincts qu'aux différentes versions d'un même test, les niveaux de performance ont le statut d'objectifs standardisés (Brown 2010 : 86). On entend par là un ensemble de compétences bien définies, fondées sur les objectifs d'un cours ou sur un programme d'apprentissage, recouvrant une ou plusieurs années (Brown 2010 : 86). Une évaluation qui repose sur des normes standardisées respecte ainsi des procédures qui ont été spécifiquement conçues dans le but d'évaluer ces compétences (Brown 2010 : 86). Dans le cas présent (POSILANG), il s'agit de concevoir un test adossé aux objectifs standardisés du *Cadre Européen Commun de Référence* pour les langues (CECRL). Les niveaux de performance qui structurent notre dispositif d'évaluation sont donc reliés à des échelles normées de compétences énoncées dans ce référentiel. Cet adossement est d'autant plus fondé que le CECRL a été conçu au départ pour être un référentiel d'évaluation. Tenir compte des critères du CECRL, niveau par niveau, est non seulement nécessaire (Noël-Jothy et Sampsonis (2006 : 4347) mais incontournable.

Dans POSILANG, une correspondance doit donc être établie entre les scores obtenus au test et les six niveaux de compétence distingués par le CECRL, comme c´est le cas pour tout test standardisé. Cependant, il faut aussi veiller à établir des procédures d'évaluation transparentes. Cela suppose non seulement qu'on explicite la relation entre le score obtenu et le niveau de compétences, mais aussi et surtout qu'on définisse les normes pour les différents niveaux de résultats exprimés par les scores (Conseil de l'Europe 2005 : 38). La définition des normes pour l'attribution des niveaux de résultats doit se traduire concrètement par la mise à disposition « des critères d'évaluation transparents » (Conseil de l'Europe 2005: 38). Les niveaux de résultats ne sont pas à confondre avec les niveaux de compétences, car les deux termes recouvrent deux notions différentes. Les niveaux de résultats sont des niveaux de performance chiffrés. Ces derniers servent de base à l'inférence de niveaux de compétences. La

définition des normes pour les différents niveaux de résultats doit obligatoirement tenir compte de l'objectif d'évaluation (Conseil de l'Europe 2005 : 38).

Deuxièmement, le test standardisé élaboré dans le cadre de ce projet de recherche doit satisfaire aux qualités de conception présentées dans le premier chapitre. On se souvient que le haut degré de praticité et de fiabilité constitue un atout typique des tests standardisés. Nous visons cette qualité, tout en réalisant un test qui respecte aussi le principe d'authenticité**.** Pour y parvenir, les items doivent être aussi proches que possible des usages réels et situés de la langue, par leur forme, par leur fonction communicative et enfin par le contexte dans lequel on les insère. Comme en art, mais en dehors de tout projet esthétique, le naturalisme auquel on se livre dans un test relève du pur artifice. Il doit être entendu que nous ne nous sommes pas rendus sur place, en territoire anglophone, pour opérer une sorte de copier-coller de la réalité interactionnelle observée. Nous nous sommes cependant inspirés de productions textuelles et visuelles authentiques, que nous avons croisées avec celles recensées dans des supports pédagogiques, pour glaner des énoncés ou abstraire des structures, que nous avons ensuite retravaillées.

Les caractéristiques d'un test standardisé, notamment la présence de normes bien définies, ainsi que l'attribution de scores précis et transparents, présentent plusieurs avantages dans une situation d'évaluation comme la nôtre. A ces avantages s'ajoute celui de la numérisation (à finaliser lorsqu'un financement sera rétabli) permettant une passation en ligne. Les atouts essentiels d'une telle configuration, dans un contexte de forte massification, sont la facilité d'administration aux candidats, la rapidité d'attribution des scores et du niveau de compétence associé. [2] En raison du nombre élevé de primo-entrants, notamment aux niveaux B1 et B2, le test doit pouvoir être exécuté en moins d'une heure lors des journées d'intégration, en amont des inscriptions pédagogiques.

Or, actuellement il n'existe aucun test qui respecte ces impératifs. DIALANG, qui a pour atouts la gratuité, l'intelligence de conception et l'adossement au CECRL, s'avère néanmoins d'accès aléatoire (saturation du serveur, incompatibilités avec certains systèmes d'exploitation, mises à jour

insuffisantes). En outre, le temps de passation est supérieur à 1h. ELAO, OXFORD et VOCABLE sont, quant à eux, des tests payants: 2 à 5 € par passage selon les volumes et les tarifs négociés. Par conséquent, aucun test de positionnement conçu dans les normes n'est utilisé par l'Université Bordeaux Montaigne à l'heure actuelle, alors que l'impératif du positionnement n'a jamais été aussi élevé pour l'admission dans les filières langues mais aussi pour la répartition dans les UE Langue obligatoires de tous les cursus. Pour cette raison, il est prévu que le nouveau dispositif comble ce manque. Bien que la priorité aille à l'anglais, d'autres langues sont visées : l'espagnol, en particulier, mais aussi l'allemand et l'italien. La matrice créée doit être compatible avec ces quatre langues.[3] Au-delà de l'absence d'un test de positionnement gratuit, accessible et respectant les bonnes pratiques, il y a une deuxième raison à notre décision d'élaborer un dispositif sur site. Nous avons vu que pour avoir une bonne qualité, un test gagnait à être élaboré localement, en tenant compte de la situation d'évaluation particulière dictée par l'environnement (Hughes 2003 : 17). C'est le seul procédé qui permette de concevoir des tests de positionnement adaptés à un placement précis des candidats (Hughes 2003 : 17).

3.1.1 Objectif 1: choix du format

En ce qui concerne le format, il est envisagé de développer un seul grand test autocorrectif de mesure de la compétence langagière, avec franchissement graduel d'étapes, dans chaque compétence. Les scores intermédiaires indiquent à l'étudiant s'il peut accéder ou non à l'étape suivante. Ce procédé coïncide avec le mode de fonctionnement d'un test adaptatif. Les tests adaptatifs constituent « la deuxième génération des environnements utilisant l'ordinateur » (Laurier 1998 : 247).

La structure du dispositif adopté doit permettre à la fois d'obtenir des scores globaux et des scores par domaine de compétence, d'attribuer un niveau de compétences à partir de ces scores, et enfin de prendre en compte des données susceptibles d'affecter le niveau d'un candidat. Ce sont, par exemple, le nombre d'années d'étude scolaire, la note obtenue au baccalauréat et les résultats de l´autoévaluation des compétences en compréhension orale et écrite ainsi qu'en production écrite.[4] Les tests qui tiennent compte du parcours

antérieur du candidat et du regard qu'il porte sur sa propre pratique de la langue facilitent l'obtention des informations recherchées sur son niveau (Laurier 1998 : 247). Notre préférence va donc à ce format, même s'il n'est pas le seul envisageable.

3.1.1.1 Tests informatisés

L'administration d'un test traditionnel suppose qu'on regroupe les participants en un même lieu. L'attribution de scores nécessite par ailleurs de soumettre les réponses à un correcteur (Laurier 1998 : 247). L'avantage principal des tests informatisés, aussi appelés tests automatisés, est l'automatisation de la passation et de la correction. Les fonctions d'administration et de correction sont confiées à l'ordinateur qui, de ce fait, accomplit le travail normalement dévolu à des intervenants humains (Laurier 1998 : 247). Ce trait distinctif des tests automatisés représente un atout considérable, notamment à « grande échelle », quand un grand nombre de candidats est concerné (Laurier 1998 : 247).

Les atouts d'un test informatisé sont nombreux. On note en premier lieu, le calcul rapide et fiable d'un score, pouvant inclure des opérations arithmétiques complexes qui pourraient sinon être sujettes à erreurs. Ensuite, une fiabilité accrue lors des étapes associées à l'attribution des scores : la saisie de la réponse, l'attribution d'un score, la transformation de ces derniers en niveaux de compétence et enfin la production de listes de résultats (Laurier 1998 : 247). Ces étapes sont souvent oubliées alors qu'elles sont tout aussi importantes. La possibilité d'une administration individualisée offerte par les tests automatisés constitue également un avantage objectif, puisqu'il n'est pas indispensable de rassembler les candidats en un même lieu (Laurier 1998 : 247). Malgré ces atouts manifestes, la conception d'un test informatisé n'est pas si différente de celle d'un test traditionnel. Dans une majorité de cas, en effet, on assiste à la simple transposition numérique de tests existant déjà au format papier. L'administration linéaire d'un ensemble de tâches est ainsi préservé (Laurier 1998 : 247).

Les tests automatisés ne sont cependant pas exempts de défauts. Un inconvénient majeur est leur caractère inadapté pour évaluer la composante communicative de la compétence langagière (Laurier 1998 : 247). Les tâches destinées à cet usage, comme la rédaction de dialogues ou de paragraphes, sont

exclues, en raison de la difficulté à traiter automatiquement les productions libres. Ce verrouillage de la parole constitue une différence majeure entre les tests automatisés et les dispositifs manuels traditionnels (Laurier 1998 : 247). Dans ces derniers, les réponses ouvertes sont non seulement possibles mais courantes. Malgré cette carence, la part des tests informatisés n'a cessé de croître sur le marché (Laurier 1998 : 247). Il en existe de plusieurs types et formats. Pour le positionnement, on note le succès d'*Oxford Online Placement Test* et *Vocable*. Pour la certification, *TOEFL* et *BULATS*, entre autres. Il faut préciser que les tests cités existent dans plusieurs formats, ce qui prouve que l'opposition support papier / support numérique ne recouvre pas une différence fondamentale de conception. Le *TOEFL* est ainsi disponible sur support papier, même si son usage est restreint par rapport au format automatisé. En ce qui concerne les tests BULATS, ils existent en trois formats : en ligne, sur CD ROM et sur papier. [5]

3.1.1.2 Tests adaptatifs

Les tests adaptatifs constituent la deuxième génération de tests informatisés (Laurier 1998 : 248). La différence principale de ces tests par rapport à la première génération tient au fait que l'administration n'est plus linéaire. Cela signifie que les tâches dans les tests adaptatifs sont choisies au cours de la passation individuelle en fonction de la performance effective d'un candidat (Laurier 1998 : 248). L'axe de progression cesse d'être rigide et prévisible. Le test s'ajuste au niveau manifeste de chacun et les candidats se retrouvent ainsi évalués de façon progressive et individualisée. La procédure consiste à extraire en cours de passation des informations permettant d'affiner l'estimation du niveau général, à partir des réponses données (Laurier 1998 : 248). A mesure qu'on avance dans le test, l'estimation du niveau s'affine. Cette estimation, en continuelle évolution, permet non seulement de prendre la mesure de ce que le candidat est capable d'accomplir mais aussi de proposer des tâches pertinentes pour cerner son niveau, à mesure qu'il avance (Laurier 1998 : 248). Il est important de noter que les tests automatisés de première génération, qui permettent de sélectionner une version en fonction du niveau présumé ne sont pas à proprement parler adaptatifs. En effet, l'estimation initiale du niveau et la proposition des tâches qui en découlent restent figées. Les réponses fournies

n'affectent en aucune manière les tâches proposées. Le test ne s'adaptant pas aux performances du candidat, il ne peut être qualifié d'adaptatif.

Les tests adaptatifs fournissent non seulement une estimation précise du niveau d'un candidat mais apportent également plusieurs autres avantages d'ordre psychométrique, psychologique et administratif (Laurier 2006 : 9-11). Les atouts psychométriques sont l'amélioration de la fiabilité et de la validité, grâce à l'adaptation précise des tâches au profil du candidat, mais aussi la découverte de constellations inhabituelles dans les réponses (Laurier 2006 : 9). Les avantages administratifs de ces tests sont également multiples. Certains coïncident avec les tests automatisés de première génération, d'autres sont caractéristiques de ce format (Laurier 2006 : 9). Les principaux avantages de ce type de test sont une durée de passation très réduite, rendue possible par l'estimation rapide du niveau d'un candidat, à l'issue de quelques tâches probantes et non redondantes, ainsi que l'administration d'items plus riches et variés correspondant aux compétences manifestes du candidat (Laurier 2006 : 9). Selon certains chercheurs, les tests adaptatifs auraient une durée de quatre à dix fois moindre que celle des tests classiques (Kingsbury & Weiss 1980 : 11).

Cet avantage pratique se double d'avantages psychologiques non négligeables, notamment la diminution de la frustration et de l'anxiété (Laurier 1998 : 248). En effet, la passation est rendue moins menaçante car les tâches qui ne correspondent visiblement pas au niveau d'un candidat se retrouvent écartées (Laurier 1998 : 248). La composante psychologique est au moins aussi importante que la composante pratique, étant donné que beaucoup de candidats sont sujets à des états émotionnels négatifs dans les situations d'évaluation, qui peuvent affecter leurs résultats de façon indésirable. Les tests adaptatifs offrent par ailleurs les mêmes fonctionnalités que les dispositifs traditionnels ou les tests automatisés de première génération. Opter pour un format adaptatif n'exige aucun renoncement. Au contraire, l'évaluation individualisée permet une lecture plus fine des performances, facilitant la comparaison avec celles d'autres apprenants ou la mise en regard avec des attendus.[6] Pour toutes ces raisons, les tests adaptatifs sont préférables aux tests traditionnels lorsqu'on est engagé dans une démarche de positionnement (Laurier 1998 : 250).

Le développement de tests adaptatifs a été rendu possible grâce aux résultats de la recherche dans le domaine de la mesure, en particulier, grâce à la

théorie de réponses aux items (Baker 1992). Dans le cadre de cette théorie, il devient plus facile de construire des banques d'items, constituées de tâches de difficulté variée mais qui mesurent la même habileté. Grâce aux banques d'items, il est possible de situer les candidats sur une échelle d'habileté, en leur faisant accomplir des tâches adaptées à leur niveau (Laurier 1998 : 248). [7] Il existe plusieurs tests de ce format sur le marché. Un test de positionnement adaptatif est par exemple, *eLAO.*[8] Un test de certification adaptatif très connu, reconnu par les institutions d'enseignement supérieur et dans le domaine professionnel, est *IELTS*.[9]

3.1.2 Objectif 2: évaluation des compétences dans le test

Le test en cours d'élaboration a pour but de fournir le score numérique global d'un candidat et donc son niveau de compétence langagière général. Au-delà de cette estimation, POSILANG entend fournir une évaluation qualitative des résultats par domaine de compétence, en compréhension orale et écrite, ainsi qu'en expression écrite. Pour parvenir à cette évaluation, il faut procéder à l'annotation de tous les items (*tagging*) par catégorie et sous-catégorie**.** Il est envisagé que les candidats puissent repérer leurs performances et leurs déficiences, par exemple, au niveau de la compréhension orale ou écrite, ainsi que leurs insuffisances grammaticales ou lexicales. [10]

Il convient de rappeler qu'un test de compétences exige une définition exhaustive des compétences spécifiques entrant dans la compétence générale (Bachman 1990 : 8). Puisque nous entendons tester la compréhension et la production écrite guidée ainsi que la compréhension orale dans notre test de langue, il faudra s'astreindre à définir ces compétences de manière exhaustive.[11] Ce procédé permet d'obtenir une définition opérationnelle de la notion de compétence langagière soumise à évaluation (Brown 2010 : 118). Au sein de chacune des compétences de communication, il est envisagé d'évaluer les compétences lexicale et grammaticale. En ce qui concerne la compétence lexicale, celle-ci recouvre l'étendue du vocabulaire et la maitrise de ce dernier (Conseil de l'Europe 2005 : 88-89). L'étendue du vocabulaire, la connaissance de certaines tournures idiomatiques seront évaluées à partir du niveau B1, tout en gardant à l'esprit qu'une bonne maîtrise des idiomes n'est véritablement attendue

qu'à partir du niveau C 1 (Conseil de l'Europe 2005 : 88). En ce qui concerne la compétence grammaticale, il est prévu d'évaluer la grammaire de la phrase ainsi que la grammaire du texte à partir du niveau B1. La connaissance de la grammaire du texte exige une compréhension plus fine, y compris, des sous-entendus et des métaphores.

3.1.3 Objectif 3: conception du test de positionnement POSILANG avec des modules complémentaires « filières » adaptables

Nous avons vu que le test devait permettre d'effectuer un positionnement par compétences, grâce à un système d'annotation et de validation des items.[13] POSILANG se veut ainsi une aide majeure à l'orientation et à la répartition des étudiants par groupes de niveaux. Cependant, les filières ont leurs spécificités pouvant exiger une évaluation plus ciblée. POSILANG doit donc être conçu de telle sorte qu'il puisse accueillir des modules thématiques complémentaires à destination des étudiants des filières littéraires, scientifiques, juridiques et économiques. Ces modules seront appelés POSI-EXT et codés par domaine, par exemple, POSI-EXT LEA. Une demande en ce sens a déjà été formulée par la direction des départements LEA et LLCE de l'Université Bordeaux Montaigne.

Au positionnement dans la langue de communication générale peut donc être adjointe une évaluation des compétences en langue et en méthodologie de spécialité.[14] Chaque module complémentaire doit être axé sur des compétences et des prérequis propres à la discipline. Les insuffisances lexicales et grammaticales, les difficultés en compréhension écrite ou orale couramment repérées par les enseignants dans leur domaine de spécialité sont répertoriées et intégrées aux items. Il a été convenu que les enseignants fournissent, dans un premier temps, un corpus de fautes ou de difficultés expressives courantes, à partir desquelles puissent être conçues les extensions complémentaires. [15] Ces dernières ne sauraient d'ailleurs se limiter à des pièges ou difficultés. Des tâches d'expression et de compréhension spécifiques au domaine concerné doivent aussi être incluses.

3.2 Définition de la compétence langagière

Avant de procéder à l'évaluation de « compétences », il est nécessaire de définir le terme lui-même et de situer celui-ci par rapport aux notions de « connaissance » et de « capacité », avec lesquelles on le confond trop souvent. *Le Cadre Européen Commun de Référence* définit les compétences de la façon suivante: « Les compétences sont l'ensemble des connaissances, des habiletés et des dispositions qui permettent d'agir » (Conseil de l'Europe 2005 : 22). Cette définition générale de la compétence révèle d'abord que les compétences sont des entités complexes. Ensuite, que celles-ci rendent l'action d'un individu possible.

Le passage suivant met en évidence le lien entre l'usage et l'apprentissage de la langue d'une part et le développement des compétences communicatives d'autre part :

> L'usage d'une langue, y compris son apprentissage, comprend les actions accomplies par les gens qui, comme individus et comme acteurs sociaux, développent un ensemble des **compétences générales** et, notamment une **compétence à communiquer langagièrement** (Conseil de l'Europe 2005 : 15).

Dans la perspective actionnelle, représentée par le CECRL, l'usage et l'apprentissage de la langue ont pour but le développement de compétences langagières communicatives. Ceci explique l'emploi du terme « approche par compétences » pour désigner l'objectif d'apprentissage visé par le CECRL (Bourguignon 2010 : 22).

Le Cadre met en garde contre le risque de confusion entre la compétence langagière et les activités langagières : « Les activités langagières impliquent l'exercice de la compétence à communiquer langagièrement, dans un domaine déterminé, pour traiter (recevoir et/ou produire) un ou des textes en vue de réaliser une tâche » (Conseil de l'Europe 2005 : 15). Puisque les activités langagières permettent de traiter les textes de manière productive ou réceptive, il est évident qu'il s'agit de la compréhension de l'écrit, de la compréhension de l'oral, de la production écrite et de la production orale (Bourguignon 2010 : 23).

Bien que le but de la compétence langagière et des activités langagières soit de « réaliser une tâche », nous ne sommes pas en présence de concepts identiques (Conseil de l'Europe 2005 : 15). La compétence en langue se manifeste par des activités langagières, nommées « skills », « savoir-faire » ou capacités. Toutefois il ne suffit pas de posséder ces capacités pour faire preuve de compétence langagière (Bourguignon 2010 : 23). Maîtriser un savoir-faire suppose qu'on sache mobiliser ce dernier de façon pertinente :

> Etre compétent suppose bien plus que la maîtrise des connaissances et savoir-faire de base (ce que l'on tend à appeler les « ressources ») : cela suppose de pouvoir les mobiliser de façon pertinente dans des situations-problèmes à résoudre ou dans des tâches complexes à effectuer » (De Ketele 2006 : 23).

La définition ci-dessus insiste sur la différence entre maîtriser « des connaissances et des savoir-faire » et être capable de les «mobiliser de façon pertinente ». La mobilisation des connaissances et des capacités consiste à utiliser ces dernières de manière adéquate (Bourguignon 2010 : 24). La différence entre les deux opérations évoquées souligne la nécessité d'appliquer des méthodes spécifiques pour les acquérir. Tandis que la maîtrise des connaissances et des capacités peut être atteinte par des exercices répétitifs, il est nécessaire d'avoir recours à des tâches complexes pour apprendre à mobiliser et à développer une compétence, y compris, une compétence langagière (Bourguignon 2010 : 24). Cette nécessité d'un recours à des tâches complexes entre dans la définition même de la compétence : « La caractéristique essentielle de la compétence, c'est qu'elle ne se « bachote » pas. Elle se développe à travers les tâches de plus en plus complexes » (Bourguigon 2010 : 24).

L'approche actionnelle repose donc sur le développement de compétences générales et langagières chez les individus, afin de rendre ces derniers capables de réaliser des « tâches complexes » (De Ketele 2006 : 22). La logique actionnelle est donc une « approche par compétence(s) » de l'apprentissage (Bourguigon 2010 : 24). Le besoin de développer des compétences n'implique pourtant pas que l'objectif d'apprendre à maîtriser des connaissances et des capacités soit abandonné (Bourguignon 2010 : 24). La maitrise des connaissances et des capacités reste tout aussi nécessaire qu'avant la mise en œuvre de l'approche par compétences, même si l'acquisition de

celles-ci cesse d'être le seul objectif de l'apprentissage. Au-delà de rendre l'apprenant capable de réaliser des tâches complexes, « l'approche par compétence(s) » entend le « rendre autonome en tant qu'utilisateur de la langue » (Bourguignon 2010 : 24) ». Or, il n'est pas possible d'atteindre ces deux buts sans maîtriser les connaissances et les savoir-faire qui représentent autant de ressources pour les apprenants. Par conséquent, les exercices répétitifs qui servent à fixer ces ressources dans la mémoire restent indispensables et doivent aussi avoir lieu dans un processus d'apprentissage focalisé sur le développement des compétences (Bourguignon 2010 : 24). Plus les ressources maitrisées sont nombreuses plus les tâches complexes nécessitant l'emploi intégré des connaissances et des capacités peuvent être réalisées (Bourguignon 2010 : 28).

3.2.1 Définition de la connaissance et de la capacité

Pour clarifier davantage le sens du concept de « compétence », il convient de distinguer celui-ci de la connaissance et de la capacité. La connaissance est relative au « code linguistique » (Bourguignon 2010 : 26). Elle relève des domaines du lexique, de la syntaxe et de la phonologie et englobe également des connaissances culturelles et littéraires liées à une thématique (Bourguignon 2010 : 26). Dans le CECRL, il est question des « compétences linguistiques » parce que les connaissances énumérées sont présentées en lien avec la « capacité à les utiliser » (Conseil de l'Europe 2005 : 87). Conformément à la définition de la compétence contenue dans le CECRL, les connaissances évoquées deviennent des compétences parce qu'elles « permettent d'agir » (Conseil de l'Europe 2005 : 22). Le Cadre européen commun indique que le terme « compétence linguistique » recouvre « la connaissance des ressources formelles » (Conseil de l'Europe 2005 : 87). Puisque les ressources formelles permettent de créer et de formuler les messages linguistiquement corrects et porteurs du sens, ils relèvent évidemment des domaines du code linguistique, évoqués ci-dessus (Conseil de l'Europe 2005 : 87).

Quant aux capacités, elles proviennent de la qualification « être capable de … » (Bourguignon 2010 : 26). Pour cette raison, les activités de production et de réception, appelées « activités de communication langagière » dans le

CECRL (2005 : 39), sont les capacités et non les compétences (Bourguignon 2010 : 26). Il y a deux raisons qui expliquent l'usage erroné des deux termes. Premièrement, les activités langagières apparaissent dans le CECRL en lien avec les niveaux de compétence et ainsi, sont à tort assimilées à des compétences. Deuxièmement, l'emploi inapproprié du terme « compétence » remonte à son usage dans l'approche communicative, dans laquelle les quatre activités langagières étaient désignées « compétences» (Bourguignon 2010 : 26). Il est important de souligner que les connaissances et les capacités ne sont pas absentes ou parfaitement maîtrisées par un apprenant. Elles peuvent être appropriées à des degrés différents par un individu (Bourguignon 2010 : 26).

3.3 Les modèles de la compétence langagière en langue seconde

Chaque test repose sur une conception plus ou moins consciente et formalisée du langage, des compétences langagières, des usages de la langue et de l'apprentissage. En l'absence de modèle explicite, ces conceptions relèvent de l'intuition (Alderson, Clapham &Wall 1999 : 13). Les spécialistes de l'évaluation en langues sont unanimes à reconnaître qu'il est nécessaire de fonder le développement et l'usage des tests sur un cadre théorique contenant une définition explicite de la compétence langagière (Bachman 1990 : 81). Les différentes habiletés incluses dans le modèle choisi doivent également être définies avec une grande précision et de manière exhaustive (Brown 2010 : 118). Ce besoin est dû au fait que la manière de définir le concept général de compétence langagière et les habiletés spécifiques qu'il renferme exerce un fort impact sur la conception d'un test et sur l'évaluation des différentes compétences des candidats (Brown 2010 : 109). Pour comprendre cet argument, on peut penser, à titre d'exemple, à un test de compréhension de l'écrit qu'il est impossible d'élaborer sans postuler, ne serait ce que brièvement, le concept de compréhension de l'écrit et les habiletés mesurées lors d'un test adéquat. Ces informations sont à fournir dans les spécifications d'un test (Alderson, Clapham & Wall 1999 : 14)

Cependant, ce n'est que récemment qu'on a reconnu la nécessité de relier explicitement tel ou tel cadre théorique à des méthodes et des technologies

destinées à mesurer la compétence langagière (Bachman & Clark 1987). En effet, le développement et l'usage des tests en langues exigent qu'on formule des définitions claires des compétences à évaluer et qu'on identifie les méthodes utilisées à cet effet (Bachman 1990 : 81). Ce besoin s'explique par le fait que le concept de compétence communicative ainsi que les méthodes appliquées sont deux facteurs majeurs dans l'interprétation et l'utilisation des scores (Bachman 1990 : 13).

3.3.1 Des exemples de modèles de communication langagière

Il existe plusieurs modèles de la compétence langagière en langues étrangères qui ont évolué à travers le temps, en parallèle avec les pratiques d'évaluation (Dervin & Suomela-Salmi 2007 : 21).[16] Jusqu'au début des années 1980, le concept de compétence langagière englobait la composante linguistique et le contexte (Dervin&Suomela-Salmi 2007 : 22). Dans les années 1980, ce concept évolue, car la composante discursive et la composante sociolinguistique sont intégrées dans la notion de compétence langagière (Dervin & Suomela-Salmi 2007 : 22). Dans ce nouveau concept de la compétence langagière, la connaissance de la langue et la capacité à utiliser celle-ci sont considérées comme indissociables (Dervin & Suomela-Salmi 2007 : 22). Cette conception a des répercussions sur les pratiques d'évaluation. L'objectif est désormais de tester la capacité à utiliser la langue dans des situations aussi proches de la réalité que possible (Dervin & Suomela-Salmi 2007 : 22). Les documents authentiques et les simulations entrent massivement dans les tests de langues (Dervin & Suomela-Salmi 2007 : 22).

Le premier modèle de la compétence langagière à avoir exercé une influence déterminante sur les tests fut celui proposé par Canale & Swain (1980). Les auteurs s'intéressaient explicitement aux pratiques d'enseignement et à l'évaluation des compétences en langue seconde (Canale & Swain 1980 : 1). Dans ce modèle, la compétence communicative est composée de trois composantes qui sont la compétence grammaticale, la compétence sociolinguistique et la compétence stratégique. La compétence grammaticale est définie comme la connaissance des règles de la grammaire, en l'occurrence de la morphologie, de la syntaxe, de la sémantique et de la phonologie. Elle inclut

également la connaissance du lexique. La compétence sociolinguistique englobe la connaissance des règles d'usage (Canale & Swain 1980 : 6). La compétence stratégique est la connaissance de la manière dont les difficultés de communication peuvent être surmontées (Canale & Swain 1980 : 6). Il faut noter que la compétence communicative est conçue comme une connaissance dans ce modèle (Canale & Swain 1980 : 30). Outre la notion de compétence communicative, ce modèle intègre le concept de performance communicative qui renvoie à l'usage de la langue dans des situations de communication réelles (Canale & Swain 1980 : 6). Selon ce modèle, un test doit contenir des tâches qui mesurent la connaissance et, de ce fait, évaluent la compétence communicative, et des tâches qui évaluent l'usage des connaissances dans la communication réelle.

Le modèle de Swain s'inscrit dans ce nouveau concept de la compétence langagière, apparu dans les années 1980. Ce modèle propose une perspective multidimensionnelle de la notion de compétence langagière, dans le but d'évaluer non seulement la connaissance langagière mais aussi la capacité d'usage de la langue (Swain 1990 : 403). Dans ce modèle, trois domaines langagiers sont distingués qui sont la grammaire, le discours et la compétence sociolinguistique. La grammaire se limite à la justesse grammaticale à l'intérieur des phrases (Swain 1990 : 403). Le discours renvoie à la cohésion et à la cohérence textuelles, tandis que la compétence sociolinguistique couvre la capacité à utiliser la langue de façon socialement juste. Ces trois domaines peuvent être évalués en sollicitant des réponses orales et écrites, soit par le biais d'items au format QCM, soit sous forme de composition écrite (Swain 1990 : 403). Il faut souligner que le modèle de Swain n'a pas été conçu pour formuler une définition exhaustive de la compétence langagière, mais pour fournir un cadre opérationnel permettant de construire des dispositifs d'évaluation des compétences (Brown 2010 : 118). Ce modèle propose une approche actionnelle de l'enseignement et de l'apprentissage de langue (Dervin & Suomela-Salmi 2007 : 22).

Le modèle théorique, proposé par Bachman & Palmer (1982a), a été empiriquement validé, à la différence des autres modèles de la compétence communicative. La validation empirique est importante, compte tenu de l'influence des modèles sur l'élaboration des programmes, sur la conception des

tests langagiers et plus généralement sur tout l'ensemble du domaine d'évaluation langagière (Bachman & Palmer 1982 a : 463). La recherche empirique rapportée sert à analyser le concept de compétence communicative. Elle extrait des informations des modèles antérieurs tout en les clarifiant (Bachman & Palmer 1982 a : 450). Un modèle de compétence communicative est déployé. Celui-ci se compose des trois composantes que sont la compétence grammaticale, la compétence pragmatique et la compétence sociolinguistique (Bachman & Palmer 1982 a : 450). La compétence grammaticale comporte la morphologie et la syntaxe qui varient dans leur étendue et leur correction. La compétence pragmatique inclut trois composantes, qui sont le vocabulaire, la cohésion et l'organisation de la cohérence. Quant à la compétence sociolinguistique, elle inclut la sensibilité aux différents registres de la langue, à la formulation authentique ainsi qu'au contrôle du langage figuré ou pourvu de références culturelles (Bachman &Palmer 1982 a : 450). Le but de cette étude n'est pas seulement d'examiner la nature des composantes spécifiques de la compétence communicative contenus dans le modèle, mais également de déterminer le construit général de la compétence communicative (Bachman &Palmer 1982 a : 451).

Le niveau des compétences grammaticale, pragmatique et sociolinguistique a été mesuré à l'aide de quatre méthodes : une interview orale, un échantillon de performance écrite, un test au format QCM et une auto-évaluation (Bachman &Palmer 1982 a : 451). La combinaison de chacune des trois habiletés avec les quatre méthodes évoquées fournit douze tests qui sont regroupés en fonction des quatre méthodes utilisées. Tous ces tests ont été pilotés sur un échantillon d'étudiants non-natifs (Bachman & Palmer 1982 a : 451). Au cours de l'étude, plusieurs modèles de compétence communicative ont été analysés à l'aide de procédures statistiques. Le modèle validé par les données statistiques inclut la composante générale et deux composantes spécifiques (Bachman &Palmer 1982 a : 460). La première des composantes spécifiques est composée des compétences grammaticale et pragmatique. La deuxième composante spécifique est la compétence sociolinguistique (Bachman &Palmer 1982 a : 462).

3.3.2 Le modèle de la compétence langagière communicative de Bachman

Parce qu'il est essentiel de définir le concept de compétence langagière de façon exhaustive avant de développer un test, le modèle servant de base à la construction de POSILANG sera défini de façon détaillée (Brown 2010 : 109). Le modèle adopté est celui de la compétence langagière communicative développé par Bachman (1990), car il fournit un fondement large autant pour le développement et l'usage des tests que pour la recherche théorique portant sur l'évaluation (Bachman 1990 : 81). Ce modèle est issu de recherches menées en linguistique théorique et appliquée. Il s'est cependant enrichi de recherches plus empiriques, en évaluation langagière (Bachman & Palmer 1982a ; Bachman 1990 : 82).

Une raison décisive pour choisir le modèle de compétence communicative langagière de Bachman (1990) est que ce modèle défend une approche actionnelle pour l'enseignement, l'apprentissage et l'évaluation des langues (Dervin & Suomela-Salmi 2007 : 22).[17] Le CECRL lui-même s'est fortement inspiré de ce modèle (Alderson& Banerjee 2002 : 81). Le concept de compétence langagière communicative s'inscrit ici dans la continuité des travaux antérieurs sur la compétence communicative menés par Widdowson (1983) et Candlin (1987). Bachman reste fidèle aux travaux précédents en déclarant que la compétence communicative implique autant la compétence langagière que la capacité à utiliser cette compétence en contexte communicatif (Bachman 1990 : 81). La compétence langagière englobe la connaissance des règles grammaticales tout en étant rendue solidaire de la capacité à utiliser la langue d'une façon qui permette d'atteindre des objectifs communicatifs particuliers (Bachman 1990 : 83). Le modèle de Bachman (1990) contient également des aspects innovants. Il établit notamment des interactions entre des composantes jusque là séparées et intègre davantage le contexte d'usage de la langue (Bachman 1990 : 81).

Le choix du modèle de compétence langagière doit aussi dépendre du type de décision qu'on entend prendre à partir des scores obtenus (Bachman&Purpura 2008 : 463). Cette nécessité s'explique par le fait qu'un test peut uniquement évaluer les connaissances et les capacités définies dans le modèle choisi. Or, les informations sur celles-ci, obtenues via l'interprétation des

scores, doivent être pertinentes par rapport à la décision envisagée (Bachman & Purpura 2008 : 463). Concernant le test POSILANG, les informations obtenues à partir des scores doivent permettre de prendre les décisions visées, la première étant le placement des étudiants dans des groupes de niveaux adaptés à leur niveau individuel, en fonction des trois domaines de compétences évalués. La deuxième décision à prendre est le repérage des points forts et des déficits de chaque candidat, en conformité avec la fonction diagnostique prévue de POSILANG. Ces deux décisions doivent avoir un impact sur le choix du modèle de compétence à la base du test, qui permettra l'obtention d'informations pertinentes.

Un aperçu général du modèle ayant été donné, nous nous proposons d'analyser avec davantage de détail la première composante du modèle de Bachman (1990) : la compétence langagière. Elle est en effet d'une grande pertinence pour l'élaboration de POSILANG. En revanche, nous laisserons de côté la compétence stratégique et les mécanismes psychophysiologiques, qui jouent un moindre rôle.

3.3.2.1 La compétence langagière dans le modèle de Bachman

La compétence langagière englobe l'ensemble des connaissances utilisées en communication langagière. Cette compétence est répartie en deux types: organisationnelle d'une part et pragmatique d'autre part (Bachman 1990 : 86-87). Chacune de celles-ci se compose de plusieurs catégories. La compétence organisationnelle se subdivise en deux types, à savoir, la compétence grammaticale et la compétence textuelle (Bachman 1990 : 87). La compétence grammaticale inclut des connaissances relativement indépendantes les unes des autres, qui sont impliquées dans l'usage langagier. Il s'agit de la connaissance du vocabulaire, de la morphologie, de la syntaxe et de la phonologie/graphologie (Bachman 1990 : 87). Celles-ci sont indispensables afin de produire ou de comprendre les représentations de l'information qui sont linguistiquement justes (Bachman 1990 : 88).

Concernant la compétence textuelle, celle-ci inclut la connaissance des conventions qui règlent l'association et la structuration des phrases à l'intérieur d'un texte. Ces conventions sont déterminées par les règles de la cohésion et de l'organisation rhétorique. La cohésion implique le marquage explicite des

relations sémantiques, par exemple, de la référence et de la conjonction. L'organisation rhétorique renvoie à la structure conceptuelle globale d'un texte, et est liée à l'effet du texte sur son utilisateur. Les méthodes de développement du discours écrit, comme la narration, la description, la comparaison ou la classification, constituent les conventions de l'organisation rhétorique (Bachman 1990 : 88).

Le concept de compétence grammaticale établi dans ce modèle a tout particulièrement attiré notre attention lors de la conception des spécifications de POSILANG. Il faut néanmoins signaler une divergence importante entre la notion de compétence grammaticale telle qu'elle est présentée dans le modèle de Bachman et la définition de ce concept donnée par le CECRL. Contrairement à la présentation intégrative faite par Bachman, la notion de compétence grammaticale exposée dans le *Cadre européen commun de référence* dissocie le vocabulaire, la phonologie, l'orthographie de la grammaire. Ces trois dernières habiletés font partie des compétences lexicale, phonologique et orthographique (Conseil de l'Europe 2005 : 87, 91-92). La notion de compétence grammaticale présentée dans le CECRL n'en demeure pas moins complexe, puisqu'elle rassemble la morphologie et la syntaxe (Conseil de l'Europe 2005 : 90).

POSILANG a tranché en choisissant de s'appuyer sur la classification des habiletés du CECRL. En conformité avec le Cadre, l'étendue et la maîtrise du vocabulaire relèvent de la compétence lexicale. Quant à la notion de compétence grammaticale, elle englobe la morphologie et la syntaxe, comme le stipule également la nomenclature du CECRL. Cependant, les décalages existant entre le modèle de Bachman et le CECRL ne sauraient être exagérés car les deux documents identifient au final les mêmes catégories. [18]

Bien que les activités de communication langagière ne soient pas explicitement évoquées dans le cadre théorique de Bachman, elles y jouent un rôle important. Puisqu'il s'agit d'un modèle de compétence communicative, la maîtrise des habiletés n'est plus considérée comme suffisante. Il importe de pouvoir mettre en œuvre les habiletés acquises dans l'usage pertinent (ou approprié) de la langue dans un contexte donné (Bachman 1990 : 84). Ce modèle établit clairement ce qui est entendu par la notion d'usage pertinent (ou

approprié) de la langue. Il devient évident que le but ultime des habiletés contenues dans la compétence organisationnelle n'est pas de contrôler la structure formelle de la langue, mais de produire et de reconnaître des phrases grammaticalement correctes, de comprendre leur contenu propositionnel et de les associer afin de produire des textes entiers (Bachman 1990 : 87). Pour un bon usage communicatif de la langue, non seulement les habiletés nommées sont nécessaires, mais aussi les capacités communicatives productives et réceptives, appelées activités de production et de réception dans le CECRL (Conseil de l'Europe 2005 : 48-55). Puisque l'évaluation de ces capacités sert au but évoqué, elle est conforme au modèle de compétence communicative de Bachman.

3.4 Matériel

Les supports utilisés lors de la conception du test sont variés. Ils peuvent être répartis en documents officiels d'une part et en sources non-officielles d'autre part.

3.4.1 Documents officiels

Le document de référence principal est le CECRL**.** Il a servi de document d'orientation lors de la conception de la maquette du test, car l'adossement du test au CECRL implique l'adossement de chaque tâche à ce référentiel**.** Lors de la création des tâches, les descripteurs au sein des niveaux de compétences ont été étudiés en détail et pris en compte. Cependant, le CECRL n'a pas été le seul repère lors de la conception des items.

Les autres documents de référence sont les Instructions Officielles (IO) pour les collèges et les lycées, dont le palier 1 a été publié en 2006, et le palier 2 en 2008 par le Ministère de l'Education nationale, de l'Enseignement supérieur et de la Recherche.[19] Elles sont destinées aux enseignants d'anglais des établissements d'enseignement secondaire français. Elles présentent le nouveau programme d'enseignement d'anglais, adossé au CECRL. La concordance des Instructions Officielles au CECRL se manifeste, premièrement, par la « progression par palier », suivie par ces documents, qui est fondée sur les

niveaux de compétences du CECRL (Miller 2007 : 1).[20] Le palier 1, composé des niveaux A1 et A2, décrit les contenus linguistiques à maîtriser pour atteindre le niveau supérieur. Le palier 2 définit les connaissances et les capacités à acquérir pour se retrouver au niveau B1 (Miller 2007 : 1). Le deuxième élément de concordance entre les Instructions Officielles et le CECRL est l'approche actionnelle (Miller 2007 : 1). En effet, l'adoption de cette approche est soulignée par les Instructions Officielles. Un sous-chapitre entier est consacré à ce sujet, au début du palier 1 (CNDP 2006 : 5). Il est expliqué dans ce passage que les compétences linguistiques et culturelles servent à accomplir des tâches et donc ne représentent pas des fins en elles-mêmes (CNDP 2006 : 6).

Les Instructions Officielles fournissent des indications qui permettent de concrétiser les compétences définies par les descripteurs du CECRL (Westhoff 2007 : 676). En raison de la spécification insuffisante des connaissances grammaticales et lexicales attendues aux différents niveaux communs de référence du CECRL, les précisions apportées par les Instructions Officielles françaises se sont révélées très précieuses. En effet, lesdites instructions précisent, langue par langue, les contenus grammaticaux, lexicaux, phonologiques et culturels concrets à maîtriser à un niveau de compétence donné. Pour donner un exemple, nous allons regarder les contenus linguistiques fournis par les IO pour le Collège, palier 2 (CNDP 2006). Bien qu'il s'agisse de connaissances et non pas de compétences stricto sensu (cf. la différence établie entre ces deux concepts en début de chapitre), ces savoirs sont désignés par le vocable « compétence » en référence aux compétences linguistiques du CECRL. Des exemples de compétence grammaticale sont le futur avec *will,* les verbes à particule adverbiale*,* les composés de *some/any/no.* La compétence sociolinguistique implique, par exemple, les consignes de sécurité, les transports en commun ainsi que la radio et les chaînes de télévision. La compétence lexicale recouvre les mots et collocations liés aux domaines sociolinguistiques abordés. La compétence phonologique recouvre, entre autres, le repérage du schéma intonatif et des mots accentués (CNDP 2006 : 21-38). Puisque les niveaux couverts par le test POSILANG s'étendent de A1 jusqu'à C1, les *Programmes d'Anglais en classe de Seconde et en classe de Première,* en supplément des *Instructions Officielles pour le Collège*, ont été consultés lors de la conception des tâches.

3.4.2 Les sources non-officielles

En dehors des documents officiels, de nombreuses sources non officielles ont été consultées lors de la conception des tâches du test de positionnement. Parmi celles-ci figurent des manuels scolaires, des articles de journaux anglophones ainsi que des sites internet. Les sources employées peuvent être divisées en non authentiques et authentiques. Les manuels scolaires représentent la première de ces catégories. Une sélection a été utilisée afin de couvrir des niveaux et des domaines de compétence différents. Parmi les sources authentiques figurent en premier lieu des articles de journaux anglophones accessibles en ligne. On été également consultés des sites internet contenant des expressions langagières authentiques, très utiles pour repérer les collocations. L'ensemble des sources utilisées est présenté sous forme de tableau, en annexe de ce chapitre. Pour chaque source, nous indiquons la tâche concernée. Précisons que seule une partie des tâches s'appuie sur ce type de matériau langagier. D'autres ont été créées sans aucun support particulier.

3.5 Méthode

La méthode de conception d'un test est l'un des quatre facteurs à fort impact sur l'interprétation et l'usage des scores obtenus (Bachman 1990 : 165). Ce facteur est l'un des plus faciles à contrôler lors du développement et de l'usage d'un instrument d'évaluation. Cette caractéristique distingue la méthode des caractéristiques personnelles des candidats et des éléments aléatoires, qui sont difficilement maîtrisables lors de la conception et de l'utilisation des instruments d'évaluation (Bachman 1990 : 13). La possibilité de contrôler la méthode impose que celle-ci soit définie avec précision car elle sert de base au développement d'un test ainsi qu'à l'interprétation et à l'usage des scores obtenus (Bachman 1990 : 8).

L'influence de la méthode choisie va au-delà de l'évaluation de la performance et de l'interprétation des scores. La méthode a un fort impact sur la performance des candidats à un test en langues, comme de nombreuses études l'ont démontré (Bachman & Palmer 1982 a : 451, Bachman 1990 : 113 ; Purpura

2004 : 101). L'étude empirique, menée par Bachman & Palmer (1982 a) et décrite dans leur ouvrage, a démontré que le choix de la méthode a des répercussions sur les habiletés mesurées (1982a : 462). Ceci signifie que la méthode utilisée explique les variations de performance et ainsi le niveau de compétences attribué. Selon cette étude, l'interview et l'auto-évaluation sont deux méthodes qui ont un plus grand impact sur la performance des candidats que la production écrite et le format questionnaire à choix multiple (Bachman & Palmer 1982a : 462). Pour cette raison, il est indispensable de déterminer l'effet de la méthode sur la performance des candidats (Bachman 1990 : 12). Or, il est difficile de le faire, malgré l'existence de formules statistiques, car chaque individu est influencé par chaque méthode à un degré variable (Bachman 1990 : 113). L'influence de la méthode dépend des caractéristiques individuelles des candidats qui sont indépendantes du test, notamment, de leurs attributs cognitifs, affectifs, de leur parcours éducatif et de leur statut socio-économique (Bachman 1990 : 113-114).

L'influence forte de la méthode sur la performance des candidats s'explique par le fait que la méthode d'un test conditionne directement le type de tâches à effectuer (Alderson 2004 : 10). Il est évident que la méthode utilisée détermine la manière dont la tâche est conçue (McNamara 2000 : 5). Il en résulte que la conception des tâches varie en fonction de la méthode appliquée, y compris, en cas d'évaluation de la même connaissance ou capacité. Pour donner un exemple, les tâches qui ont pour objectif d'évaluer la compréhension écrite au travers de réponses fixes ou de réponses rédigées sont radicalement différentes. Une autre explication de l'influence de la méthode utilisée est qu'elle sert à contrôler le contexte dans lequel la performance langagière a lieu (Bachman 1990 : 111). Selon Firth, le contexte est le construit central dans l'étude de la langue, car les éléments du contexte déterminent l'usage de la langue dans les différentes situations de communication (Firth 1957 : 182). Il est clair que la méthode choisie a le pouvoir de contrôler le contexte de la performance langagière.

Malgré l'impact évident de la méthode sur la performance des candidats, le but d'utiliser une méthode particulière est d'obtenir des informations sur les connaissances, les capacités et les compétences langagières des candidats, et

non pas sur leur maîtrise d'une méthode particulière. Pour cette raison, la méthode est uniquement un moyen de déclencher un comportement particulier, appelé performance des candidats, qui révèle leurs capacités langagières (Hughes 2003 : 75). Puisque les scores attribués à une performance particulière sont normalement interprétés comme des indicateurs de compétences langagières, il est nécessaire de minimiser l'influence de la méthode utilisée sur les résultats obtenus. Il est clair qu'on ne peut réduire l'effet de la méthode qu'en choisissant celle-ci avec une grande attention et en considérant en détail les éléments qui la déterminent (Alderson 2004 : 10). Pour donner un exemple, les scores à un test utilisant le format QCM doivent permettre d'évaluer les compétences langagières des candidats et non leur maîtrise de ce format particulier (Bachman 1990 : 12). Il faut par conséquent prêter attention à cet aspect, déjà évoqué dans la sous-partie consacrée à la conception des items au format QCM. Quelle que soit la méthode utilisée, celle-ci doit respecter les quatre lignes directrices suivantes (Hughes 2003 : 75). Premièrement, la méthode employée doit déclencher le comportement qui est un indicateur valable d'une capacité en langue cible. Deuxièmement, le comportement réponse obtenu doit permettre l'attribution de scores fiables. Troisièmement, la méthode appliquée doit être pratique et économe sur le plan du temps et de l'effort. Enfin, la méthode choisie doit avoir un impact positif sur l'enseignement et l'apprentissage si le test est préparé en cours ou si sa passation conditionne les activités proposées en classe de langue (Hughes 2003 : 75).

Il est important de signaler que la méthode n'est pas une unité monolithique, mais qu'elle se compose de plusieurs caractéristiques qui exercent une influence sur la performance (Bachman 1990 : 113). Ceci est le cas aussi pour les méthodes permettant une évaluation objective (Bachman 1990 : 116). Certaines caractéristiques sont en général données, par exemple, l'objectif du test, le rôle des participants à la situation d'évaluation, la position subordonnée du candidat et supérieure de l'examinateur. Cependant, même ces qualités peuvent être changées. Il est possible de rapprocher les conditions de passation des caractéristiques de l'usage réel de la langue en rendant le test ainsi plus authentique (Bachman 1990 : 113).

3.5.1 La modélisation par facettes de Bachman

Cette thèse s'appuie sur la modélisation par facettes développée par Bachman (1990) pour rendre compte de la méthode de conception d'un test. La démarche n'est pas sans rappeler celle adoptée pour construire son modèle de la compétence langagière.[21] Le modèle proposé ici englobe cinq catégories : l'environnement d'évaluation, la rubrique d'évaluation, la nature de l'input reçu, la nature de la réponse attendue à cet input ainsi que la relation entre l'input et la réponse (Bachman 1990 : 119). Chaque catégorie se décline en plusieurs paramètres. En ce qui concerne l'environnement d'évaluation, il comporte quatre paramètres qui sont le degré de familiarité avec le lieu et l'équipement, le personnel, le moment de la journée et les conditions physiques (Bachman 1990 : 119). Le degré de familiarité est un facteur qui a un impact positif sur la performance car il réduit le potentiel menaçant d'une situation d'évaluation. Puisque les tests sur support papier sont plus courants que les dispositifs automatisés, il est possible que le format traditionnel induise une meilleure performance (Bachman 1990 : 119). L'effet bénéfique sur la performance se manifeste également en cas d'administration du test par un personnel connu des candidats. Concernant le test POSILANG, les deux premières facettes n'auront pas d'impact positif sur la performance car ni le lieu ni l'équipement utilisés, ni le personnel ne seront connus de la majorité des candidats. En revanche, les deux dernières facettes, le moment et les conditions matérielles de passation sont contrôlables. Il est prévu de créer des conditions optimales pour les candidats afin d'assurer la meilleure performance possible.

La deuxième catégorie, la rubrique du test, énonce un certain nombre de paramètres qui précisent la manière dont les candidats doivent procéder lors de sa passation. Les trois principaux sont l'architecture interne du test, l'allocation du temps et les instructions (Bachman 1990 : 120). Concernant l'organisation interne, on note que la structuration de la grande majorité des tests procède par division en parties, qui sont soit des tâches individuelles soit des ensembles des tâches. La performance des candidats est influencée par les trois éléments constitutifs de l'organisation du test, à savoir, la saillance des parties, leur ordre et enfin l'importance relative (Bachman 1990 : 120). En ce qui concerne le premier de ces éléments, la performance des candidats dépend de la saillance

des parties en tant qu'entités distinctes et des descriptions fournies par le concepteur. Ainsi, certains tests se composent de plusieurs sous-tests distincts. Ces derniers sont souvent identifiés en tant que tels par des intitulés et des descriptions brèves du domaine d'évaluation. Dans certains dispositifs, ces descriptions fonctionnent comme autant d'étiquettes appliquées aux différentes sous-parties (Bachman 1990 : 120). C'est le cas de POSILANG, par exemple, dont les trois sous-parties portent des étiquettes qui signalent le domaine de compétence évalué : la compréhension orale, la compréhension et la production écrites. Un dispositif dans lequel les sous-tests sont explicitement marqués mais ne sont ni étiquetés ni pourvus de descriptions est par exemple *Energy Placement Test*. Seules les mentions *Student A Test* et *Student B Test* sont fournies.

L'ordre dans lequel les différentes tâches sont présentées peut également avoir un impact sur la performance individuelle des candidats (Bachman 1990 : 120). Cet ordre est généralement fonction du niveau de difficulté et reflète l'intention du concepteur. Lorsqu'il n'y a pas de différence de niveau, l'ordre peut effectivement être aléatoire (Bachman 1990 : 120). L'ordre des tâches constitue un élément de contrôle de la performance exercé par le concepteur du test, mais auquel les candidats peuvent échapper en adoptant leurs propres stratégies de réponse (Bachman 1990 : 121). Cependant, dans les tests composés de plusieurs sous-tests, les candidats n'ont généralement pas la possibilité de répondre aux tâches dans l'ordre qu'ils souhaitent, à plus forte raison lorsqu'une contrainte temporelle est imposée. Le dispositif standardisé *Test of English as a Foreign Language* est un exemple de cet état de fait (Bachman 1990 : 121). L'ordre dans lequel sont disposées les tâches joue un rôle particulier dans les tests adaptatifs, comme il a été précisé au début du chapitre. La séquence est déterminée par les réponses précédentes du candidat (Bachman 1990 : 121). Lors de l'administration de POSILANG, il est prévu de se conformer au fonctionnement des tests adaptatifs, qui consiste à soumettre les tâches des niveaux inférieurs avant celles des niveaux supérieurs à chaque candidat, afin de déterminer le niveau individuel avec précision.

Les sous-parties constitutives d'un test sont également susceptibles d'avoir un impact sur la performance des candidats, mais seulement si les

candidats en sont conscients. En connaissant l'importance de chaque tâche ou de chaque sous-partie, les candidats peuvent adapter leurs stratégies de réponse, par exemple, en répondant d'abord aux tâches de plus grande importance. Dans le cas de POSILANG, il n'est pas prévu de fournir cette information aux candidats parce que chaque tâche a la même importance relative, le même score étant attribué partout. Cependant, les candidats n'en seront pas informés.

Le deuxième paramètre, l'allocation de temps, a également un impact sur la performance des candidats (Bachman 1990 : 122). Il faut préciser que ceci concerne les tests dans lesquels une contrainte temporelle est imposée. Dans ce genre de dispositif, les scores sont déterminés non seulement par le niveau de compétence des candidats, mais également par la vitesse de leurs réponses. Dans les dispositifs sans contrainte temporelle, les résultats obtenus sont en premier lieu fonction du niveau de compétence des candidats (Bachman 1990 : 123). Dans notre test de positionnement, une contrainte temporelle globale sera imposée mais grâce au format adaptatif seul le niveau de compétence sera évalué. Ce format permet de soumettre des tâches séparément et dans un certain ordre sans contrôle trop étroit de la vitesse.

Le dernier paramètre joue un rôle crucial dans la performance : il s'agit des instructions utilisées (Bachman 1990 : 123). Celles-ci ont pour fonction de renseigner sur les conditions de passation du test, les procédures à suivre et les types de tâches à accomplir. Les variables incluent la langue et le canal utilisés, la spécification des procédures et des tâches, l'explicitation des critères de correction des réponses (Bachman 1990 : 123). Les instructions peuvent être données en langue source ou en langue cible, ou encore sous forme combinée. En ce qui concerne POSILANG, le choix a été fait de fournir les instructions en français au niveau A1, pour assurer une compréhension par tous les candidats. Dans les niveaux supérieurs, en revanche, les instructions sont en anglais, les candidats étant censés maîtriser suffisamment la langue. Dans tous les cas, le canal de transmission est l'écrit uniquement, y compris dans les tâches d'évaluation de l'oral.

Les instructions (ou consignes) précisent en général les procédures à suivre par les candidats et les tâches à accomplir (Bachman 1990 : 124). Les procédures informent sur la manière de répondre, tandis que la spécification des tâches renseigne sur leur type et sur leur forme (Bachman 1990 : 124). Les consignes de notre propre test de positionnement ont en général une double fonction. Une consigne typique pour les tâches de compréhension de l'écrit est « Lisez les phrases suivantes et choisissez les bonnes formes de l'adjectif à insérer » (Tâche 2, Compréhension de l'écrit, Niveau A1). La procédure à suivre est fournie dans la première partie de l'instruction (lire les phrases). Le type de tâche à effectuer est indiqué dans la deuxième partie (sélectionner les réponses correctes). Certaines instructions dans POSILANG indiquent uniquement le type de tâche à accomplir, sans préciser la procédure à suivre. C'est le cas, par exemple, de la consigne de la tâche 5 « Quelle est l'idée principale du texte ? » (Tâche 5, Compréhension de l'écrit, Niveau A1). La consigne citée n'est pas la seule dans laquelle la procédure à suivre n'est pas indiquée car évidente. Ceci est légitime dans la mesure où la spécification des procédures et des tâches n'est pas indispensable mais seulement plus fréquente (Bachman 1990 : 123).

L'énonciation claire des critères de correction a un impact sur la performance des candidats. La raison à cela est que les candidats peuvent adapter leurs stratégies de réponse aux critères de correction. Cette adaptation ne signifie pas pour autant qu'ils atteignent un meilleur score (Bachman 1990 : 124). Pour donner un exemple, si la correction grammaticale est explicitement marquée comme un critère dans les instructions, cette information pourra avoir un effet inhibant sur les candidats qui n'oseront pas montrer la meilleure performance en expression écrite dont ils sont capables (Bachman 1990 : 124). Dans le cas de POSILANG, les critères de correction ne seront pas indiqués aux candidats pour ne pas influencer leur performance.

Les catégories constitutives de la méthode d'un test sont l'input et la réponse attendue (Bachman 1990 : 125). L'input est l'information contenue dans une tâche donnée. Il détermine, à côté des consignes et des spécifications de la tâche, la réponse attendue (Bachman 1990 : 126). La réponse attendue ne correspond pas forcément à la réponse effectivement donnée. Cette dernière résulte non seulement des paramètres évoqués, mais de facteurs qui sont au-

delà du contrôle du concepteur, comme la compétence langagière du candidat. L'input et la réponse attendue sont présentés ensemble en raison de leurs facettes communes, qui sont le format et la nature de la langue employée (Bachman 1990 : 126).

En ce qui concerne le format de l'input, les paramètres sont le canal, le mode, le format, le véhicule et la langue de présentation, sans oublier la problématique et la vitesse de présentation (Bachman 1990 : 127). La composante la plus générale est le format de présentation qui peut comprendre du matériau langagier, non-langagier ou une combinaison des deux. La forme de la réponse attendue peut être langagière, non langagière ou non verbale. La différence entre une réponse non langagière et non verbale est que la seconde n'implique ni langage ni paralangage (Bachman 1990 : 130). La réponse non verbale implique seulement une forme de marquage, comme dans les tâches à réponse fixe, où le candidat se contente de cocher une case. Les QCM de POSILANG appellent des réponses de ce type.

Le canal de présentation peut être visuel ou oral, voire une combinaison des deux. Dans POSILANG, les tâches évaluant la compréhension et l'expression écrites contiennent un input visuel, tandis que les tâches de compréhension orale affichent une combinaison des deux canaux. Les textes sont présentés à l'oral alors que les items eux-mêmes sont écrits. La combinaison de deux canaux au sein d'une même tâche constitue un procédé tout à fait possible (Bachman 1990 : 127). Le canal de réponse attendue est visuel dans toutes les tâches, y compris dans celles de compréhension orale, car la production orale n'est pas sollicitée dans ce test. Tandis que le mode de présentation peut être réceptif ou productif, il est uniquement réceptif dans POSILANG, y compris dans le domaine de production de l'écrit. L'adoption d'un mode d'expression unique s'explique par les restrictions imposées par le format adaptatif. Il aurait certes été plus authentique d'évaluer la production écrite en sollicitant des réponses plus articulées, mais cela aurait requis l'élaboration de tâches à réponse construite**,** ce qui est difficile à mettre en œuvre dans un test de ce type.

Le véhicule concerne le canal de présentation de l'oral. On distingue l'input à vive voix de l'input enregistré. Dans POSILANG, les textes des tâches de compréhension orale sont enregistrés en langue cible (Bachman 1990 : 127). L'identification des problèmes concerne la spécificité avec laquelle on identifie les difficultés dans les tâches et avec laquelle on attire l'attention du candidat sur celles-ci (Bachman 1990 : 127). Dans les tâches de POSILANG, les problèmes sont identifiés avec précision : grâce à un input au format QCM et grâce à des instructions qui attirent l'attention des candidats sur un enjeu bien précis.

La dernière composante du format est la vitesse de présentation de l'input. Dans la mesure où il a été décidé d'imposer une contrainte temporelle dans POSILANG, l'input est présenté à un certain rythme. Les réponses attendues doivent par ailleurs être données à une vitesse particulière. La vitesse est encore plus précise et sensible dans les tâches de compréhension de l'oral parce que les textes sont lus une seule fois, sans que le candidat ait la possibilité de les lire.

Bien que la notion de format soit commune à l'input et à la réponse, le format de la réponse a ceci de particulier qu'il se différencie selon la nature de la réponse attendue. On distingue deux types de réponses : la réponse à sélectionner et la réponse à construire (Bachman 1990 : 129). La réponse sélectionnée implique toujours un choix parmi plusieurs options possibles. Dans POSILANG, toutes les réponses sollicitées sont de ce type. La réponse à construire, en revanche, implique une production langagière, plus ou moins longue et structurée.

Le deuxième paramètre commun à l'input et à la réponse est la nature de la langue. Celle-ci détermine le degré d'intelligibilité de l'input et de la réponse d'un candidat (Bachman 1990 : 130). La nature de la langue engage la longueur du texte, son contenu propositionnel, ses caractéristiques organisationnelles et ses traits illocutoires. [22] Quant à la longueur de l'input, elle peut aller d'un seul mot à un discours constitué de plusieurs phrases. Dans POSILANG, l'input est de longueur variable, allant d'une phrase à un discours étoffé. L'exemple du deuxième cas de figure est cité ci-dessous :

> The Smiths live in a flat in York. It is a big flat with three bedrooms. Mr. and Mrs. Smith have got a small bedroom. Their son Patrick has got the second bedroom. His sisters Cathy and Jenny must share the third bedroom. And that is the problem. They quarrel

about everything all the time. They disagree about furniture, music and books (tâche 5, compréhension de l'écrit, niveau A1).

Le critère de longueur ne s'applique en revanche jamais à la réponse dans notre test, puisque nous avons choisi la modalité choix multiples (QCM).

Le contenu propositionnel de l'input et de la réponse attendue doit être décrit en tenant compte du vocabulaire, du degré de contextualisation, de la distribution de l'information, du thème et du genre. Le vocabulaire utilisé peut varier suivant la fréquence, le domaine d'usage ainsi que le nombre de références culturelles et d'expressions figurées. Le domaine d'usage joue un rôle parce que celui-ci détermine le degré de spécificité du vocabulaire, qui est élevé dans les domaines techniques notamment. Or, la spécificité et la fréquence du vocabulaire ont une influence sur la difficulté des tâches, qui est d'autant plus élevée que le vocabulaire est moins fréquent et plus spécifique (Bachman 1990 : 131). Ces deux qualités sont liées car plus le vocabulaire est spécialisé, moins il est courant donc susceptible d'être maîtrisé par la masse des candidats. Les références culturelles et les figures de rhétorique ont également un effet sur la difficulté des tâches. Leur évaluation est généralement réservée aux niveaux supérieurs bien que nous ayons posé quelques jalons dès le niveau B1 dans nos propres réalisations.

Dans POSILANG, le vocabulaire peu fréquent est inclus dans certaines tâches à partir du niveau B1. Pour donner un exemple, nous citerons l'énoncé de la tâche qui contient le mot composé peu fréquent « vehicle tires »:

Jean used to work in a company which produces motor vehicle tires.
What does this sentence express?

A. Jean still works in the company.
B. Jean produces motor vehicle tires.
C. Jean worked in that company.
D. Jean is going to work for the company
(Tâche 3, compréhension de l'oral, niveau B1)

A partir du niveau B2, certaines tâches sont entièrement construites à partir de langue spécialisée plus ou moins vulgarisée. Le degré de technicité peut cependant être élevé, comme c'est souvent le cas dans la presse et comme dans l'exemple ci-après, extrait de POSILANG, qui relève du domaine biomédical.

Stem cells can be compared to blank slates. Unlike regular cells, which can only replicate to create more of their own kind, stem cells are pluripotent: they can develop into any type of cell in the human body. In addition, they also have the ability to reproduce themselves many times over.
There are two types of stem cells: embryonic –and adult stem cells. Embryonic stem cells come from an embryo-the mass of cells in the earliest stage of human development that, if implanted in a woman´s womb, will eventually grow into a fetus.
Which description of stem cells is wrong?

A. Some of the cells are derived from the mass of cells in the earliest stage of human development.
B. The cells can turn into any cell type in the human body.
C. Stem cells can reproduce themselves several times.
D. Stem cells resemble regular cells of the human body.
(Tâche 4, compréhension de l'écrit, niveau B2)

Les figures de rhétorique, quant à elles, sont également contenues dans notre test de positionnement. Ainsi, une expression idiomatique à caractère métaphorique apparaît déjà au niveau B1, dans l'énoncé ci-après:

Jack declared that he is innocent and so he stood his ground in spite of the repeated accusations.

A. If you ***stand your ground****,* you keep your position and refuse to give up.
B. If you ***stand your ground****,* you stick to the ground.
C. If you ***stand your ground****,* you don´t move.
D. If you ***stand your ground,*** you are extremely smart.
(Tâche 1, Expression de l'écrit, Niveau B1)

Un autre exemple de figure de rhétorique à la fois simple et complexe, très spécifique et néanmoins courante, est le proverbe. Les proverbes sont également inclus dans notre test, par exemple, dans la tâche suivante :

Your mom tells you the following proverb after you report about your new job:
"(A) burden of one's own choice is not felt.
What does the proverb mean?

A. You work hard even if you choose what you like.
B. It´s not always good to make your own choices.
C. To make one´s own choice is a burden.
D. Something difficult seems easier when it is done voluntarily
(Tâche 5, Compréhension de l'oral, Niveau B2).

La question des proverbes nous mène logiquement à celle, plus large, des références culturelles, qui sont aussi présentes dans les tâches de POSILANG à partir du niveau B1. Un exemple de ces références culturelles est « *The Big Apple* » qui renvoie évidemment à New York (Tâche 4, Expression écrite, Niveau B1). Cependant, la compréhension de cette référence culturelle n'est pas

obligatoire pour pouvoir choisir la bonne réponse. En revanche, les références culturelles dans certaines autres tâches sont essentielles pour la compréhension de l'input. Ceci est le cas pour la référence culturelle « *American Dream* » contenue dans la tâche suivante :

> *You turn on the radio and listen to a part of a speech by Colin Powell. He was formerly Secretary of State of the Unites States*:
>
> The American dream is something that every immigrant brought to this country, as my parents did, and that is the ability to go as far as you can in life, limited only by your own dreams and willingness to work hard. And above all, the American Dream for these folks meant that your children will have the opportunity to do better than you will.
>
> Find a synonym for the word underlined.
> A. future
> B. chance
> C. freedom
> D. luxury
> (Tâche 3, Compréhension de l'oral, Niveau B2).

La deuxième composante du contenu propositionnel est le degré de contextualisation des tâches. Ce terme est défini comme la part d'information contextuelle dans l'information globale que renferme l'instance de discours considérée (Bachman 1990 : 131). L'emploi d'une langue riche en références contextuelles engendre un discours à haut degré de contextualisation, tandis que l'usage d'une langue faiblement contextualisée génère un discours à faible degré de contextualisation (Bachman 1990 : 131). Il existe trois types d'information contextuelle dont deux seulement peuvent être mises en œuvre dans notre test de positionnement.[23]

Le premier se manifeste dans les souvenirs suscités par l'information contenue dans l'input (Bachman 1990 : 133). La tâche décrivant le « Rêve Américain» citée ci-dessus en constitue un bon exemple. S'agissant d'une référence culturelle très connue, cette expression déclenche des souvenirs particuliers chez les candidats leur permettant d'accéder au contenu informationnel de l'input tout en activant probablement d'autres connaissances ou expériences en lien avec cette référence.

Le deuxième type d'information contextuelle est développé dans l'input lui-même. Ceci n'est possible que dans un énoncé d'une certaine longueur

(Bachman 1990 : 133). Un exemple de ce type de contextualisation est fourni par la tâche suivante, extraite de POSILANG :

> *Ecoutez la conversation suivante. Où est-ce que Jack doit aller en dernier pour trouver la poste ?*
> - Jack: Excuse me, sir, could you tell me the way to the next post office, please?
> - Sir: Yes, of course. First, you go along Miller Street. Then, you turn left into Grove road. After that you turn right into the Cherry Avenue and you see the yellow post office.
>
> .
>
> A. He has to go along Miller Street.
> B. He has to turn right into the Cherry Avenue.
> C. He has to turn left into Grove road.
> D. He has to pass a crossroad.
> (Tâche 4, Compréhension de l'oral, Niveau A1).

Dans la tâche citée, l'information contextuelle est fournie dans la question où mention est faite du bureau de poste. De ce fait, la réponse qui sera donnée à cette question représente une utilisation contextualisée du langage.

Il est important de signaler que de nombreuses tâches proposées dans POSILANG sont peu contextualisées et que cette décontextualisation est un choix assumé de notre part. En premier lieu parce que la mise à disposition d'information contextuelle dans un test crée des différences entre candidats, en raison de l'inégale capacité des individus à repérer et à traiter les données situationnelles (Bachman 1990 : 134). Or, notre intention lors de la conception de POSILANG était de minimiser les décalages d'expérience et de culture entre les candidats. Ceci afin d'éviter de possibles biais allant à l'encontre de l'équité du test (Kunnan 2010 : 3). Afin de réduire les disparités entre candidats, la bonne pratique veut qu'on élabore des tâches réduites en contexte (Bachman 1990 : 133). C'est ce que nous avons fait. Par ailleurs, il est plus difficile de répondre à des tâches moins contextualisées. En effet, la quantité d'information contextuelle a un impact notoire sur la capacité d'un candidat à interpréter le contenu propositionnel du discours et à y répondre (Bachman 1990 : 133). Plus l'input est ancré dans un contexte, plus il est facile au candidat de répondre correctement au contenu propositionnel d'une tâche, en réduisant la part des formes langagières elles-mêmes (Bachman 1990 : 133). Or, il n'est pas dans notre intention de faciliter la tâche des candidats en leur fournissant toute l'information contextuelle dont ils ont besoin pour inférer et non décoder du sens. Notre but reste d'évaluer une capacité à comprendre et à manipuler les formes de la

langue, même si nous reconnaissons volontiers que la compétence pragmatique et les capacités d'inférence situationnelle font partie intégrante des compétences langagières générales.

La distribution de l'information est la manière dont celle-ci est répartie dans l'input et doit être traitée par les candidats. Un discours dans lequel l'information nouvelle est distribuée pendant une période de temps très brève sera qualifié de « compact ». Dans le cas contraire, le discours sera appelé « diffus » (Bachman 1990 : 134). Dans POSILANG, l'information nouvelle est très majoritairement distribuée en mode compact aux candidats. La durée étant imposée, les candidats n'ont ni la possibilité de faire répéter ni celle de faire ralentir pour reconstruire le sens de ce qu'ils lisent ou entendent (Bachman 1990 : 135).

Le thème de l'input doit être choisi avec la plus grande attention car ce paramètre influence la performance des candidats de manière considérable (Bachman 1990 : 136). Les sujets qui se prêtent bien à l'usage des tests de langue doivent à l'évidence être intéressants et pertinents pour tous les candidats. Cependant ces sujets doivent aussi éviter de favoriser certaines personnes et ce faisant de constituer un biais. Il y a là un dilemme pour tous les concepteurs des tests dont il est difficile de s'extraire. Une issue possible est le déploiement d'un large éventail de sujets (Bachman 1990 : 136). Conformément à cette préconisation, POSILANG inclut un vaste répertoire de thèmes et de situations. Ainsi trouve-t-on une grande diversité de sujets dans chaque domaine de compétences et à chaque niveau. Les sujets couvrent les quatre domaines d'usage distingués par le CECRL : public, professionnel, éducatif et personnel (Conseil de l'Europe 2005 : 18). Pour illustrer ce point, il suffit de considérer les tâches de compréhension écrite de niveau A1. Les trois premières tâches recouvrent le domaine public car elles décrivent des échanges sociaux. Les trois tâches suivantes relèvent du domaine personnel : elles décrivent des pratiques sociales individuelles, notamment les relations familiales. Le domaine de compétence présenté, à savoir la compréhension de l'écrit, inclut ainsi des tâches qui relèvent de deux domaines d'usage sur quatre. Partout dans POSILANG, l'input renvoie à une grande diversité des sujets. Cette diversité a été créée intentionnellement dans le but de rendre le test intéressant pour les

candidats et, en même temps, afin de minimiser le biais associé à un choix de sujets trop étroit.

Au thème de l'input sont associés un type d'information et des réponses attendues. Le type d'information peut être classé selon trois dimensions, exprimées par les oppositions concret/abstrait, positif/négatif, factuel/contrefactuel (Bachman 1990 : 135). Concernant la première dimension, pertinente pour notre test, on trouve des inputs renfermant des informations abstraites dès le niveau A1, mais la proportion va grandissant au fur et à mesure que le niveau augmente [24]. Ainsi, au niveau A1, seules trois tâches contenant des informations abstraites sont proposées. En revanche, au niveau C1, la proportion est inversée et seules trois tâches s'avèrent porteuses d'informations concrètes. Le pourcentage croissant d'information abstraite au fur et à mesure que le niveau s'élève s'explique en théorie par la plus grande difficulté à comprendre ce type d'information (Bachman 1990 : 136). Cependant, il est permis de s'interroger sur le cas des candidats francophones et plus généralement des locuteurs de langues romanes. Le latin et le français ayant largement contribué à construire le lexique abstrait de l'anglais, la difficulté n'est peut-être pas si grande et pourrait même s'inverser. Les francophones sont parfois plus embarrassés pour se repérer concrètement dans l'espace ou dans le temps, pour choisir une préposition ou une particule adverbiale avec un verbe de mouvement que pour raisonner et abstraire en anglais.

Le dernier paramètre affectant l'input est le genre auquel appartient le discours. Celui-ci peut être défini à partir de caractéristiques formelles identifiables, traditionnellement reconnues (Hymes 1972 : 65). Hymes énumère quelques exemples: « poem, myth, tale, proverb, riddle, curse, prayer, oration, lecture, commercial, form letter and editorial » (Hymes 1972 : 65). On peut distinguer une grande variété de genres dans POSILANG. On trouve notamment des interviews, des reportages, des lettres, des proverbes, des annonces, des entrées de journal de bord, etc. Dans ce test, le genre est souvent explicité dans les consignes de la tâche. Le premier exemple est une carte postale:

Read the following post card and choose between simple past and present perfect.

Dear Anna,

I am writing to you from Barcelona. It is great here! It´s sunny and lively. Yesterday, we 1.____ (go) to a great restaurant and 2._____(eat) tasty tapas. I 3.____ (drink) so much sangria in my life.
Hope to see you soon back in London.
Big hug,
Luise

Gap.1.	Gap2.	Gap3.
A. go	A. eaten	A. never drink
B. went	B. have eaten	B. never drank
C. have gone	C. ate	C. never drinked
D. going	D. eated	D. have never drunk

(Tâche 1, Expression écrite, Niveau A1)

L'exemple suivant est également rattaché à un genre de discours particulier, mais celui-ci n'est pas explicité dans les instructions.

Read the following paragraph and select where it is taken from:

We have got lots of activities for young people under 16. Sport does not just mean football! We have a big choice of different sports for you! Special prices for pupils. Come and get information about our offers.

A. book
B. newspaper article
C. prospectus
D. instruction booklet

(Tâche 5, Expression de l'écrit, Niveau A1)

C'est à dessein que le genre n'est pas indiqué dans la consigne car il faut que les candidats l'identifient eux-mêmes s'ils veulent trouver la bonne réponse. Cependant, force est d'admettre que le genre n'est pas un critère pertinent dans de nombreuses tâches proposées dans POSILANG. Néanmoins, certains chercheurs estiment que le genre est omniprésent dans les tests, que la dictée, le questionnaire à choix multiple constituent en eux-mêmes des genres identifiables (Bachman 1990 : 138). Les tests qui empruntent ces modes d'expression et d'interaction avec les candidats, facilitent la tâche aux candidats familiers de ces genres particuliers. En revanche, une méconnaissance (ou une expérience insuffisante) de ces genres rend l'effectuation de la tâche plus difficile (Bachman 1990 : 139). A ce titre, la batterie de QCM déployée dans POSILANG est censée faciliter la passation pour une majorité de candidats, puisque le genre adopté est commun. Même s'il ne l'est pas pour tout le monde, les instructions sont claires et détaillées. De ce fait, il n'y a pas besoin de connaître ou d'avoir

passé d'autres tests au format QCM pour répondre correctement aux items contenus dans POSILANG.

Le paramètre suivant concerne l'organisation (ou la structuration) du discours.[25] L'organisation se décline en trois catégories : la grammaire, la cohérence et la rhétorique. La langue utilisée dans notre test est articulée à partir de ces trois critères pour assurer une intelligibilité optimale (Bachman 1990 : 139). Quand les éléments de structuration du discours ne sont pas respectés dans POSILANG, en l'occurrence les règles de grammaire, c'est que les fautes ont été sciemment incluses pour être repérées et corrigées par les candidats :

> *Read the following police report and choose the sentence which contains a grammatical mistake:*
>
> A. A red Mercedes was stolen on 3rd March between 3 and 5 pm.
> B. The suspect has been arrested yesterday.
> C. The man was accused of the crime last night.
> D. He has already been presented to the judge
>
> (Tâche 3, Expression écrite, Niveau B1).

Concernant les réponses attendues, des restrictions sont imposées à plusieurs niveaux. D'abord, au niveau du format et du canal de la réponse, puisque seules des réponses écrites au format QCM sont acceptées. Ensuite, au niveau du temps et de la longueur. La contrainte temporelle est essentiellement imposée pour des raisons pratiques, même si la vitesse peut être considérée ailleurs comme un indicateur d'aisance ou de performance. Le cadrage de la réponse est fonction du format choisi. Il est extrême dans le cas du QCM ou de l'exercice lacunaire. Cela peut paraître très artificiel mais il ne faut jamais oublier que l'usage réel de la langue est lui aussi sujet à des contraintes. Cela étant admis, il est clair qu'un excès de formatage et trop de restrictions frappent la majorité des tests de langues. L'une des conséquences de cet excès est une réduction de la variabilité, de la spontanéité et de l'authenticité de la performance langagière (Bachman 1990 : 148).

La dernière catégorie constitutive du cadre présenté est la relation entre l'input et la réponse. Dans notre test, la relation est non réciproque en raison d'un manque d'interaction et de feedback. En effet, non seulement les tests au format QCM, mais la majorité de dispositifs d'évaluation, sont non réciproques. Cela ne remet pas en question leur validité, fiabilité ou utilité. Etant donné que l'usage

langagier est en général souvent non réciproque, ces tests constituent une approche légitime de l'évaluation des compétences langagières (Bachman 1990 : 150).

3.5.1.1 Application du cadre méthodologique et de ses composantes à l'évaluation langagière

Dans tout type de test, la méthode utilisée influence la performance des candidats (Purpura 2004 : 101). Or, il n'est pas possible de comprendre la manière dont les techniques utilisées affectent cette performance sans procéder au préalable à l'examen du cadre méthodologique et de ses facettes (Bachman 1990 : 156). Une grande variété de techniques est en effet employée dans les tests de langue (Bachman 1990 : 156). Le cadre méthodologique est conçu comme un guide permettant d'analyser les tests, de les développer, de contrôler leur usage et plus généralement de mener une réflexion sur l'évaluation en langues (Bachman 1990 :156). Le rôle que ce cadre méthodologique est susceptible de jouer dans tous ces domaines mérite d'être brièvement expliqué.

Concernant le recensement des dispositifs d'évaluation, le cadre peut être utile pour décrire et comparer des tâches qui sont contenues dans un ou plusieurs tests, ou encore référencées dans un programme (Bachman 1990 : 156). [26] Il est également nécessaire de tenir compte de ce cadre lors de la conception d'un nouveau test. Dans ce cas, les éléments de la méthode doivent être intégrés au descriptif et aux objectifs du test. Outre les compétences visées par un test donné, le descriptif doit établir la méthodologie avec précision (Bachman 1990 : 156). En effet, la caractérisation précise des tâches n'est possible que si on détermine avec rigueur les éléments de méthode. La nature et la fonction de l'input, la réponse attendue sont indispensables pour concevoir des tâches individuelles cohérentes et pour relier celles-ci à des séquences d'enseignement et d'apprentissage (Bachman 1990 : 154). Cette préconisation méthodologique, qui fait partie des bonnes pratiques, a été intégralement respectée lors de l'établissement des spécifications de POSILANG. Cela semble aller de soi mais il n'est pas inutile de rappeler ici que de nombreux concepteurs de tests, pris par l'urgence et faute de contact avec la recherche, font l'économie de cette étape.

L'explicitation du cadre méthodologique fait partie intégrante du processus de validation d'un test (Bachman 1990 : 155). Quel que soit le test, l'identification des compétences ne suffit pas. La méthode d'évaluation choisie, ses possibles effets sur les performances observées, les biais envisageables doivent impérativement faire l'objet d'une recherche (Bachman 1990 : 155). Cette recherche est censée permettre d'élaborer les tests dans lesquels la performance reflète bien les compétences en langue préalablement ciblées et non pas la maîtrise d'une procédure par les candidats.

3.6 L'élaboration d'un test

L'élaboration d'un test peut parfois être facile ou au contraire s'avérer complexe, longue et coûteuse. Ce second cas se présente lorsque le test revêt une certaine importance pour l'avenir des candidats ou tente d'évaluer leurs compétences de façon exhaustive (Douglas 2010 : 6). Quelle que soit la situation, le processus de création d'un test exige toujours des concepteurs qu'ils se posent une série de questions préliminaires et qu'ils y apportent des réponses satisfaisantes. Il faut bien avoir conscience qu'il n'existe pas de réponse bonne ou mauvaise à des questions comme : Quel type de test utiliser ? Quelle est la longueur idéale? Comment interpréter au mieux les résultats ? La réponse nécessite la prise en compte de critères comme l'usage auquel le test est destiné, la manière d'interpréter et d'utiliser les résultats obtenus, ainsi que les conditions dans lesquelles il sera administré (Bachman 1990 : 1).

La construction d'un test standardisé représente dans tous les cas un grand défi, quelle que soit sa taille. Trois raisons principales sont identifiables (Brown 2010 : 120). La première tient à la nécessité de fonder le dispositif sur des normes de performance bien construites qui ne contiennent pas de biais. Cela exige la collecte et l'analyse empirique des données de performance et des objectifs poursuivis par l'institution. La deuxième tient à la complexité des spécifications. Un test exige la définition et la validation du concept de compétence langagière, elle-même composée de sous-compétences particulières. En dehors de la validation du concept de compétence langagière, située à la base du test, l'élaboration des spécifications demande la prise en compte du principe de praticité (Brown 2010 : 120). La troisième difficulté lors de

la construction d'un test standardisé provient du long travail nécessaire à la conception et à la validation des items ainsi qu'à l'établissement des procédures d'attribution et d'évaluation des scores. L'optimisation de ces données demande souvent plusieurs essais d'amélioration et l'élaboration d'un certain nombre d'esquisses. Cependant, quel que soit le type de test envisagé, les étapes lors de la conception et de la mise en œuvre sont identiques (Brown 2010 : 120).

La conception et l'implémentation d'un nouveau test nécessitent plusieurs étapes distinctes et successives avant de pouvoir atteindre la phase d'opérationnalisation. Ces étapes sont la conception, la construction et la mise au point d'un instrument d'évaluation (McNamara 2000 : 23). En apparence, ce processus de développement est linéaire, mais en réalité il ne l'est pas. Il s'agit en fait d'un processus cyclique car toutes les étapes sont intégrées (Fulcher & Davidson 2007 : 62). Toutefois, il faut avoir l'humilité de reconnaître que quel que soit le nombre de sessions de pilotage et d'évaluation qu'on engage avant la mise en œuvre d'un test, il n'y a au final que l'usage collectif et effectif du test qui fournisse de réelles preuves de sa qualité (McNamara 2000 :23). Une qualité et une adéquation qui ne sont jamais définitives, puisqu'il faut être prêt à adapter le test à des contextes sans cesse évolutifs, même après opérationnalisation. Les modifications imposées par la collecte des données d'usage [27] (Fulcher & Davidson 2007 : 62) sont d'une importance capitale car les performances effectives des candidats permettent de se prononcer sur leur capacité réelle d'utilisation de la langue, dans les domaines et les situations définis (Fulcher & Davidson 2007 : 62). Une tâche est validée dès lors que les conclusions qu'on peut tirer de sa bonne réalisation sont conformes aux paramètres précis, établis dans les spécifications du test (Fulcher& Davidson 2007 : 62). Dans les pages suivantes, les différentes étapes ayant présidé à l'élaboration du test POSILANG seront donc décrites et expliquées avec un certain détail.

3.6.1 Le respect des principes valables pour un test

Les six principes applicables à l'évaluation d'un test, que nous avons présentés dans notre premier chapitre, s'appliquent également à la construction des dispositifs d'évaluation. L'objectif majeur d'un test est de fournir des informations

utiles et fiables sur les compétences des candidats (Bachman & Palmer 1996 : 231-233).

Puisque il a été décidé d'élaborer et de mettre en œuvre un test de positionnement à l'Université de Bordeaux, il est particulièrement important d'assurer la validité prédictive de ce test. Un dispositif fait preuve de validité prédictive s'il détermine la probabilité du succès des candidats de manière précise (Brown 2010 : 33). Etablir cette validité est de toute première importance pour un test de positionnement comme POSILANG.

3.6.2 Les étapes de l'élaboration d'un test

Les étapes présentées comme nécessaires à la conception et à la mise en œuvre d'un test varient d'un spécialiste à l'autre. Selon certains chercheurs, trois grandes étapes successives sont considérées comme nécessaires et suffisantes pour l'élaboration et la mise en œuvre d'un dispositif (Alderson, Clapham & Wall 1995 : 5). La première consiste à élaborer les spécifications sur lesquelles le test se fonde. La seconde étape consiste à concevoir les tâches faisant partie de la maquette du test. La troisième étape est celle du pilotage du test, qui consiste à tester les items de la maquette en amont de l'opérationnalisation du dispositif (Alderson, Clapham & Wall 1995 : 5). L'élaboration de POSILANG respecte ces trois étapes.

D'autres spécialistes estiment que le processus de développement d'un test doit forcément comprendre cinq étapes afin qu'il soit un instrument d'évaluation efficace (Brown 2010 : 55). Selon eux, ces étapes sont indispensables pour la conception, l'administration, l'attribution des scores et l'évaluation générale de la passation. La première étape consiste à formuler l'objectif principal du test. Certains experts parlent ici de « *test usefulness* », ce qui peut être traduit par le terme *utilité* en français (Bachman & Palmer 1996 : 17-19). L'utilité s'impose comme une notion pertinente car l'objectif principal d'un test, quelle que soit la complexité de son élaboration, est de s'imposer comme un instrument utile du point de vue pratique, en livrant des informations fiables sur les compétences des candidats (Bachman & Palmer 1996 : 231-233).

Après avoir déterminé le but principal d'un test, il importe de clarifier ses objectifs. Il faut notamment décider des critères permettant de déterminer que les candidats ont bel et bien atteint lesdits objectifs. Il est important de noter que les conséquences du test pour le destin scolaire ou professionnel des candidats font partie de ces objectifs. Quelles qu'en soient la nature et l'étendue, il est indispensable de formuler l'ensemble des objectifs de façon aussi claire et explicite que possible (Brown 2010 : 108). La formulation des objectifs du test et des critères marquant leur réalisation par les candidats nécessite une détermination des contenus. Ces contenus doivent être en cohérence avec les objectifs visés (Brown 2010 : 108). Ainsi s'opère une manière de bouclage : fonction-objectifs-contenus.

La deuxième étape prévoit l'élaboration des spécifications d'un test. Celles-ci doivent prendre en considération le but principal et les objectifs du test. La conception des spécifications prévoit plusieurs étapes successives. Les spécifications contiennent plusieurs éléments nécessaires à la conception du test, l'attribution des scores, l'évaluation des résultats ainsi que la manière d'en écrire un rapport aux candidats (Brown 2010 : 109). Il faut commencer par décrire le contenu envisagé du test, c'est-à-dire les domaines et les composantes langagières sur lesquels celui-ci va porter (Brown 2010 : 109).

La phase de conception des items doit être suivie de leur révision. Lors de l'étape de révision, la vérification de plusieurs aspects affectant la validité et la fiabilité des items est indispensable. La clarté des instructions doit être absolue du point de vue du contenu et de l'expression langagière. Le niveau de difficulté des tâches doit être étendu, tout en restant adapté à ce que les candidats sont potentiellement capables d'accomplir. Il faut en outre veiller à ce que chaque item mesure un objectif précis et que le test dans son entier reflète l'intégralité des objectifs d'évaluation (Brown 2010 : 77).

La quatrième étape concerne l'administration du test. Il s'agit de définir les informations à fournir aux candidats en amont, comme les contraintes temporelles ainsi que les matériaux autorisés. Il convient également d'informer les candidats du type d'items utilisés, des critères d'évaluation. Il est également important d'expliciter les stratégies permettant d'optimiser leur performance

(Brown 2010 : 78). Lors de l'administration du test, il est indispensable de faire preuve d'un haut niveau d'organisation afin de prévenir les incidents liés à un manque de précautions (Brown 2010 : 78).

En dernier lieu, il convient de décider de la manière d'attribuer des scores aux différentes tâches en répartissant celles-ci en fonction de la difficulté et de la complexité relatives des différentes parties du test (Brown 2010 : 80). Une fois le barème défini, il reste encore à déterminer la manière de donner le feedback. Il existe une multitude de possibilités pour un même test. Les principales formes de feedback pratiquées sont les suivantes. On peut d'abord choisir d'attribuer un score total ou dissocié en fonction des différentes parties ou selon les compétences visées (Brown 2010 : 80). Les scores peuvent être diagnostiques, c'est-à-dire renvoyer à des catégories langagières particulières. Cependant, le feedback peut également être moins formel. Un commentaire peut être donné, sous forme orale ou écrite. On peut également identifier des domaines ou des compétences qui nécessiteraient d'être améliorés, au vu du résultat, en énonçant des stratégies de remédiation. Les échanges entre les étudiants ou entre un étudiant et un enseignant constituent une autre façon de donner un retour. La difficulté consiste à choisir une stratégie de feedback qui soit non seulement faisable mais aussi utile (Brown 2010 : 80). Il faut garder à l'esprit que le retour doit avoir un impact positif sur les candidats pour satisfaire au cinquième principe de la conception des tests, qui est l'impact bénéfique de l'évaluation sur l'apprentissage futur (Brown 2010 : 80). Le développement de chaque test nécessite le parcours de ces cinq étapes car celles-ci garantissent que la mesure des capacités et les conclusions tirées à partir des performances seront aussi précises, équitables et utiles que possible (Douglas 2010 : 63).

3.6.3 Etape 1: L'élaboration des spécifications d'un test

Les spécifications énoncent de façon formelle et officielle les domaines évalués et la méthodologie d'évaluation du test. Elles sont essentielles pour plusieurs raisons. En premier lieu, elles constituent le plan à suivre par les concepteurs du test (McNamarra 2000 : 31). Ensuite, elles sont indispensables pour la validation des compétences évaluées. Le développement et la publication des spécifications constituent une partie cruciale du processus de construction et d'évaluation d'un test (Alderson, Clapham & Wall 1995 : 5).

Les spécifications se distinguent d'un programme par leur contenu, qui est plus détaillé, et par le public auquel elles s'adressent. Un programme est destiné aux enseignants et aux candidats souhaitant se préparer à un test, ou encore à des professionnels prenant des décisions sur la base des scores obtenus. Les spécifications, en revanche, s'adressent aux concepteurs ainsi qu'aux spécialistes chargés de déterminer si le test a rempli les objectifs visés. En outre, le contenu et le format des spécifications varient selon les destinataires (Alderson, Clapham & Wall 1995 : 9). Ainsi, les spécifications destinées aux concepteurs du test doivent être aussi détaillées que possible, car elles servent de guide lors de la construction d'un dispositif d'évaluation (Alderson, Clapham & Wall 1995 : 10). Ce format doit répondre aux douze questions principales qui couvrent les différents aspects du test (Alderson, Clapham & Wall 1995 : 9). Pour illustrer ces catégories, les questions auxquelles ce genre de spécifications doit répondre, seront présentées ci-après (Alderson, Clapham & Wall 1995 : 10). Ces questions serviront de base à l'élaboration des spécifications du test POSILANG présentées par la suite.

Premièrement, il est essentiel de connaître le but principal du test. Celui-ci doit être le premier renseignement figurant dans les spécifications car l'objectif détermine la catégorie dont un dispositif donné fait partie. [28] Comme il a été expliqué dans le chapitre précédent, le but sert de critère de répartition des tests entre un nombre de catégories limitées (Alderson, Clapham & Wall 1995 : 9- 10). Le deuxième point concerne le public auquel le test est destiné. Idéalement, il faut connaître l'âge et le sexe de la personne, autant que sa langue d'origine, le niveau de compétence en langue évaluée, mais aussi le niveau de culture

générale et d'éducation, la raison pour passer le test, ainsi que les intérêts personnels et professionnels (Alderson, Clapham & Wall 1995 : 10-11). Troisièmement, il est nécessaire de réfléchir à la structure du test, c'est-à-dire au nombre de parties contenues, à la durée de chacune des sections et à la manière de répartir ces dernières au sein du test, de façon séparée ou intégrée. En quatrième lieu, il faut décider quelle situation en langue cible est envisagée par le test et si celle-ci est simulée en quelque sorte par le contenu et les méthodes choisis (Alderson, Clapham & Wall 1995 : 10-11). Le cinquième ensemble de réflexions concerne le type de textes utilisés : le canal de communication (oral ou écrit), les sources, le public envisagé et la longueur. Cette catégorie de questions englobe également le degré d'authenticité et de difficulté des textes, ce qui dépend en partie de la longueur et aussi de la complexité de la langue utilisée (Alderson, Clapham & Wall 1995 : 11). L'aspect suivant concerne les compétences langagières soumises à évaluation ainsi que la meilleure façon de répartir cette évaluation entre les tâches. Ici se pose la question de tester les compétences individuellement ou de façon intégrée. L'enjeu suivant explore les éléments langagiers à tester. Ici doit être décidé de spécifier ou non les structures grammaticales et le répertoire lexical évalués (Alderson, Clapham & Wall 1995 : 11). La huitième question se réfère aux types de tâches requises, par exemple, authentiques, intégrées, objectivement évaluables, etc. (Alderson, Clapham & Wall 1995 : 11). La neuvième interrogation concerne le nombre de tâches requises pour chaque section, ainsi que la pondération de chaque tâche. La pondération peut se faire de manière égale ou différenciée, en fonction de la difficulté. La dixième question s'intéresse aux méthodes utilisées en détail (Alderson, Clapham & Wall 1995 : 11). Le onzième aspect concerne les instructions fournies aux candidats. Enfin, la douzième et dernière question se réfère aux critères d'évaluation utilisés par les correcteurs. Ce système de classification a été appliqué à notre test de positionnement. Il est illustré dans le sous-chapitre suivant.

3.6.3.1 Les spécifications du test POSILANG

Le but principal de notre test est le positionnement des étudiants par groupes de niveaux de compétences. Ces groupes de niveaux de compétences, organisés en aval de la passation du test par les autorités pédagogiques, vont accueillir les

étudiants d'un même niveau de compétences. En ce qui concerne le public du test, il varie selon le niveau de compétences en anglais, qui peut s'étendre du niveau A1 au niveau C1. Les candidats se distinguent également selon leurs intérêts professionnels et personnels. En revanche, d'autres paramètres sont comparables, par exemple, la langue d'origine, le niveau de culture générale et d'éducation, aussi bien que l'âge des candidats.

Structuralement, POSILANG se compose de trois parties : compréhension orale, compréhension écrite et expression écrite, réitérées pour chaque niveau de compétence. Ces parties sont séparées et peuvent donc être passées de façon autonome. Une situation en langue cible est envisagée pour une partie des tâches. Les situations cibles se distinguent de celles évoquées pour un test basé sur la performance parce que POSILANG est un test standardisé. En tant que tel, il n'évalue pas les compétences langagières dans un acte de communication (Mc Namarra 2000 : 6).

Notre test combine deux canaux de présentation, visuel et oral. La longueur, l'authenticité et évidemment la difficulté varient selon les tâches. Il y a des tâches qui se composent d'une seule phrase comme la suivante:

> *Lisez cette annonce trouvée dans un journal scolaire. Ensuite, choisissez un synonyme pour le mot souligné :*
>
> For winning the football match the class of the 10th grade won the popular award.
> A. Prize
> B. Food
> C. Book
> D. Judgement
>
> (Tâche 2, Expression écrite, A1).

La longueur de certains autres textes s´élève à une demi-page. Concernant l´authenticité des textes, certains sont authentiques, par exemple, tous les articles extraits de journaux. Les textes contenant des figures de rhétorique sont authentiques mais adaptés, comme le suivant :

> *Read the following paragraph and explain the idiom (fixed expression) underlined* out of the context given:

You meet your friend Jack. You haven´t seen him for a while, but you have always been one of his best friends. Today Jack seems to be sad and very thoughtful. You ask him why he feels this way. Then Jack can´t keep things to himself any longer. He decides to bare his soul to you.

(Tâche 3, Compréhension de l´écrit, Niveau B2)

Une partie des textes est adossée à des manuels scolaires. Dans ce cas, ils sont soit authentiques et adaptés soit inventés. Un exemple de texte inventé est le suivant :

Jenny says the following about her sister:

"Samantha left school. She didn´t go to university. She joined the army."

Please use these sentences to form complex ones which make sense.

(Tâche B2, compréhension de l´oral, niveau B2)

Les textes issus de notre imagination, non inspirés par l'une des nombreuses sources non-officielles que nous avons consultées, sont entièrement fictifs.

You take a walk in your neighborhood and witness the following quarrel between two neighbours:
- Neighbour 1: "You have no right to cut the branches of my tree! The tree is growing on my land, not on yours!"
- Neighbour 2: "The branches are on my land, they reach my kitchen window!"
- Neighbour 1: "Ok, that´s enough, Mr.!! I´m going to complain to the police! You are destroying my property!"

(Tâche 2, Compréhension de l´oral, Niveau B2).

Les énoncés isolés figurant dans les choix multiples sont soit authentiques et adaptés soit fictifs (pour majorité). Un bon exemple de phrase authentique adaptée est [29] :

A. Although Samantha left school, she didn´t go to university, but to the army.
B. Since Samantha left school, she didn´t go to university even though she joined the army.
C. While Samantha left school, so she didn´t go to university, but joined the army.
D. After Samantha left school, she didn´t go to university, but instead she joined the army.

(Tâche B2, Compréhension de l´oral, Niveau B2)

Comme nous l'avons signalé, les compétences évaluées dans POSILANG sont la compréhension orale, la compréhension écrite et la production écrite

guidée. Chaque compétence est testée par un ensemble de tâches qui font partie d'un domaine de compétences à un niveau donné. Les formes de l'expression grammaticale (catégories, structures) sont précisées en s'appuyant sur la nomenclature des *Instructions Officielles* de l'enseignement secondaire français, par exemple, l'impératif, les prépositions ou les adjectifs possessifs. Le répertoire lexical n'est pas spécifié, mais conditionné par les thèmes traités, eux-mêmes rapportés à un niveau de compétences donné.

Chaque section contient entre cinq et six tâches qui sont pondérées de la même manière. L'attribution des scores ne dépend pas du niveau de difficulté des tâches dans POSILANG. La méthode utilisée dans le test entier est le QCM. Les instructions utilisées sont diverses et contiennent, en partie, les procédures à suivre et les types de tâches à accomplir. Dans une partie des instructions, en revanche, uniquement les types de tâches à effectuer sont indiqués. [30]

Les spécifications doivent également informer l'utilisateur des procédures d'attribution des scores. Celles-ci sont objectives dans notre test. L'attribution des scores objectifs prévient tout désaccord (Hughes 2003 : 62). Nous avons décidé d'attribuer un point pour chaque tâche dans chaque domaine de compétence. Il en est de même pour les tâches complexes nécessitant plusieurs réponses. Un point est attribué pour l'ensemble de la tâche et subdivisé selon le nombre de bonnes réponses demandées.

En ce qui concerne le seuillage des niveaux, il a été décidé d´établir un seuil de passation de 60%. Pour atteindre un certain niveau, il faut avoir accompli au moins 60% des tâches correctement, dans chaque domaine de compétence d'un niveau donné. Pour calculer le score total de chaque niveau, les points sont tout simplement additionnés, par domaine de compétence. Pour avoir le droit de continuer le test et donc de passer au niveau supérieur, il faut avoir acquis 60 % du niveau concerné. Pour valider le niveau A1, dans POSILANG, il faut ainsi avoir obtenu un score minimum de 11 points, correspondant à 60 % du score total qui s'élève à 18 (3x6) points. On peut alors passer au niveau A2, mais pour valider ce dernier, il faut à nouveau atteindre 60 % correspondant à un score minimum de 22 sur 36. Au-delà du calcul du score global, il est prévu d'évaluer le

niveau dans chaque domaine de compétences séparément et de fournir un feedback aux candidats sur ce point.

3.6.4 Etape 2 : conception de la maquette

Après l'énoncé des spécifications, la création de la maquette du test doit constituer l'étape suivante. Celle-ci implique la sélection et la mise en ordre des tâches (Brown 2010 : 110). Avant de pouvoir créer ou sélectionner les tâches, il faut prendre position sur leur contenu (McNamara 2000 : 25). Cela constitue la première étape pratique de la conception d'un test. Cette étape présuppose l'élaboration préalable du « construit » qui renvoie aux connaissances ou aux compétences évaluées par le dispositif (McNamara 2000 : 13).

Pour pouvoir déterminer le contenu d'un test, il faut avoir un point de vue particulier sur l'usage de la langue dans ce dispositif, ainsi que sur la relation entre la performance au test et les contextes réels de l'usage de la langue (McNamara 2000 : 25). Le contenu du test peut être défini de deux manières et dépend de la définition du construit. La première façon de définir ce construit est opérationnelle, en tant qu'ensemble de tâches pratiques et réelles. Dans ce cas, établir le contenu d'un test implique qu'on choisisse les tâches les plus représentatives du domaine testé, en fonction de leur fréquence ou de leur importance (McNamara 2000 : 25). [31]

Une autre manière de définir le contenu d'un test est abstraite**.** On prend appui sur une théorie de la connaissance ou des capacités dans le domaine langagier (McNamarra 2005 : 25). Les domaines de connaissance sont le système grammatical, le lexique ou la prononciation. Les capacités sont celles requises pour effectuer les activités de communication en réception et en production. On raisonne à partir d'échantillons de structures grammaticales ou d'items lexicaux associés à tel ou tel niveau, ou encore on raisonne à partir d'activités de communication considérées comme pertinentes (McNamara 2000 : 26).

En ce qui concerne le processus de conception des types d'items, il varie beaucoup dans sa complexité d'un test à l'autre, en fonction du type de test et de ses objectifs. Cependant, il n'existe qu'un répertoire limité de stimuli et de

réponses (Brown 2010 : 61-62). Par ailleurs, il existe une correspondance entre les modes de stimulus et les modes de réponse, de sorte que seulement une partie des modes de réponse convienne à chaque mode de stimulus (Brown 2010 : 61-62).

Il est important de comprendre la nature et les principes de construction des items pour construire de nouveaux tests ou procéder à des mises à jour (lorsqu'on décide d'adapter ou de faire évoluer des dispositifs déjà construits). Entrent en jeu la justesse des formules langagières utilisées, le choix du niveau approprié de difficulté et le pouvoir de discrimination des items. Il y a toujours de bonnes raisons à vouloir saisir les principes de construction des items, mais les objectifs principaux restent l'amélioration de la validité et la réduction de la longueur (Belanger 2002 : 4). Il peut aussi être nécessaire de comprendre la nature et les principes de conception des items pour des raisons qui ne sont pas liées aux tests. Les raisons envisageables de le faire sont l'évaluation des difficultés d'apprentissage, la modification ou l'amélioration de l'enseignement (Belanger 2002 : 4). Il faut noter que pour comprendre la nature et les principes de construction des items, il est indispensable d'analyser ces derniers.

3.6.4.1 Les directives pour la conception des items à choix multiples

POSILANG est un test qui fait appel au questionnaire au choix multiple (QCM). Il est donc important que nous prêtions attention à ce qui fait la spécificité de ce format**.** Les items au format QCM peuvent avoir plusieurs formes, mais tous se composent d'une base, également désignée input ou stimulus, qui appelle une réponse (Purpura 2001 : 1). La réponse doit être choisie parmi un éventail de propositions. Le plus souvent, une seule de ces réponses possibles est la bonne. On la désigne par le terme « clé », les autres renvoyant à des choix incorrects (Purpura 2001 : 1).

Cette structure fondamentale étant établie, plusieurs réalisations sont possibles. On peut partir d'une phrase, qu'il faut soit remplir soit compléter. On peut également être amené à compléter un dialogue. Enfin, on peut demander que des erreurs soient identifiées dans un stimulus. Les formes potentiellement erronées sont signalées par des lettres différentes (Purpura 2001 : 1). Cette

dernière déclinaison du format QCM est beaucoup plus rare que les trois autres. On peut citer:

Cities have been suffering of air pollution since the industrial revolution.

[a] [b] [c]

Now is it getting better? No error. (Purpura 2001: 1)

[d] [e]

Bien qu'il puisse paraitre simple de construire des tests contenant des items au format QCM, ils sont en fait très délicats à construire correctement. Les limites de ce format n'ont été reconnues que récemment (Cohen 1998 : 98). Parmi les points faibles de leur conception figurent l'existence de plusieurs réponses correctes ou bien d'aucune, l'inefficacité des options incorrectes et la présence d'indices de la bonne réponse dans les options (Hughes 2003 : 76). L'inefficacité des options incorrectes peut résulter de leur absurdité, ce qui mène à leur élimination automatique. Quant aux indices de bonne réponse dans les options, ils peuvent être explicites, mais également implicites, car le choix de la réponse correcte peut faire partie de la connaissance du monde (Hughes 1989 : 60). La sélection de la réponse clé par élimination des options absurdes ou grâce à la connaissance du monde constitue un réel problème. L'item n'évalue plus les connaissances et les capacités langagières des candidats, mais uniquement leur sens commun et leur connaissance du monde. Pour cette raison, le test qui contient des items porteurs de ces caractéristiques ne satisfait pas au principe de validité (Hughes 1989 : 60). Un autre danger caractéristique des QCM est l'évaluation de plusieurs domaines de compétences par un seul item, par exemple, de la compétence grammaticale aussi bien que de la compétence lexicale. Eviter ces erreurs demande beaucoup de temps et d'expertise lors de la construction des items et implique la nécessité d'en élaborer un plus grand nombre que ceux qui seront retenus lors de la procédure de validation (Hughes 2003 : 76).

Au-delà des problèmes structurels possibles, les items au format QCM présentent d'autres inconvénients (Hughes 2003 : 76). Ce format peut mener à des problèmes lors de l'administration du test, lors de l'évaluation et lors de

l'apprentissage. Il a été constaté que les candidats s'arrêtaient de lire quand ils ont l´impression d´avoir trouvé la bonne réponse. En outre, les items ont tendance à être lus de façon hâtive et superficielle (Cohen 1998 : 98). En ce qui concerne la phase d'administration, le test au format QCM encourage une tendance à deviner et à frauder chez les candidats (Hughes 2003 : 76-77). En effet, cette technique permet de répondre correctement à un certain pourcentage d'items, sans nécessairement avoir la connaissance ou la capacité testée (Hughes 2003 : 76). Puisque le pourcentage de bonnes réponses s'élève à 33% en cas de présence des trois options, il est indispensable de fournir au moins quatre alternatives, afin de diminuer l'effet de cette possibilité. En outre, il faut s'assurer que les options contenues dans un item aient toutes été choisies par un certain nombre de candidats, car dans le cas contraire, les options incorrectes ne rempliront pas leur fonction, qui est de distraire le candidat (Hughes 2003 : 77).

En outre, le format QCM tend à limiter le champ d'évaluation (Hughes 2003 : 77). Pour pouvoir tester un aspect langagier, il faut trouver les options plausibles. Or, celles-ci ne sont pas toujours disponibles. Ce problème concerne notamment l'évaluation de certaines structures grammaticales, par exemple, la distinction entre *simple past* et *present perfect* en anglais (Hughes 2003 : 77). Le deuxième problème est que ce format évalue seulement les capacités des candidats en posture de réception. Si ces capacités divergent des capacités de production individuelles, la performance au test ne reflète pas les capacités véritables. L'interprétation qui peut être faite de la performance n'est alors plus valable (Hughes 2003 : 77). Si les phénomènes langagiers évalués par un test ne sont pas à la base de leur usage productif, le test manque de validité de construit (Hughes 2003 : 77). En ce qui concerne l'enseignement et l'apprentissage, le format QCM risque d'avoir un impact négatif sur les pratiques (Hughes 2003 : 78). L'usage des tests à ce format réduit parfois le répertoire des activités pédagogiques proposées à l'entraînement (Hughes 2003 : 78).

Les limites et dérives que nous venons d'évoquer ne doivent pas masquer les atouts de ce format qui sont non moins réelles. Le QCM respecte des principes de praticité et de fiabilité. Les réponses correctes sont prédéterminées. La procédure d'évaluation est homogène, économique et cohérente (Hughes 2003 : 78). Pour cette raison, il faut rapporter les bénéfices de ce format aux

efforts et au temps de conception nécessaires (Brown 2010 : 67). En fonction des points forts et des défauts identifiés, il convient de déterminer si le format QCM est vraiment la meilleure option dans tel ou tel contexte d'évaluation particulier. A cause de la praticité et de la fiabilité élevée des questions à choix multiple, leur usage est indéniablement utile dans les tests standardisés (Brown 2010 : 67). Ce format est notamment conseillé pour des tests destinés à usage peu fréquent, par un grand nombre de candidats (Hughes 2003 : 78).

Bien qu'une réflexion méthodologique sérieuse soit souhaitable pour tous les dispositifs d'évaluation, la nécessité s'avère primordiale pour les tests au format QCM (Alderson 2004 : 10). En effet, les options proposées dans les réponses font ici partie du texte à traiter par les candidats (Alderson 2004 : 10). Le niveau de difficulté d'un item au format QCM ne dépend pas uniquement du stimulus, mais également des options proposées (Alderson 2004 : 10). Les facteurs déterminant le niveau de difficulté sont nombreux. On y trouve notamment l'élaboration et le contenu des options, l'usage du lexique, l'ordre des options proposées, le nombre d'informations abordées dans le stimulus reprises dans les options (Alderson 2004 : 10). La prise en compte de ces facteurs est très importante lors de l'évaluation du niveau de difficulté des items (Alderson 2004 : 10).

Pour éviter les défauts communs du format, il convient de suivre plusieurs lignes directrices, valables pour la conception de toutes les questions à choix multiple (Purpura 2001 : 1-3). Les lignes directrices ci-après doivent impérativement être respectées. Premièrement, un item doit viser un seul objectif : tous les choix proposés ne doivent tester qu'un seul phénomène langagier (Purpura 2001 : 2). Cette règle concerne non seulement les phénomènes qui font partie des différentes compétences langagières, mais également les structures révélatrices d'une seule compétence, par exemple, de la compétence grammaticale. Ainsi, il est interdit d'évaluer les pronoms et les auxiliaires dans un seul item, bien qu'ils relèvent tous les deux de la syntaxe (Purpura 2001 : 2).

Deuxièmement, le stimulus et les options doivent être formulés de façon aussi simple et directe que possible (Brown 2010 : 69). Cela implique,

premièrement, d'éviter les phrases trop longues qui n'aident pas à choisir la réponse correcte, et deuxièmement, d'éviter la répétition dans les options d'un mot déjà présent dans le stimulus (Brown 2010 : 69). Les instructions doivent également être aussi claires et précises que possible. Les items délicats sont à écarter car les locuteurs natifs de la langue ne doivent pas avoir de difficulté à répondre (Purpura 2001 : 2-3).

Troisièmement, il faut veiller à ce que la réponse correcte prévue soit la seule qui soit bonne. Les autres options ne doivent pas être recevables (Brown 2010 : 69). Par ailleurs, et comme nous l'avons déjà signalé, il faut prendre garde à ne pas faciliter le choix de la réponse clé en jouant sur certains facteurs comme la cohérence ou la longueur (Purpura 2001 : 4). En quatrième lieu, les options fournies doivent remplir d'autres critères sur le plan du contenu que celui d'être plausibles. Tous les contenus consignés dans les spécifications du test doivent être exploités. Ainsi, pour un test de positionnement, les items doivent renvoyer à l'ensemble du programme et être disposés dans un ordre de difficulté croissante (Purpura 2001 : 4). Le contenu des items doit être adapté à l'âge des candidats et éviter les biais (Purpura 2001 : 2). Rappelons que l'absence de biais est une qualité indispensable pour l'élaboration d'un dispositif équitable (Kunnan 2010 : 3), comme nous l'avons vu dans notre premier chapitre.

En ce qui concerne les instruments d'évaluation standardisés, à référence normative, il est obligatoire de comparer les items selon trois indices, qui sont leur degré de facilité ou de difficulté, leur index de discrimination et l'efficience des options fausses (Brown 2010 : 70). Cette ligne directrice est cependant facultative pour les autres catégories de tests (Brown 2010 : 70). Le degré de facilité est déterminé par le pourcentage d`étudiants ayant répondu correctement à un item. Cette valeur permet d'apprécier si un item est facile ou difficile pour un groupe de candidats. L'index de discrimination d'un item concerne le pouvoir que possède un item de différencier les candidats de niveau supérieur des candidats de niveau inférieur (Brown 2010 : 71).

La position des chercheurs vis-à-vis des items au format QCM varie. Certains se montrent très critiques, tandis que d'autres, au contraire, en valident l'utilité (Fulcher & Davidson 2007 : 63). Les critiques les plus vives proviennent

des partisans de l'approche communicative. Ces derniers estiment qu'une évaluation doit proposer des tâches reflétant l'usage de la langue en situation réelle (Fulcher & Davidson 2007 : 63). Ce point de vue est exposé dans l'ouvrage de Noël-Jothy & Sampsonis (2006) qui dénoncent les QCM, estimant que la seule compétence évaluée à travers ce format est la capacité à utiliser le format lui-même (Noël-Jothy & Sampsonis 2006 : 44). En d'autres termes, le QCM n'évaluerait pas les compétences langagières. Les chercheurs poursuivent leur critique en affirmant que les QCM favorisent les candidats qui y sont habitués (Noël-Jothy & Sampsonis 2006 : 44). La position de ce groupe de chercheurs paraît logique si on tient compte de leur seule acceptation des tâches authentiques, c'est-à-dire, de celles imitant l'usage réel de la langue (Noël- Jothy & Sampsonis 2006 : 44).

D'autres chercheurs n'attendent pas des tâches proposées dans un test qu'elles soient authentiques. Ils ne dénoncent donc pas les QCM en tant que tels, sur la seule base de leur artificialité (Fulcher & Davidson 2007 : 63). Ils réfutent l'argument selon lequel ces items seraient non probants uniquement parce qu'ils feraient fi de l'usage réel de la langue (Fulcher & Davidson 2007 : 63). Cela ne signifie pas pour autant que le format QCM soit automatiquement validé. Seule est proclamée la nécessité de se fonder sur une analyse rationnelle des réponses pour apprécier la performance (Fulcher & Davidson 2007 : 63).

3.6.5 Etape 3 : validation des tâches

Une fois la maquette conçue, il est essentiel de valider les tâches. La validation est un processus de collecte de données qui permet de démontrer le lien entre effectuation et compétence (Douglas & Hegelheimer 2008 : 123). Le processus de validation permet de vérifier si la nature et la difficulté des tâches correspondent bien au niveau de référence présumé. La validation des tâches est une procédure complexe car l'existence d'une relation entre le test et le CECRL n'est pas un fait directement observable, mais relève d'une affirmation pour laquelle le concepteur d'examen devra apporter des preuves, aussi bien au niveau théorique que pratique (Conseil de l'Europe 2009 : 7). La procédure par laquelle on obtient ces preuves est désignée «validation of the claim » dans la version originale du *Manual*, à traduire par la «validation de l'affirmation »

(Conseil de l'Europe 2009 : 7). Ainsi, l'attribution d'un des six niveaux de compétences du CECRL aux résultats obtenus constitue une affirmation qui a besoin d'être validée (Douglas & Hegelheimer 2008 : 123).

Le processus de validation des items doit obligatoirement comporter deux étapes distinctes et successives qui sont l'estimation du niveau des items et leur pilotage. L'estimation du niveau des items sert à leur validation rationnelle (Alderson, Clapham & Wall 1995 : 171).[32] En revanche, le pré-testing des items, également désigné pilotage, succédant à l'estimation de leur niveau commun de référence, est effectué afin d'opérer une validation empirique. Ce procédé sert à vérifier que le niveau de difficulté des tâches correspond bien au niveau de référence présumé (Alderson 2004 : 13). Ces deux étapes, l'estimation du niveau de référence des items et leur pilotage, constituent un procédé de validation des items qui est incontournable, comme cela a été souligné par l'équipe travaillant sur la conception de la grille :

> An important component of the linking process is to characterize the content of test items and tasks, and the Project Team believes that this process can be facilitated by completing this Grid. However, an equally important component is the analysis of the empirical results of the test on suitable samples of the intended population. Only once the results of the content analysis have been combined with the standards set by a standard setting process as described in the Manual based on the empirical results of the test, can a claimed link to the CEFR be explored.[33]

Les deux étapes qui servent à valider les items ne doivent pas s'effectuer de manière isolée, mais s'inscrire parmi d'autres procédures (Alderson 2004 : 13). L'estimation du niveau des items doit être précédée par la description des textes et des items. Des grilles d'analyse de contenu des items de compréhension écrite et orale ont été développées par un projet de recherche néerlandais.[34] Le pilotage doit être suivi par les trois étapes suivantes, constitutives du processus d'élaboration des items, qui sont le calibrage, la fixation des standards ainsi que l'attribution d'un niveau définitif aux items (Alderson 2004 : 13).

En principe, il n'y a pas de différence entre la procédure de validation des tests informatisés et des tests sur support papier. Dans un cas comme dans l'autre, la validation nécessite l'apport de preuves pour corroborer la performance individuelle au test. Cependant, l'usage de tests automatisés a été l'objet de méfiance dès le début, la crainte étant que ces dispositifs contiennent des

défauts propres aux technologies numériques (Douglas & Hegelheimer 2008 : 123). Six dangers ont ainsi été évoqués par Chapelle & Douglas (2006 : 41). Le premier risque pour la validité concerne la différence en matière de capacité mesurée par les tests automatisés par rapport aux dispositifs administrés sur papier (Chapelle & Douglas 2006 : 41). Les variations en termes de capacité ont été effectivement repérées par la recherche empirique, mais elles se sont avérées minimes. En revanche, il a bien été démontré que la familiarité avec les technologies numériques pouvait affecter la performance au test de manière significative (Douglas & Hegelheimer 2008 : 123).

Le deuxième risque que l'évaluation assistée par ordinateur fait peser sur la validité concerne la variété des items. Les types d'items envisageables ne sont pas les mêmes (Chapelle & Douglas 2006 : 41). Certains chercheurs soulignent les limites imposées par les technologies numériques sur les types de tâches possibles (Douglas & Hegelheimer 2008 : 125). Les limites sont manifestes, premièrement, au niveau de l'imitation de l'usage réel de la langue et deuxièmement, au niveau de l'arrangement graphique des tâches. Concernant la première restriction, l'ordinateur peut traiter seulement certains types des tâches, malgré le potentiel élevé de ce medium pour présenter les tâches issues de l'usage réel de la langue (Douglas & Hegelheimer 2008 : 125). Par exemple, il n'a pas encore été possible de développer un logiciel qui permette une simulation du langage vraiment interactive, ce qui empêche l'évaluation de la capacité d'interaction (Douglas & Hegelheimer 2008 : 125). Concernant la disposition graphique des tâches, la restriction s'applique en raison de l'espace réduit sur l'interface de l'ordinateur par rapport à celui de la page au format papier (Douglas & Hegelheimer 2008 : 125). Par ailleurs, les limites affectant les types de tâches concevables sont bien réelles. Bien qu'elles soient considérées comme un danger potentiel pour la validité des tests informatisés, la limitation des tâches ne doit pas remettre en cause la validité des tests. Afin d'assurer cette qualité essentielle, il faut premièrement, en tant que concepteur et utilisateur de test, considérer la situation d'usage de la langue ciblée par le dispositif créé. Ensuite, il est nécessaire de déterminer si les types de tâches compatibles avec le format numérique permettent de tirer des conclusions valables à partir des scores et de prendre les décisions souhaitées (Douglas & Hegelheimer 2008 : 125).

Le troisième enjeu, jugé comme une grave menace pesant sur la validité des tests informatisés, concerne la sélection des tâches par un algorithme lorsque les tests sont adaptatifs (Chapelle & Douglas 2006 : 41). La première crainte est que le choix des tâches ne représente pas un échantillon satisfaisant du contenu du test et, deuxièmement, que ce choix puisse engendrer des états émotionnels négatifs chez les candidats comme l'anxiété (Douglas & Hegelheimer 2008 : 125). La menace est considérée comme élevée lorsque l'algorithme repose sur un seul paramètre, qui est le niveau de difficulté des tâches, ou alors lorsqu'il y a une évaluation séparée des différents domaines de compétences (Douglas & Hegelheimer 2008 : 125). Le risque pour la validité du test provient du fait que ces deux facteurs réduisent probablement le choix des tâches soumises à chaque candidat lors de la passation du test (Douglas & Hegelheimer 2008 : 125).

Le quatrième aspect considéré comme une menace pour la validité est lié à l'attribution automatisée des scores (Chapelle & Douglas 2006 : 41). Ceci suscite la crainte qu'une telle procédure n'attribue pas de points aux qualités de la réponse qui sont pourtant pertinentes pour la mesure du construit du test (Chapelle & Douglas 2006 : 41). La notation automatique est problématique pour les réponses partiellement correctes qui exigeraient normalement l'attribution de scores partiels. Ceci soulève deux questions, à savoir, la pondération de chaque réponse et la base de la décision (Douglas & Hegelheimer 2008 : 125). Pour cette raison, les réponses qui peuvent être classées en deux catégories tranchées, correctes ou fausses, sont nettement préférables lorsqu'on procède à une allocation automatisée de scores.

Le cinquième point jugé comme un risque pour la validité concerne les enjeux de sécurité, qui englobent l'identité des candidats, la perméabilité des tests et l'exposition des tâches dans les tests adaptatifs (Chapelle & Douglas 2006 : 41). Le risque de mémorisation est réel (Douglas & Hegelheimer 2008 : 126). Les risques liés à la sécurité des dispositifs automatisés sont généralement admis des chercheurs, mais ne compromettent pas pour autant leur validité (Douglas & Hegelheimer 2008 : 126). Pour évaluer l'utilité d'usage des tests automatisés, le risque lié à la sécurité doit être comparé aux bénéfices de leur utilisation dans un contexte d'évaluation donné, compte tenu de l'objectif d'évaluation particulier (Douglas & Hegelheimer 2008 : 126).

Enfin, l'usage de tests informatisés peut avoir des conséquences néfastes pour les apprenants, les programmes d'apprentissage et la société. Les coûts d'implémentation sont souvent élevés (Chapelle & Douglas 2006 : 41). Par ailleurs, les tests informatisés peuvent avoir un impact négatif sur les pratiques d'enseignement et d'apprentissage. Il est tentant de se focaliser sur les connaissances et les capacités directement évaluées par le test, plutôt que d'effectuer un travail plus large sur les activités d'expression (Douglas & Hegelheimer 2008 : 127). L'impact négatif sur l'enseignement et l'apprentissage d'une langue n'est pourtant pas propre aux tests informatisés, puisque tous les tests peuvent avoir un effet négatif sur les pratiques pédagogiques, réduisant le nombre de connaissances et de capacités ciblées à celles évaluées par le test.

Malgré les risques dont sont porteurs les tests informatisés, un degré satisfaisant de validité reste atteignable. Comme cela se fait pour les tests papier, il est possible de procéder à la validation en deux étapes de tous les items.

3.6.6 Validation rationnelle des tâches

La validation rationnelle des items est effectuée moyennant l'estimation de leur niveau (Alderson 2004 : 13). Celle-ci nécessite d'établir un lien entre chaque tâche constitutive d'un test et les niveaux communs de référence, en s'appuyant sur deux instruments qui constituent une « nouvelle génération » de documents élaborés à partir du CECRL (Conseil de l'Europe 2009 : 4). Le premier de ces instruments est le *Manuel* pour *Relier les examens de langues au Cadre européen commun de référence pour les langues (CECR)* qui énumère et explique les différentes techniques utilisées à cet usage. Le deuxième référentiel est constitué par le système de classification fourni par les grilles d'analyse du contenu des items de compréhension écrite et orale (Alderson 2004 : 13).

En ce qui concerne le Manuel (2009), un grand nombre de tests ont été reliés aux niveaux communs de référence ce qui a permis de réviser et de compléter la version pilote du manuel (Conseil de l'Europe 2009 : 5).[35] Ce document décrit cinq étapes liées qui sont à respecter afin d'établir un lien entre un test et un niveau commun de référence (Conseil de l'Europe 2009 : 9). Les procédures en question sont la familiarisation, la spécification, la familiarisation

avec les normes, la mise aux normes et finalement, la validation. Bien que ce soient les étapes mutuellement liées, elles suivent un ordre logique (Conseil de l'Europe 2009 : 9). En raison du lien logique entre ces étapes, celles-ci sont également placées dans l'ordre chronologique, à l'exception des activités de validation. En effet, la validation gagne à être considérée comme un processus continu d'assurance de qualité qui vise à vérifier si les procédures précédentes ont permis d'atteindre les buts respectifs (Conseil de l'Europe 2009 : 11). Il ne faut pas confondre les activités de validation avec la validité d'un test, qui est une qualité essentielle de chaque dispositif, expliquée dans le premier chapitre de cette thèse.

La validité est le concept central lors du processus d'établissement du lien entre un test et le *Cadre européen commun de référence* (Conseil de l'Europe 2009 : 9). Avant de suivre les étapes visées et de relier un test au CECRL, il faut s'assurer de la validité interne du dispositif et de la fiabilité qui y est étroitement associée. Les normes internes au test seront opérationnalisées de façon cohérente uniquement si cette validité est acquise. Dans le cas contraire, l'établissement du lien d'un test à un niveau commun de référence ne pourra pas être validé en raison de l'incohérence des normes internes (Conseil de l'Europe 2009 : 9).

Chacune des cinq étapes évoquées comporte une série de procédures, qui doivent être choisies en fonction de leur pertinence ou de leur adéquation au contexte local. Ce choix doit être fait par le concepteur du test ou de l'examen (Conseil de l'Europe 2009 : 13). Le manuel attire l'attention de ses utilisateurs sur la nécessité de choisir et d'appliquer les techniques relevant de toutes les étapes inclues dans le processus de validation. Ceci est essentiel afin d'apporter les preuves d'un lien existant entre un test ou un examen et un niveau commun de référence.

3.6.6.1 Validation rationnelle des tâches de POSILANG

Dans le cas de POSILANG, l'estimation du niveau de référence des tâches a été effectuée à l'aide des grilles d'analyse de contenu des items de compréhension écrite et orale développées dans le cadre d'un projet néerlandais. Comme prévu par l'équipe d'élaboration de cet instrument de classification, l'estimation du

niveau de référence des tâches est précédée par une spécification de leurs caractéristiques. Il est prévu de ne fournir ici qu'un échantillon d'illustrations, au travers d'une sélection de tâches issues de POSILANG. Chaque niveau commun de référence et chaque domaine de compétence évalué sera ainsi illustré par une tâche [36].

Par exemple, la première tâche du domaine « Compréhension orale » du niveau A1 est la suivante :

At the zoo. The zoo-keeper shouts:

« Don't feed the animals! "
What did you hear?

A. Don't need the animals
B. Don't feed the animals
C. Don't nourish the animals
D. Don't lead the animals.

La tâche contient seulement du vocabulaire fréquent, des structures grammaticales simples et du contenu concret. L´articulation est claire et la vitesse lente ce qui contribue à l´attribution du niveau A1. Comme dans tout le test, l´item proposé est au format QCM. L'opération cognitive à effectuer consiste à reconnaître l'intention communicative à partir de l'information explicite.[37] La question suivante fait partie de la compréhension écrite de niveau A1 :

« Benny arrives at Heathrow after a week in Spain. He is waiting for his friend´s mother to fetch him.

1. Max´s mother can´t meet Benny at the airport…
2. There are taxis from the airport to the city centre…
3. If Benny has problems at the airport…
4. There is only one tube line from Heathrow...

A. ...but they are expensive
B. ...he can go to an information desk
C. ...because she works until 6 o`clock
D. ... and that is the Picadilly Line

Cette tâche est également de niveau A1, comme elle parle d'un thème courant et concret, d'un voyage. Le vocabulaire utilisé est fréquent et les structures grammaticales simples. L´item est au format appariement. Il requiert une identification des liens sémantiques entre parties pour former des phrases qui font un sens.

La tâche ci-après fait partie du domaine « Expression de l´écrit » en A1.

Which word best describes the children's « work »?
In this family, the children must tidy up their rooms, lay the table and stack the dishwasher.

A. Homework
B. Housework
C. Guesswork
D. Teamwork

Comme la tâche précédente, elle traite d´un contenu courant et concret: la maison et le ménage. Le vocabulaire et les structures utilisés sont simples. L´opération demandée dans l´item est de reconnaitre l´idée principale d'une information explicite. Concernant le niveau A2, un bon exemple de tâche centrée sur la compréhension de l´oral est le suivant :

"You *will hear a question and four answers. Select the correct answer to this question. What are you going to do on the weekend? "*

A. I ski in the mountains.
B. I want stay at home.
C. I am going to visit my aunt in Sheffield.
D. I am planning travel to London.

Le contenu de cette tâche est concret et le sujet courant : il s'agit de projets pour le weekend. La tâche contient, comme celles de niveau A1, des structures simples et du vocabulaire fréquent. Comme en A 1, l'accent est standard. La différence entre A1 et A2 est le phénomène grammatical évalué, qui est ici le *go-future*. Les expressions de l´avenir sont à connaitre au niveau A2 (CNDP 2005 : 23).

Une tâche qui évalue la compréhension de l´écrit au niveau A2 est par exemple:

You read the following headline in the daily newspaper:

"The annual <u>divorce</u> rate is rising in Britain."

Which of the following words has a similar meaning as the underlined one?

A. illness

B. separation

C. unemployment

D. social conflicts

La tâche donnée parle des relations entre les gens, ce qui renvoie à un thème courant. De même, un vocabulaire et des structures simples sont employés. La différence est qu´un texte authentique apparait dans cette tâche. Comme il s´agit du niveau A2, un texte authentique (par opposition à fabriqué) est adapté.

Dans le domaine « Expression de l´écrit » on trouve la tâche suivante :

Make opposite pairs. Find the words with an opposite meaning to the underlined ones.
I like people who are strong and kind.

1. strong: 2.kind:

A. Interesting, A. Interesting,
B. Glad, B. Glad,
C. Weak C. Weak,
D. Nasty D. Nasty

Cette tâche contient des mots fréquents mais pas autant qu'au niveau A1. De plus, le contenu exprimé est relativement abstrait. Ces deux aspects font la différence avec A1.

La tâche citée ci-dessous se trouve dans la compréhension de l´oral au niveau B1:

Listen to Jean´s statement about his professional life:

Jean used to work in a company which produces motor vehicle tires.
What does this sentence express?

A. Jean still works in the company.
B. Jean produces motor vehicle tires.
C. Jean worked in that company.
D. Jean is going to work for the company.

Contrairement à ce qui se passe au niveau précédent, cette tâche B1 contient un vocabulaire assez étendu. La phrase inclut une structure complexe qui est une proposition relative.

La tâche suivante fait partie du domaine « Compréhension de l´écrit » au niveau B1 :

You watch a Chart show on TV and the following idea comes to your mind:
"There are few singers_____ names I´ve never heard before."

Select the correct pronoun for the gap.
A. who´s

B. that
C. who
D. whose

Du vocabulaire courant y figure. Toutefois, on note la présence de plusieurs structures grammaticales complexes: la première phrase est composée (elle inclut une coordonnée) "You watch a Chart show on TV and the following idea comes to your mind". La deuxième contient une proposition relative en *whose* qui est déjà d'un certain niveau de difficulté.

La tâche finale de niveau B1 en «Expression écrite» est la suivante:

Anna tells her friend Ben what she thinks about her parents.
Choose the correct reflexive pronoun from those below:

-Anna: "My mum always does so much for me. That's really incredible!
-Ben: "How do you mean it?"
-Anna: "I mean parents sometimes also need to have a good time and enjoy _____.

A. ourselves
B. theirselves
C. themselves
D. himselves

Cette tâche contient plusieurs structures complexes. La première est une interrogative indirecte: « what she thinks about her parents. » La deuxième structure complexe est une proposition subordonnée dans la troisième phrase, qui n'est pas introduite par le pronom « that » : « parents sometimes also need to have a good time and enjoy… » . De plus, cette tâche contient un verbe réfléchi « to enjoy oneself » qui est un indicateur de niveau B1.

En ce qui concerne le niveau B2, nous proposons comme illustration de tâche de « Compréhension orale » :

Your mom tells you the following proverb after you report about your new job:
"(A) burden of one's own choice is not felt.
What does the proverb mean?

A. You work hard even if you choose what you like.
B. It´s not always good to make your own choices.
C. To make one´s own choice is a burden.
D. Something difficult seems easier when it is done voluntarily.

Contrairement à ce qui se passe en B1, l´input dans cette tâche inclut un contenu abstrait. Il s´agit d´un proverbe dont il faut comprendre le sens. En outre, le vocabulaire est relativement étendu. S'agissant d´un proverbe, l´information

est transmise de manière implicite. Par conséquent, l´item demande à reconnaitre l´intention communicative à partir de l´information implicite.

Une tâche choisie dans le domaine « Compréhension écrite » de niveau B2 est celle-ci :

You read the following article in the newspaper commemorating Alexander Fleming´s birthday:

> Only 80 years ago, it was not unusual for many children to die before adulthood. One major reason for this was that there were no cures for common bacterial illnesses such as pneumonia, diphtheria or scarlet fever. Thanks to biologists like Scotsman Alexander Fleming (1881-1955), whose discovery of penicillin helped to launch the field of antibiotic medicine, contracting bacterial illnesses was no longer a death sentence-at least for those who had access to good medical treatment.

Which of the following statements is <u>false?</u>

A. Alexander Fleming was a biologist.
B. Alexander Fleming´s invention of penicillin was a starting point for research on antibiotics.
C. A disease like pneumonia was a common cause of death.
D. Fleming discovered the antibiotic Penicillin more than 200 years ago.

Cette tâche constitue la première tâche complètement authentique. Elle est construite à partir d'un extrait d'un article de presse. Le texte sélectionné contient de nombreuses structures grammaticales complexes: plusieurs propositions relatives et un gérondif. De plus, le vocabulaire est étendu. Ces deux aspects sont des indicateurs de niveau B2. Ici, il faut comprendre les détails pour être capable de répondre correctement.

La tâche suivante évalue l´expression de l´écrit en B2:

You find the following text in your university paper:
After all, by many measures, there´s never been a better time to be a woman. In places like Scandinavia and Britain, a third or more corporate managers are now women. Latin America has seen a 50 percent jump in the number of women politicians in the last decade. However, in the EU Parliament, only 23 out of 162 members are female. Even though more than 50 percent of students in higher education in many parts of the world are women, the representation of women in corporate leadership has been stagnant for the last few years.
Please choose the statement which best explains the word "corporate":

A. Corporate designates commercial companies.
B. Corporate designates nonprofit organizations.
C. Corporate designates political leadership.

D. Corporate designates higher education.

Comme dans les autres tâches de niveau B2, il s´agit d´un texte authentique, qui provient d´un journal anglophone. Le texte se distingue par un vocabulaire large et des structures complexes.

Une illustration de tâche rattachée au domaine de la « Compréhension orale » de niveau C 1 est la suivante :

You are listening to a radio programme about theatres in capitals world-wide:

The Mecca of British theatre-goers is London, which boasts more than 200 professional theatres, many of which are concentrated in the West End and on the South Bank. The centre of the theatre scene in the US is Broadway in New York. New stage productions are presented every year, many of them costing several million pounds or dollars. Production costs however, are not a guarantee of box-office success. Most large theatres are repertory theatres with a stock company of actors performing a repertoire of plays at these particular theatres only. Provincial theatres are mainly served by touring companies as they usually cannot afford their own repertory companies.
Which of the following statements is false?

A. London has more than 200 different theatres.
B. Production costs are normally a guarantee for box office success.
C. Provincial theatres don´t have the funds for their own repertory companies.
C. London theatres are mostly found in the West End and the South Bank.

Comme on le constate, cette tâche C1 est nettement plus longue que celle de niveau inférieur. En plus, un large choix de vocabulaire est utilisé ainsi qu'un certain nombre de structures complexes: plusieurs subordonnées relatives et causales, plusieurs constructions au gérondif ainsi que des structures au passif. Pour être capable de répondre à la question, il faut comprendre les détails du texte.

Une tâche qui cible la compréhension de l´écrit au niveau C1 est celle reproduite ci-après.

Sarah reads in her history book the chapter about American history.

The backbones of US democracy and the system of government are the Declaration of Independence (1776) and the Constitution of the USA (1787) with its seven original articles and 26 amendments.

What does the underlined word stand for?

A. memoires

B. spine

C. principle

D. idea

On note qu'ici le contenu est abstrait tout en empruntant un mot très concret (« backbone »). Le candidat qui répond correctement saisit qu'il s'agit d'une métaphore qui renvoie à l'idée de principe fondateur et structurant. A l'évidence, il s'agit d'une notion abstraite mais imagée. Le vocabulaire est étendu bien qu´il y ait plusieurs mots transparents.

Une tâche qui cible l´expression écrite au niveau C1 est par exemple:

You come across the following statement in a British newspaper:

Feeling abandoned and ignored the North of the UK accuses the government in London of being biased in favor of the South which is better off economically.

What does the underlined phrase want to express?

A. The government in London favors the North.
B. The government in London favors the South.
C. The government in London is irresolute.
D. The government in London is prejudiced against the North.

Tout comme au niveau B2, le texte est ici authentique. Son contenu est abstrait comme pour les autres tâches de niveau C1. Le vocabulaire est également large et idiomatique, comme le souligne l'expression « to be better off ». Les constructions grammaticales sont complexes, en raison de l'emploi du gérondif, du participe passé et d'une proposition relative.

3.6.7 Passation des items en amont de leur opérationnalisation

La passation des items en amont de l'administration opérationnelle du test, désignée « pre-testing » en anglais, a pour but de procéder à une première forme de validation empirique (Alderson, Clapham & Wall 1995 : 74). La validation empirique est une méthode de validation des items qui se distingue de la validation dite « rationnelle ». Toutefois, la division traditionnelle entre « validation empirique » et « validation rationnelle » est aujourd'hui remise en cause car les deux méthodes de validation intègrent en fait les deux dimensions (Alderson, Clapham & Wall 1995 : 171). Les chercheurs contemporains préfèrent

donc opérer une autre distinction, entre « validation interne » et « validation externe » (Alderson, Clapham & Wall 1995 : 171). La validation interne renvoie aux recherches sur le contenu d'un test et l'effet perçu du contenu. La validation externe consiste à comparer les scores obtenus au test avec une ou plusieurs mesures de compétence obtenues ailleurs (Alderson, Clapham & Wall 1995 : 171). Cette autre mesure de la compétence, qui sert de référence, doit être exprimée de manière empirique en dehors du test lui-même. La mesure peut être le score obtenu à une version du même test ou alors à un tout autre test. La mesure peut également renvoyer à une évaluation de la compétence du candidat réalisée par d'autres moyens, dans un autre cadre. Enfin, elle peut même être issue d'une auto-évaluation du candidat portant sur sa propre compétence en langue (Alderson, Clapham & Wall 1995 : 177-178).

Le pilotage est le seul moyen de vérifier si le contenu des tâches, leur niveau de difficulté et leur pouvoir discriminant correspondent bien au niveau de référence présumé. On relève de très grandes variations parmi les spécialistes concernant les paramètres à adopter (Alderson, Clapham & Wall 1995 : 73). Pour ce qui est des items au format QCM, l'estimation de la performance se révèle particulièrement délicate. L'interprétation des réponses correctes ou incorrectes est facteur d'ambiguïté et de désaccord (Alderson, Clapham & Wall 1995 : 73). Ce type d'items nécessite donc un pilotage plus intense que les items aux autres formats (Alderson, Clapham & Wall 1995 : 75). Mais ce qui est vrai des QCM l'est en réalité de tout type d'items (Alderson, Clapham & Wall 1995 : 75).

Bien qu'il soit souhaitable de valider tous les paramètres, il est au minimum nécessaire de valider le niveau de difficulté de chaque item (Laurier 1998 : 253). Il y a plusieurs raisons de vouloir contrôler ce niveau lors du pilotage. La première est la difficulté croissante des items au sein du dispositif. Pour que la séquence des tâches proposées soit correctement ordonnée, il faut que le niveau de difficulté des tâches ait été correctement estimé au préalable. Une autre raison est le gain en assurance des candidats, lorsqu'ils sont soumis à une progression graduelle. Cette assurance a un impact très positif sur la validité faciale d'un test [38] (Bélanger 2002 : 9).

La dimension de l'échantillon nécessaire dépend du nombre de paramètres qu'on cherche à valider. Une calibration avec un modèle à trois

paramètres peut requérir jusqu'à 500 candidats (Laurier 1998 : 253). Il faut noter que l'échantillon minimal des candidats participant au pilotage des items est normalement une centaine (Laurier 1998 : 253). Le niveau des items doit correspondre au niveau de compétence estimé des candidats. En d'autres termes, il faut soumettre des items faciles aux étudiants débutants et des items (très) difficiles aux étudiants (très) avancés (Laurier 1998 : 253). Comme nous le verrons plus loin, des items trop faciles, ou alors trop difficiles, fournissent peu d'information sur le niveau de difficulté et rendent le résultat éminemment prévisible (Laurier 1998 : 253). La technique habituelle consiste à expérimenter les items à partir d'une version papier-crayon (Laurier 1998 : 252).

Il faut souligner que la passation préalable d'un test n'est pas identique au « pilotage ». Ce dernier terme ne se réfère qu'à la première étape, en amont de l'usage opérationnel (Alderson, Clapham & Wall 1995 : 74). Aux côtés de « pilotage », le terme « pre-testing » englobe les principaux essais requis avant la mise en œuvre opérationnelle du dispositif (Alderson, Clapham & Wall 1995 : 74). Les deux étapes de passation du test avant son opérationnalisation sont décrites dans les pages qui suivent.

3.6.7.1 Le pilotage des items

Le pilotage des items d'un test a pour finalité de reconnaître et d'éliminer les problèmes majeurs avant les essais principaux du dispositif. Moins formelle que l'étape suivante, l'étape du pilotage se compose de deux phases. La première, relativement informelle, consiste à demander à quelques amis ou collègues de passer le test afin d'évaluer la clarté des consignes et du langage utilisés, ainsi que la pertinence des bonnes réponses (Alderson, Clapham & Wall 1995 : 74). La deuxième phase du pilotage consiste à faire passer la première version révisée du test à un groupe d'environ vingt candidats, proches des candidats cibles par leurs caractéristiques générales et par leur niveau de compétence langagière (Hughes 2003 : 64). Pour certains linguistes, les participants à cette phase de pilotage doivent être des locuteurs natifs de la langue (Hughes 2003 : 64). Cette étape du pilotage est en mesure de fournir des informations multiples sur tous les aspects importants liés à la passation d'un test : l'intelligibilité et la pertinence des options pour les items à réponse fixe, la clarté des consignes, la facilité d'administration du test, le temps de passation nécessaire et l'utilité de

l'échelle d'évaluation (Alderson, Clapham & Wall 1995 : 75). Cette analyse empirique se veut aussi objective que possible, mais selon North (2000) ne peut l'être entièrement car elle implique forcément des jugements et des prises de décisions engageant la subjectivité des concepteurs (North 2000 : 221).

3.6.7.2 Principaux essais de passation des items

Cette deuxième étape est tout aussi nécessaire que la précédente avant l'opérationnalisation d'un test. Son envergure et le type d'analyse exigé en aval varient toutefois d'un test à l'autre. L'importance et le but du test, le degré d'objectivité de la notation sont ici des facteurs déterminants (Alderson, Clapham & Wall 1995 : 75). Les items notés objectivement, notamment au format *QCM*, demandent davantage d'essais avant leur mise en œuvre opérationnelle que les items au format « réponse ouverte » (Alderson, Clapham & Wall 1995 : 75). Ici encore, les candidats participant aux essais doivent ressembler aux candidats cibles du test par leur niveau de compétence langagière et par leurs caractéristiques générales (Hughes 2003 : 64-65). En outre, il est indispensable que les candidats convoqués pour ces essais prennent le test au sérieux, même si le dispositif est encore en phase d'élaboration. Dans le cas contraire, la procédure de validation des items risque d'être invalidée (Alderson, Clapham & Wall 1995 : 76).

Les items doivent être présentés de la même manière que celle envisagée pour la passation du test final. Le seul aspect qui peut être modifié par rapport à la passation opérationnelle est la durée. Car pour estimer avec précision la fiabilité du test, la durée de temps nécessaire à l'effectuation des tâches doit être accordée aux candidats (Alderson, Clapham & Wall 1995 : 76).

Lorsqu'un test est à évaluation objective, il se construit autour d'une norme. Dans ce type de test, un score total peut être calculé. C'est ce score qui constitue la norme, par rapport à laquelle les résultats individuels seront appréciés. Les résultats des candidats sont situés sur une échelle ordonnée et comparés entre eux. Des critères aussi objectifs que possibles sont donc appliqués (Alderson, Clapham & Wall 1995 : 76-77). On parle alors de norme ou de référence « critériée » (Alderson, Clapham & Wall 1995 : 77).

3.6.8 Analyse des résultats du pilotage

Pour parvenir à valider les tâches, il est nécessaire d'effectuer une analyse méthodique (Laurier 1998 : 252). L'analyse des tâches est incontournable, peu importe qu'il s'agisse de validation interne ou externe. L'analyse, désignée également «calibration », vise à établir les «caractéristiques» ou les «paramètres» de chaque tâche (Laurier 1998 : 252). Il existe deux types d'analyse des tâches : l'analyse qualitative et l'analyse quantitative.

3.6.8.1 Analyse qualitative

L'analyse qualitative des tâches a trois fonctions. La première consiste à vérifier la validité de contenu d'une tâche, c'est-à-dire à déterminer dans quelle mesure le contenu de cette dernière est représentatif du domaine ou de la capacité à évaluer (Bélanger 2002 : 4) [39]. La deuxième fonction de l'analyse qualitative des tâches est la vérification de leur « rédaction adéquate», qui se traduit par un format approprié (Bélanger 2002 : 5). La troisième fonction de l'analyse qualitative est de repérer une mauvaise compréhension des tâches par les candidats ou de déceler des réponses inattendues mais recevables in fine, durant la phase de pilotage. Dans la mesure où la présence de réponses problématiques signale une mauvaise conception des items, il est indispensable d'intégrer l'examen des réponses à l'analyse qualitative des tâches. Au final, soit on décide de modifier ce qui s'avère problématique, soit on élimine purement et simplement la tâche (Hughes 2003 : 65).

De son côté, l'analyse quantitative des tâches a pour but d'établir la difficulté et le «pouvoir discriminant » des tâches (Bélanger 2002 : 5). Dans la mesure où l'analyse quantitative recourt à des méthodes statistiques, elle est également appelée analyse statistique (Hughes 2003 : 65). Ce type d'analyse est décrit plus loin.

3.6.8.2 Analyse quantitative des items

Il est d'usage de procéder à deux mesures par item : la première concerne sa valeur de facilité et la seconde, son index de discrimination (Alderson, Clapham & Wall 1995 : 82). Ces paramètres sont essentiels pour l'analyse des items et méritent donc d'être explicités. Il est entendu que le calcul n'est possible que si

les items sont notés de manière cohérente, qu'il s'agisse d'items à réponse fixe ou ouverte (Alderson, Clapham & Wall 1995 : 86). Lorsque le format est de type QCM, le calcul intègre bien évidemment les distracteurs.

3.6.8.3 La valeur de facilité des items

Cette valeur indique le niveau de difficulté d'un item en mesurant le pourcentage des candidats ayant fourni une bonne réponse (Alderson, Clapham & Wall 1995 : 82). La valeur de facilité d'un item est corrélée négativement avec son niveau de difficulté, car plus la valeur de facilité est élevée plus son niveau de difficulté est bas (Hughes 2003 : 225). Pour donner un exemple, une valeur de facilité de 50%, souvent notée sous forme de proportion .5 ou 0.5, indique que le taux de candidats ayant répondu correctement à cet item s'élève à 50% (Alderson, Clapham & Wall 1995 : 82). Les items très faciles ou au contraire très difficiles ont, dans les faits, un très faible pouvoir discriminant : presque tout le monde répond correctement ou alors presque personne. Ces items ne permettent donc pas de départager les candidats (Alderson, Clapham & Wall 1995 : 83). On voit donc qu'il existe un lien étroit entre la valeur de facilité d'un item et son index de discrimination. Cela n'implique pas pour autant qu'il faille supprimer les items avec une valeur de facilité très élevée ou très basse. La décision concernant le maintien ou la suppression de ces items doit dépendre de l'objectif du test, c'est-à-dire de l'usage auquel ce test est destiné (Hughes 2003 : 225). Dans le cas d'un test de positionnement couvrant plusieurs niveaux de compétences, destiné à répartir les candidats par groupes de compétences, il est judicieux d'élaborer des items avec une large palette de valeurs de facilité (Hughes 2003 : 225). C'est le cas de POSILANG, aussi avons-nous veillé à ce que l'éventail des valeurs de facilité soit aussi large que possible. En revanche, si nous avions conçu un test de compétence dans le seul but d'identifier les excellents étudiants, l'inclusion d'items à haute valeur de facilité n'aurait présenté absolument aucun intérêt (Hughes 2003 : 225). Cependant, quel que soit l'usage prévu d'un test, il convient d'interpréter la valeur de facilité d'un item avec la plus grande prudence car celle-ci est non seulement liée à la variation des compétences des candidats mais aussi aux disparités entre items au niveau de leur difficulté intrinsèque (Alderson & Huhta 2005 : 313). Etant déterminée par ces deux facteurs, la valeur de facilité

s'avère en fin de compte une notion plus évasive et complexe qu'elle ne le semble de prime abord.

3.6.8.4 L'index de discrimination des items

Cette variable concerne le pouvoir de discrimination d'un item, c'est-à-dire sa capacité à différencier les candidats de niveaux de compétence différents (Alderson, Clapham & Wall 1995 : 81). Puisque l'index de discrimination n'est rien d'autre qu'un coefficient de corrélation, pour calculer celui-ci on compare le score obtenu à l'item avec le résultat obtenu au test tout entier. A l'évidence, le résultat au test ne doit pas inclure le score obtenu à l'item faisant l'objet d'une analyse ciblée. L'index de discrimination d'un item est d'autant plus élevé que la corrélation entre les deux scores est importante (Hughes 2003 : 226). Dans le cadre de cette procédure, il est fréquent de comparer la performance à l'item en question aux autres parties du test qui évaluent la ou les mêmes compétences langagières. Ainsi, le résultat des candidats à un item particulier de compréhension de l'oral, par exemple, est souvent comparé à la performance obtenue aux items de compréhension de l'oral ailleurs. Pour cette raison, les candidats sont souvent répartis par groupes de niveaux, selon les compétences (Alderson, Clapham & Wall 1995 : 84). La comparaison des scores n'est cependant pas acceptée par tous les spécialistes. Certains considèrent comme illogique de comparer la performance des candidats à un item particulier avec les résultats obtenus au test tout entier, alors même que le test n'est pas stabilisé et fiabilisé (Alderson, Clapham & Wall 1995 : 84). Pour échapper à cette incohérence, il est possible de répartir les candidats en différents niveaux de compétences sur la base d'une mesure externe au test, par exemple, à partir d'évaluations subjectives opérées par un professeur (Alderson, Clapham & Wall 1995 : 84).

Une autre manière de calculer l'index de discrimination, qui a l'avantage d'être plus simple, consiste à comparer la proportion de réponses correctes obtenue par le tiers le plus fort des candidats avec le pourcentage de bonnes réponses obtenu par le tiers le plus faible (Alderson, Clapham & Wall 1995 : 84). Pour donner un exemple, si sept candidats sur dix, dans le meilleur groupe, ont fourni une bonne réponse et seulement deux sur dix dans le groupe le plus faible, l'index de discrimination sera calculé de la manière suivante : .7-.2 = +.5

(Alderson, Clapham & Wall 1995 : 84). L'index de discrimination maximal s'élève à 1.0. Dans ce cas, l'intégralité des candidats appartenant au tiers le plus élevé a répondu correctement à un item, alors qu'aucun candidat n'y est parvenu dans le tiers le plus faible. Les items qui ont un index de discrimination maximal de 1.0 ont une valeur de facilité comprise entre 33% et 66%. Les items qui ont une valeur de facilité comprise entre ces deux taux permettent de distinguer maximalement entre les candidats (Alderson, Clapham & Wall 1995 : 83). En réalité, il est rare qu'un item possède un index de discrimination aussi puissant. Les concepteurs de tests se satisfont le plus souvent d'items dont l'index de discrimination atteint au moins 0.4. Cela étant dit, il n'existe pas de règle stipulant que tel ou tel index de discrimination est un index plancher parce que l'index dépend non seulement de la qualité de l'item mais aussi de son type et du répertoire des capacités des candidats (Alderson, Clapham & Wall 1995 : 83). Les items dont l'index de discrimination est très bas ne sont pas nécessairement défectueux. La valeur de facilité d'un item est le facteur central déterminant l'index de discrimination (Hughes 2003 : 228). Ainsi, les items très faciles et très difficiles, ont systématiquement un index de discrimination bas, même si c'est pour des raisons différentes (Hughes 2003 : 228). Les items très faciles sont traités correctement par la plus grande partie des candidats et n'ont donc pas d'effet discriminant entre candidats. Les items très difficiles ont également un index de discrimination très bas, car ils permettent uniquement d'identifier les meilleurs candidats en éliminant tous les autres, qui se retrouvent mis à l'écart en bloc, sans nuance (Hughes 2003 : 228). Les items à faible discrimination peuvent être nécessaires s'ils correspondent à l'objectif particulier d'un test, par exemple, déceler les très bons candidats uniquement (et écarter tous les autres) ; ou au contraire, s'assurer que des prérequis élémentaires sont validés par la majorité des candidats. Ces exemples servent à montrer que les items dont l'index de discrimination est bas ne doivent pas être systématiquement écartés ou corrigés, dès lors qu'ils ne sont pas défectueux (Hughes 2003 : 227). Signalons qu'il existe désormais de multiples programmes informatiques permettant de calculer l'index de discrimination automatiquement.

Il est important de signaler que l'index de discrimination peut prendre une valeur négative (Hughes 2003 : 225). Cela se produit lorsqu'il y a davantage de réponses correctes données par le tiers le plus bas des candidats que par le tiers

le plus compétent (Hughes 2003 : 226). Dans la mesure où l'index de discrimination est censé avoir un effet positif sur la fiabilité du test, plus le nombre d'items possédant un index de discrimination élevé est grand, plus le test est fiable dans son entier (Hughes 2003 : 226-227). A l'opposé, la présence d'items avec un index de discrimination négatif compromet gravement la fiabilité d'un dispositif (Hughes 2003 : 45). Voilà pourquoi les items avec des index de discrimination négatifs ne sauraient être acceptés: loin d'accroitre la fiabilité du test, de tels items la réduisent.

Quelle que soit la procédure adoptée pour le calcul de l'index de discrimination, il est nécessaire de répartir les candidats en groupes de niveaux de compétences au préalable (Alderson, Clapham & Wall 1995 : 84). L'objectif principal d'un test de positionnement, qui est la répartition des candidats selon leurs niveaux de compétences, exige la prise en compte de ces niveaux dans le calcul de l'index de discrimination des items.[40]

3.6.8.5 Analyse de distracteurs

Lorsqu'on analyse des items au format QCM, il est nécessaire d'analyser la performance des distracteurs, en complément des deux valeurs présentées plus haut. La raison à cela est que dans ce type d'items un index de discrimination bas est parfois imputable au mauvais fonctionnement des distracteurs[41]. L'alarme est donnée lorsque le tiers le plus fort des candidats choisit le distracteur (et non la bonne réponse), ou alors lorsque personne ne tombe dans le piège (Alderson, Clapham & Wall 1995 : 82). L'item de compréhension de l'oral suivant contient deux distracteurs problématiques :

> We are going to a film tonight. Do you want to come along?
> A. Where are you going tonight?
> B. Are you going to see a film tonight?
> C. Thanks. What time is it?
> D. Are you going along now? (Alderson, Clapham & Wall 1995 : 85).

Le distracteur B ne remplit pas sa fonction car il n'a été choisi par personne. Il en est de même pour le distracteur D qui a été majoritairement choisi par le tiers le plus fort (Alderson, Clapham & Wall 1995 : 86).

3.7 Analyse du test entier

Au-delà de l'analyse des items individuels, il faut obtenir des informations statistiques précises sur le test entier. Les différents types de mesure exposés ci-après permettent d'atteindre ce but.

3.7.1 Le Tableau de fréquence

L'élaboration d'un tableau de fréquence constitue la première étape de l'analyse statistique d'un test. Ce tableau présente les différents scores obtenus et les rapporte au nombre de candidats. Cette information générale est utile parce qu'elle facilite la compréhension des effets des différents seuils entre les niveaux et le marquage du score limite nécessaire à la réussite (Hughes 2003 : 220). Toutefois, ce tableau sert uniquement à donner un aperçu global de performance.

3.7.2 Les Mesures de tendance centrale : la Moyenne, le Mode et la Médiane

Les mesures de tendance centrale représentent les scores typiques obtenus à un test. Il en existe trois. La plus commune est la moyenne de tous les scores obtenus lors d'une session. Cette valeur est calculée en ajoutant tous les scores et en divisant le score total par le nombre de participants au test (Hughes 2003 : 220). Les autres mesures de tendance centrale sont le mode et la médiane. Tandis que le mode est le score le plus commun, la médiane est située au milieu d'une séquence de scores individuels. S'il y a un nombre pair de scores, la médiane est calculée en ajoutant les deux scores du milieu et en les divisant par deux (Hughes 2003 : 221).

3.7.3 Fiabilité

La fiabilité est une qualité qui peut être calculée. Puisque la fiabilité d'un dispositif dépend du nombre d'items, elle peut être augmentée ou réduite en les ajoutant ou en les enlevant, sous condition que les items soient indépendants de tous les autres inclus dans le même test (Hughes 2003 : 223). Il est possible de quantifier la fiabilité d'un test au moyen d'un coefficient. Ce dernier peut prendre une valeur comprise entre zéro et 1. Un coefficient égal à zéro implique qu'un test donne deux ensembles de résultats complètement déconnectés l'un de l'autre. En

revanche, le montant 1 est le coefficient idéal parce qu'il suggère l'obtention de deux ensembles de résultats exactement identiques à deux différentes passations du test (Hughes 2003 : 39). Bien que théoriquement envisageable, l'obtention de ces deux coefficients extrêmes n'est pas possible en pratique. De même, il n'est pas possible de nommer un montant particulier à viser pour tous les tests en langue, quel que soit leur usage. Le coefficient de fiabilité souhaitable dépend du ou des domaines de compétences évalué(s) par le dispositif donné, de l'usage réel ou prévu. Le degré d'importance des décisions prises sur la base des scores est un facteur déterminant pour le coefficient de fiabilité. En effet, la fiabilité sera d'autant plus élevée que les décisions seront importantes (Hughes 2003 : 39).

Il existe des méthodes différentes pour calculer le coefficient de fiabilité. La plus évidente consiste à faire passer le même test deux fois par le même groupe de candidats. Cette méthode est connue sous l'appellation « test-retest method » (Hughes 2003 : 40). Il en existe une variante, « the alternate forms method », qui permet de soumettre deux formes différentes du même test à un ensemble de candidats (Hughes 2003 : 40). Toutefois, ces deux méthodes sont rarement appliquées, premièrement, pour des raisons pratiques et, deuxièmement, en raison de la disponibilité de méthodes plus économiques qui permettent également de calculer le coefficient de fiabilité. Ces dernières exigent une seule passation du test. Chaque participant obtient deux scores, respectivement attribués aux deux moitiés du test. A partir de ces deux scores, un coefficient de fiabilité interne, analogue à celui calculé au moyen de deux passations du test, est obtenu (Hughes 2003 : 40). Pour que cette méthode fonctionne, il faut que le test soit divisé en deux moitiés tout à fait équivalentes. Il a été empiriquement démontré que cette méthode fournit des coefficients très proches de ceux qui s'appuient sur deux formes différentes d'un même test (Hughes 2003 : 40).

3.8 Item Response Theory

L'analyse statistique des items peut s'effectuer en s'appuyant sur d'autres méthodes que l'analyse traditionnelle. De nouvelles méthodes d'analyse ont été développées ces dernières années sous le nom *Item Response Theory* (Hughes

2003 : 228). Cette théorie, couramment désignée par l'acronyme IRT, permet de calibrer le niveau de difficulté d'un item, sans faire entrer en jeu la compétence des candidats qui y ont répondu (Alderson & Huhta 2005 : 313). En dehors de cette théorie, le niveau de difficulté est influencé par la compétence des participants au test (Alderson & Huhta 2005 : 313). La méthode constitutive de la théorie IRT la plus utilisée en évaluation langagière est l'analyse Rasch (Hughes 2003 : 228). Cette analyse repose sur l'idée que chaque item a un niveau de difficulté particulier et que chaque candidat se situe à un niveau de compétence fixe. Ces deux informations servent à créer un modèle de réponses des candidats aux items constitutifs du test. Bien que l'analyse Rasch présume un niveau de compétences stable pour chaque candidat, elle tient compte du fait que la performance d'un individu ne reflète jamais complètement sa compétence (Hughes 2003 : 228). L'écart entre la performance réelle d'un candidat et la performance prédite par le modèle est accepté. Cette analyse identifie toutefois les candidats et les items dont la performance est significativement différente de celle attendue (Hughes 2003 : 228). Le degré d'adaptation de tous les items est évalué, ce qui permet de repérer les supports et les tâches inadaptés à la performance attendue. En outre, la performance de chaque candidat à chaque item est calculée ce qui permet de repérer les performances inadaptées au modèle (Hughes 2003 : 230).

Les deux méthodes, l'analyse Rasch et l'analyse traditionnelle, doivent être conçues comme complémentaires dans la mesure où elles contribuent au développement de dispositifs de meilleure qualité. En effet, l'une et l'autre mettent en évidence les défauts des items constitutifs d'un test (Hughes 2003 : 232). Toutefois, l'analyse Rasch possède quelques nouvelles fonctionnalités qui permettent un examen plus fin de la performance des items et des candidats individuels. La possibilité d'identifier et d'exclure les items inadaptés ou les apprenants ayant une performance insuffisante constituent un exemple des nouvelles fonctions apportées (North 2000 : 221). Malgré ces apports, le modèle Rasch requiert des prises de décision à tous les points clés pendant la procédure d'analyse des items. La formulation de jugements bien fondés est facilitée par les éléments suivants : la mise à disposition des descripteurs ayant passé le processus du pilotage, l'accès à certaines informations sur les candidats, le degré de familiarité des évaluateurs avec les domaines de compétences

mesurés (North 2000 : 221). Ces trois conditions étaient remplies lors du pilotage du test de positionnement POSILANG, ce qui a permis des prises de décision réfléchies.

Notes:

[1] Lapaire : Bilan étape 2012

[2] Lapaire : Bilan étape 2012

[3] Lapaire : Bilan étape 2012

[4] Pour cette raison, une fiche d'auto-évaluation a été élaborée qui sera soumise à chaque candidat avant la passation du test. Les rubriques qui sont contenues dans la fiche d'auto-évaluation sont les suivantes : 1.Votre Parcours linguistique ; 2. Estimation déclarative de votre niveau. La 1. rubrique comporte 3 questions: 1) Combien d'années avez-vous étudié l'anglais au total au collège et au lycée ? 2) Est-ce que vous avez déjà sejourné dans des pays anglophones ?; Quelle note avez-vous obtenu au bac ?

[5] http://www.britishcouncil.fr/examen/bulats

[6] La comparaison du résultat individuel avec d'autres candidats s'appelle « interprétation normative » tandis que celle par rapport à une performance attendue est désignée « interprétation critériée » (Laurie 2007 : 248).

[7] Le terme « habileté » est utilisé ici délibérément et non pas celui de « compétence ». Ceci est dû au fait que l'habileté est reliée à l'activité de mesure. Comme expliqué dans le premier chapitre « Pratiques et Dispositifs-quelques exemples », on ne peut pas mesurer un niveau de compétence, mais seulement celui d'une connaissance ou d'une habileté (Brown 2010 : 3). Ces termes seront définis et expliqués en détail ci-dessous.

[8] http://www.comefica.com/frcontenu/elao.htm.

[9] http://www.ielts.org/test_takers_information/what_is_ielts.aspx.

[10] Lapaire : bilan étape 01.2013.

[11] La compréhension et la production écrite assistée, ainsi que la compréhension orale, renvoient aux « activités de communication langagière » du CECRL (2005 : 48). Celles-ci sont appelées les « compétences de communication » dans l'approche communicative. Elles sont en réalité des « savoir-faire langagiers » ou des « *skills* » (Bourguignon 2010 : 22).

[12] Lapaire : Bilan étape 2013

[13] Le positionnement prévu englobe l'attribution d´un score et d´un niveau de compétence selon le CECRL

[14] Lapaire : Bilan étape 2013.

[15] Lapaire : Bilan étape 2013.

[16] Dans la littérature, les termes « modèle » et « cadre théorique » sont souvent utilisés de façon interchangeable avec plusieurs significations différentes (Fulcher & Davidson 2005 : 36). Un tel usage est facteur de confusion. En réalité, ces deux termes renvoient à des concepts distincts. Le modèle est une description théorique, englobante et relativement abstraite, de la capacité à communiquer en langue seconde. En revanche, le cadre de référence est une sélection de capacités extraites d'un modèle, qui sont pertinentes dans un contexte d'évaluation spécifique (Fulcher & Davidson 2005 : 36).

[17] L'approche actionnelle est définie et décrite en détail dans le deuxième chapitre de la thèse. Le concept de tâche est central dans cette approche (Tagliante 2005 : 36).

[18] Dans le modèle de Bachman, les composantes de la compétence grammaticale sont les connaissances et non pas les capacités (Bachman 1990 : 87). Cette désignation établit un parallèle avec le CECRL qui définit « les composantes principales de la compétence linguistique » en tant que « connaissance de ressources formelles » (Conseil l'Europe 2005 : 87). La connaissance de ces ressources permet l'usage correct et significatif de la langue (Conseil l'Europe 2005 : 87).

[19]http://www2.cndp.fr/lesScripts/bandeau/bandeau.asp?bas=http://www2.cndp.fr/doc_administrative/programmes/accueil.htm.

[20]http://www.alsace.iufm.fr/web.ressources/web/ressources_pedagogiques/productions_pedagogiques_iufm/anglais/2nddegre/palier1.pdf.

[21] Bachman indique que ce modèle n'est ni une liste définitive ni une liste exhaustive. Il est prévu pour servir de guide dans toute recherche empirique (Bachman 1990 :117).

[22] Puisque les caractéristiques illocutionelles font partie de la compétence pragmatique d'un candidat, elles ne seront pas présentées par la suite. La raison à cela est que la compétence pragmatique n'est pas évaluée dans ce test.

[23] Le premier type d'information contextuelle est celui fourni par le contexte physique immédiat. Or, il n'est pas possible de transmettre d'information contextuelle de ce type dans POSILANG, faute de contexte physique immédiat.

[24] Alors que l'information abstraite peut être représentée dans des modes abstraits ou linguistiques uniquement, l'information concrète peut être représentée dans des modes visuels, sonores, tactiles et kinesthésiques (Bachman 1990 : 135- 136).

[25] Comme il a été expliqué dans le sous-chapitre traitant du modèle de compétence langagière communicative de Bachman, la compétence organisationnelle englobe les capacités qui contrôlent l'organisation formelle du discours (Bachman 1990 : 139).

[26] Ce cadre n'a pas été utilisé dans le premier chapitre, lors de l'analyse des tests de positionnement, parce que le but de l'analyse n'était pas d'analyser les méthodes en détail. Pour cette raison, la méthode ne constitue qu'une seule catégorie parmi les 19 incluses dans la grille d'évaluation utilisée pour l'analyse des tests.

[27] Les spécifications doivent évoluer continuellement, au fur et à mesure des précisions apportées sur la définition des compétences évaluées. Elles résultent de la collecte de données portant sur le lien entre les compétences et les tâches contenues dans un test (Fulcher & Davidson 2002 : 62).

[28] Le lien direct entre l'objectif principal du test et sa catégorie a été présenté et expliqué au début du premier chapitre de cette thèse.

[29] Une tâche se compose d'un texte et d'un item. Cependant, seuls les items sont cités ci-dessous parce que les textes sont cités auparavant.

[30] Pour des explicitations détaillées de la méthode appliquée et des instructions, voir le sous-chapitre consacré à la méthode.

[31] Le terme *domaine du test* est employé dans deux sens. Le premier désigne l'ensemble des tâches et des types de comportement lors de l'usage réel de la langue, qui correspondent aux compétences visées par le construit de ce test. Le domaine du test désigne également l'ensemble des tâches dans le test lui-même (McNamarra 2005 : 25).

[32] Le terme de validation rationnelle, par opposition à validation empirique, constitue une distinction traditionnelle entre les deux méthodes de validation. Cette distinction traditionnelle est remplacée par une autre dichotomie à l'heure actuelle qui fera l'objet d'un sous-chapitre distinct.

[33] http://www.lancaster.ac.uk/fss/projects/grid/

[34] Ces grilles d'analyse sont présentées et expliquées dans le chapitre précédent, « Adosser un test au CECRL ».

[35] La publication de la version pilote du Manuel en 2003 a déclenché une série des projets qui consistaient à relier les tests existants au CECRL (Manuel 2009 : 5).

[36] Ce procédé satisfait aux limites pratiques imposées par la thèse. En vérité, le contenu de chaque tâche sera décrit selon les catégories de la grille et leur niveau commun de référence sera estimé.

[37] La grille contenant toutes les caractéristiques de toutes les tâches présentées se trouve dans l'annexe à ce chapitre.

[38] Le concept de validité faciale, expliqué dans le premier chapitre, signifie l'impression par rapport à la validité du test (Brown 2010 : 35).

[39] Il faut noter que la notion de «validité de contenu » s'oppose à « la théorie platonicienne des traits », car selon cette théorie, toute habileté (y compris langagière), ne peut se manifester qu'imparfaitement. En conséquence, le contenu des items constitue une représentation imparfaite d'une ou de plusieurs habiletés langagières à mesurer (Belanger 2002 : 4). En revanche, la notion de « validité de contenu » est en accord avec la théorie fonctionnaliste parce que celle-ci ne contient pas la notion de représentation imparfaite d'une habileté évaluée (Belanger 2002 : 4).

[40] Il convient de préciser qu'afin de calculer l'index de discrimination, il suffit de répartir les candidats en trois tiers selon leur score individuel au test. En revanche, l'objectif d'un test de positionnement comme POSILANG est de répartir les candidats en groupes de niveaux qui correspondent précisément à leur niveau de compétences.

[41] Comme évoqué dans le sous-chapitre consacré à la conception des items au format QCM, les distracteurs sont les options incorrectes fournies à côté de la bonne réponse, désignée la réponse-clé.

[42] http://www.alte.org/attachments/files/alte_cando.pdf

[43] http://www.alte.org/attachments/files/alte_cando.pdf

[44] Un trait est une habileté ou un autre phénomène cible évalué à partir des scores obtenus aux items. Une distribution normale cumulée se manifeste sous forme d'une ogive (Bélanger 2002 : 7).

Chapitre 3 - Annexe I : sources utilisées

SITES INTERNET	TACHES
www.usingenglish.com/reference/idioms/	- Idiomes (à partir de B1) - Expressions idiomatiques - Proverbes
http://www.learn-english-today.com/idioms/idioms_proverbs.html	- Idiomes - Expressions idiomatiques - Proverbes
www.english-idioms.de/	- Idiomes

MANUELS	Niveaux	Taches
Schwarz et al. (Hrsg.): English G Band A1 für das 5.Schuljahr an Gymnasien, 1.Auflage Berlin: Cornelsen 1985	A1	- Description du trajet - Pronoms personnels et possessifs - Construction du comparatif des adjectifs - Marqueurs temporels
Créativité personnelle	A1	- Compréhension d´un texte cohérent - Impératif - Pronoms interrogatifs - Degrés de l'adjectif - Synonymes de noms - Usage correct de pronoms personnels - Choix des temps (*simple present* vs. *present progressive*)
Weisshaar, H. (Hrsg.)Green Line 2, Klett, Stuttgart/ Leipzig 2006. Textbook and workbook	A2	- Compréhension du sens d´un texte cohérent - Exercice de grammaire (usage du *go-future*) - Exercice de grammaire *(simple past*)
Otte, M.D. (Hrsg.) Englisch komplett/5.8.Schuljahr. Stuttgart, Klett 2004. 4	A2	- Compréhension d´un texte cohérent - Exercice de lexique (lexique adverbial) - Exercice de grammaire (adverbes indéfinis) - Choix des temps *(simple past vs. present perfect*)
Créativité personnelle	A2	- Exercice de lexique (nominal) - Choix des synonymes

		- Compréhension du vocabulaire dans un texte cohérent - Exercice de grammaire (ordre des mots dans une phrase)
Weisshaar, H. (Hrsg.) Green Line 3. Stuttgart/ Leipzig, Klett 2007. Textbook and workbook	B1	- Compréhension d'un texte cohérent - Exercice de grammaire (préposition) **-** Exercice de grammaire (discours indirect et concordance des temps au passé)
East, P.&B. McCredie (Hrsg.)Englisch. 9/10 Klasse. Köln, Komet 2007 5	B1	- Compréhension d'un texte cohérent - Exercice de grammaire (pronoms relatifs) dans une phrase isolée
Creativité Personelle	B1	- Compréhension d´une phrase dans un paragraphe - Exercice de grammaire (*had better+ infinitive*, verbes modaux, passif, *used to* + infinitif) - Exercice de lexique (lexique nominal et nom indénombrable, collocation)
Bernhard Knop, Corienne Naumann-Breeze et al (Hrsg.): Words in context : thematischer Oberstufenwortschatz, Klett, Taschenbuch, 1. Januar 1998 6	B1	- Exercice de lexique (noms composées) dans un paragraphe
Schwarz, H. (Hrsg.): English G 21 Workbook, Berlin: Cornelsen 2008	B1	- Exercice de grammaire: *passives, indirect speech*
Christie, D (Hrsg.) : Gateway: Englisch für berufliche Schulen: Stuttgart: Klett: 2006	B1	- Compréhension d'un texte cohérent
Words in context	B1	- Compréhension d'un texte cohérent (USA)
Words in context	B2	- Compréhension d'un texte cohérent (*South Africa, religion*) -Exercice de lexique (lexique nominal)
Creativité personnelle	B2	- Exercice de lexique (lexique verbal)

Thaler,E. (Hrsg.) : The new Summit : Text and Methods, Bildungshaus Braunschweig 2007	B2/C1	- Compréhension du sens d'un paragraphe (*American culture*) - Compréhension du sens d'une phrase (contenant une expression idiomatique) - Exercice de lexique (lexique nominal) dans un paragraphe (*American Dream, Women in the world of work*)
Schwarz,H. (Hrsg.) English G. 2000 A5 Berlin : Cornelsen 2001	B2	- Exercice de grammaire (pronoms indéfinis et possessifs)
Weisshaar, H. (Hrsg.) :Green Line Oberstufe Klasse 10, Stuttgart, Leipzig : Klett 2010	B2	- Compréhension d'un texte cohérent (Médicine)
Creativité personnelle	B2	- Compréhension du sens d´un paragraphe (Dialogue entre voisins)
Weisshaar, H. (Hrsg.) :Green Line Oberstufe, Stuttgart, Leipzig : Klett 2009	B2/C1	- Compréhension d'un texte cohérent (Stem cells) - Comprendre une expression idiomatique - Comprendre le sens d´un discours
Green Line: Oberstufe Skill and Examtrainer, Nordrhein- Westfalen: Stuttgart, Leipzig : Klett 2009	B2/C1	- Compréhension du sens des phrases - Exercice de grammaire (conjonctions de subordination, adverbe, comparatif/superlatif)
Words in context	C1	- Compréhension d'un texte cohérent (racial issue in the USA) - Exercice de lexique (lexique nominal) à l'aide du contexte (Théâtre)
Newspaper article The Guardian, Wednesday 22, August 2012	C1	- Compréhension du sens d´un article
Newspaper article Reuters.com, Mon Aug 20, 2012	C1	- Compréhension du sens d´un article (*Pussy riot*)
http://time.com/world/ by Tony Karon Aug. 22, 2012	C1	- Compréhension du sens d´un discours (*As South Africa Reels from Mine Shootings, Social Inequality Threatens to Undo the Post-Apartheid 'Miracle*)
Newspaper article http://healthland.time.com	C1	- Compréhension du sens d´un discours (*Older dads linked to kids autism and shizophrenia*)

By Alice ParkAug. 23, 2012		
Newspaper article http://articles.philly.com/2012-08-17/news/33249254_1_natural-gas-fall-in-coal-prices-energy-department	C1	- Compréhension du sens d´un discours (*Carbon dioxide surprise sees decreased levels being released*)

Chapitre 3 - Annexe II

A1:

1.Comprehension orale

1. At the zoo. The zookeeper shouts:
« Don't feed the animals! "

What did you hear?

A. Don't need the animals
B. Don't feed the animals
C. Don't nourish the animals
D. Don't lead the animals.

Description	Dimension
Text source	Routine commands (instruction)
Authenticity	Not authentic
Discours type	Mainly instructive
Domain	Public
Topic & Content	Free time, entertainment
Nature of content	Only concrete content
Text length	Max. 10 sec.
Vocabulary	Only frequent vocabulary
Grammar	Only simple structures
Text speed	slow
Number of participants	2 participants
Accent/ standard	Standard accent
Clarity of articulation	Clearly articulated
How often played	Played once
View CEFR scales	By level
Text likely to be comprehensible by learner at CEFR level	A1

Description	Dimension
Item type	Multiple choice
Operations	Recognize communicative purpose from explicit information
Item Level Estimated	A1

Task level estimate: A1

2. Compréhension écrite

5. Reliez les deux parties des phrases suivantes:

Benny arrives at Heathrow after a week in Spain. He is waiting for his friend´s mother to fetch him.

1. Max´s mother can´t meet Benny at the airport…

2. There are taxis from the airport to the city centre…
3. If Benny has problems at the airport…
4. There is only one tube line from Heathrow...

A. ...but they are expensive
B. ...he can go to an information desk
C. ...because she works until 6 o`clock
D. ... and that is the Piccadilly Line

Description	**Dimension**
Text source	Text books, reader
Authenticity	Pedagogic
Discours type	Mainly narrative, stories
Domain	personal
Topic	Travel
Nature of content	Only concrete content
Text length in words	19 words
Vocabulary	Only frequent vocabulary
Grammar	Only simple structures
View CEFR scales	by level
Text likely to be comprehensible by learner at CEFR level	A1

Description	**Dimension**
Item type	Multiple choice
Operations	Recognise text structure/connections between parts from explicit information
Item Level Estimated	A1

Task level estimate: A1

3. Expression écrite

3. Which word best describes the children's « work »?:
In this family, the children must tidy up their rooms, lay the table and stack the dishwasher.

A. Homework
B. Housework
C. Guesswork
D. Teamwork

Description	**Dimension**
Text source	Textbooks, material
Authenticity	Pedagogic
Discours type	Mainly descriptive, impressionistic descriptions
Domain	Personal
Topic	House and home, environment
Nature of content	Only concrete content
Text length in words	17 words
Vocabulary	Only frequent vocabulary

Grammar	Only simple structures
View CEFR scales	by level
Text likely to be comprehensible by learner at CEFR level	A1

Description	Dimension
Item type	Multiple choice
Operations	Recognise main idea from explicit information
Item Level Estimated	A1

Task level estimate: A1

A2

1.Compréhension orale

3. *You will hear a question and four answers. Select the correct answer to this question.*
What are you going to do on the week-end?

A.I ski in the mountains.
B. I want stay at home.
C. I am going to visit my aunt in Sheffield.
D. I am planning travel to London.

Description	Dimension
Text source	Interpersonal dialogues and conversations
Authenticity	Not authentic
Discours type	Mainly expository, explications, talks
Domain	personal
Topic & Content	Free time, entertainment
Nature of content	Only concrete content
Text length	Max. 10 sec.
Vocabulary	Only frequent vocabulary
Grammar	Only simple structures
Text speed	slow
Number of participants	2 participants
Accent/ standard	Standard accent
Clarity of articulation	Clearly articulated
How often played	Played once
View CEFR scales	by level
Text likely to be comprehensible by learner at CEFR level	A2

Description	Dimension
Item type	Multiple choice
Operations	Recognise communicative purpose from explicit information
Item Level Estimated	A2

Task level estimate: A2

2. Compréhension écrite

5. *You read the following headline in the daily newspaper:*

The annual divorce rate is rising in Britain.

Which of the following words has a similar meaning as the underlined one?

A. illness

B. separation

C. unemployment

D. social conflicts

Description	Dimension
Text source	Newspapers
Authenticity	Adapted, simplified
Discours type	Mainly narrative, e.g. news reports
Domain	public
Topic	Relations with other people
Nature of content	Only concrete content
Text length in words	8 words
Vocabulary	Only frequent vocabulary
Grammar	Only simple structures
View CEFR scales	by level
Text likely to be comprehensible by learner at CEFR level	A2

Description	Dimension
Item type	Multiple choice
Operations	Recognise main idea from explicit information
Item Level Estimated	A2

Task level estimate: A2

3. Expression écrite

2. *Make opposite pairs. Find the words with an opposite meaning to the underlined ones.*
I like people who are strong and kind.

1. strong: 2.kind:

A. Interesting, A. Interesting,
B. Glad, B. Glad,
C. Weak C. Weak,
D. Nasty D. Nasty

Description	Dimension
Text source	Textbooks, readers
Authenticity	pedagogic

Discours type	Mainly argumentative, opinions, comments
Domain	personal
Topic	Personal identification
Nature of content	fairly abstract content
Text length in words	8 words
Vocabulary	Mostly frequent vocabulary
Grammar	Mainly simple structures
View CEFR scales	by level
Text likely to be comprehensible by learner at CEFR level	A2

Description	**Dimension**
Item type	Multiple choice
Operations	Recognise main idea from explicit information
Item Level Estimated	A2

Task level estimate: A2

B1.

1.Compréhension orale

3. *Listen to Jean´s statement about his professional life:*
Jean used to work in a company which produces motor vehicle tires.
What does this sentence express?

A. Jean still works in the company.
B. Jean produces motor vehicle tires.
C. Jean worked in that company.
D. Jean is going to work for the company.

Description	**Dimension**
Text source	Interpersonal dialogues and conversations
Authenticity	Not authentic
Discours type	Mainly expository, explications, talks
Domain	personal
Topic & Content	Other : work
Nature of content	Only concrete content
Text length	Max. 10 sec.
Vocabulary	Rather extended
Grammar	Mainly simple structures
Text speed	slow
Number of participants	1 participant
Accent/ standard	Standard accent
Clarity of articulation	Clearly articulated
How often played	Played once
View CEFR scales	by level
Text likely to be comprehensible by learner at CEFR level	B1

Description	**Dimension**
Item type	Multiple choice

Operations	Recognise main idea/gist from explicit information
Item Level Estimated	B1

Task level estimated: B1

2.Compréhension écrite

4. You watch a Chart show on TV and the following idea comes to your mind:
"There are few singers_____ names I´ve never heard before."
Select the correct pronoun for the gap.
A. who´s
B. that
C. who
D. whose

Description	Dimension
Text source	Exercise materials
Authenticity	Adapted, simplified
Discours type	Mainly expository, explications, talks
Domain	personal
Topic	Free time, entertainment
Nature of content	Only concrete content
Text length in words	10 words
Vocabulary	Only frequent vocabulary
Grammar	Limited range of complex structures
View CEFR scales	by level
Text likely to be comprehensible by learner at CEFR level	B1

Description	Dimension
Item type	Multiple choice
Operations	Recognise detail from explicit information
Item Level Estimated	B1

Task level estimate: B1

3.Expression écrite

5. Anna tells her friend Ben what she thinks about her parents.
Choose the correct reflexive pronoun from those below:

-Anna: "My mum always does so much for me. That's really incredible!
-Ben: "How do you mean it?"
-Anna: "I mean parents sometimes also need to have a good time and enjoy _____."
A. ourselves
B. theirselves
C. themselves
D. himselves

Description	Dimension
Text source	Text books, readers
Authenticity	pedagogic
Discours type	Mainly expository, explications, talks
Domain	personal
Topic	Relations with other people

Nature of content	Mostly concrete content
Text length in words	32 words
Vocabulary	Mainly frequent vocabulary
Grammar	Limited range of complex structures
View CEFR scales	By level
Text likely to be comprehensible by learner at CEFR level	B1

Description	Dimension
Item type	Multiple choice
Operations	Recognise detail from explicit information
Item Level Estimated	B1

Task level estimate: B1

B2

1. Comprehension orale

5. Your mom tells you the following proverb after you report about your new job:
"(A) burden of one's own choice is not felt.
What does the proverb mean?

A. You work hard even if you choose what you like.
B. It´s not always good to make your own choices.
C. To make one´s own choice is a burden.
D. Something difficult seems easier when it is done voluntarily.

Description	Dimension
Text source	Interpersonal dialogues and conversations
Authenticity	authentic
Discours type	Mainly argumentative, comments
Domain	personal
Topic & Content	Other : work
Nature of content	Mainly abstract content
Text length	Max. 15 sec.
Vocabulary	Rather extended
Grammar	Mainly simple structures
Text speed	normal
Number of participants	2 participants
Accent/ standard	Standard accent
Clarity of articulation	Clearly articulated
How often played	Played once
View CEFR scales	by level
Text likely to be comprehensible by learner at CEFR level	B2

Description	Dimension
Item type	Multiple choice
Operations	Recognise communicative purpose from implicit information
Item Level Estimated	B2

Task level estimate: B2

2. Comprehension écrite

1. You read the following article in the newspaper commemorating Alexander Fleming´s birthday:

Only 80 years ago, it was not unusual for many children to die before adulthood. One major reason for this was that there were no cures for common bacterial illnesses such as pneumonia, diphtheria or scarlet fever. Thanks to biologists like Scotsman Alexander Fleming (1881-1955), whose discovery of penicillin helped to launch the field of antibiotic medicine, contracting bacterial illnesses was no longer a death sentence-at least for those who had access to good medical treatment.

Which of the following statements is false?

A. Alexander Fleming was a biologist.
B. Alexander Fleming´s invention of penicillin was a starting point for research on antibiotics.
C. A disease like pneumonia was a common cause of death.
D. Fleming discovered the antibiotic Penicillin more than 200 years ago.

Description	Dimension
Text source	Newspapers
Authenticity	genuine
Discours type	Mainly expository, interpretations, article
Domain	public
Topic	Health and bodycare
Nature of content	Mostly concrete content
Text length in words	90 words
Vocabulary	extended
Grammar	Wide range of complex structures
View CEFR scales	by level
Text likely to be comprehensible by learner at CEFR level	B2

Description	Dimension
Item type	Multiple choice
Operations	Recognise detail from explicit information
Item Level Estimated	B2

Task level estimate: B2

3. Expression écrite

5. You find the following text in your university paper:
After all, by many measures, there´s never been a better time to be a woman. In places like Scandinavia and Britain, a third or more corporate managers are now women. Latin America has seen a 50 percent jump in the number of women politicians in the last decade. However, in the EU Parliament, only 23 out of 162 members are female. Even though more than 50 percent of students in higher education in many parts of the world are women, the representation of women in corporate leadership has been stagnant for the last few years.
Please choose the statement which best explains the word "corporate":

A. Corporate designates commercial companies.
B. Corporate designates nonprofit organizations.
C. Corporate designates political leadership.
D. Corporate designates higher education.

Description	Dimension
Text source	Newspapers
Authenticity	genuine
Discours type	Mainly expository, interpretations, article
Domain	occupational
Topic	Other: work
Nature of content	Mostly concrete content
Text length in words	114 words
Vocabulary	extended
Grammar	wide range of complex structures
View CEFR scales	by level
Text likely to be comprehensible by learner at CEFR level	B2

Description	Dimension
Item type	Multiple choice
Operations	Recognise detail from explicit information
Item Level Estimated	B2

Task level estimate: B2

C1
1. Compréhension orale

1. You are listening to a radio program about theatres in capitals world-wide:

The Mecca of British theatre-goers is London, which boasts more than 200 professional theatres, many of which are concentrated in the West End and on the South Bank. The centre of the theatre scene in the US is Broadway in New York. New stage productions are presented every year, many of them costing several million pounds or dollars. Production costs however, are not a guarantee of box-office success. Most large theatres are repertory theatres with a stock company of actors performing a repertoire of plays at these particular theatres only. Provincial theatres are mainly served by touring companies as they usually cannot afford their own repertory companies.
Which of the following statements is false:

A. London has more than 200 different theatres.
B. Production costs are normally a guarantee for box office success.
C. Provincial theatres don´t have the funds for their own repertory companies.
C. London theatres are mostly found in the West End and the South Bank.

Description	Dimension
Text source	TV/ radio documentaries
Authenticity	authentic
Discours type	Mainly narrative, reports
Domain	public
Topic & Content	Free time, entertainment
Nature of content	Mostly concrete content
Text length	Approx. 45-50 sec.
Vocabulary	extended
Grammar	wide range of complex structures
Text speed	normal
Number of participants	1 participant
Accent/ standard	Standard accent

Clarity of articulation	Clearly articulated
How often played	Played once
View CEFR scales	by level
Text likely to be comprehensible by learner at CEFR level	C1

Description	Dimension
Item type	Multiple choice
Operations	Recognise detail from explicit information
Item Level Estimated	C1

Task level estimate: C1

2. Comprehension écrite

6. Sarah reads in her history book the chapter about American history.

The backbones of US democracy and the system of government are the Declaration of Independence (1776) and the Constitution of the USA (1787) with its seven original articles and 26 amendments.

What does the underlined word stand for?

A. memoires

B. spine

C. principle

D. idea

Description	Dimension
Text source	Text books, readers
Authenticity	Adapted/ simplified
Discourse type	Mainly expository, explications
Domain	Educational
Topic	Other: history
Nature of content	Fairly abstract content
Text length in words	39 words
Vocabulary	extended
Grammar	limited range of complex structures
View CEFR scales	by level
Text likely to be comprehensible by learner at CEFR level	C1

Description	Dimension
Item type	Multiple choice
Operations	Recognise main idea from explicit information
Item Level Estimated	C1

Task level estimate: C1

3. Expression écrite

6. You come across the following statement in a British newspaper:

Feeling abandoned and ignored the North of the UK accuses the government in London of being biased in favor of the South which is better off economically.

What does the underlined phrase express?

A. The government in London favors the North.
B. The government in London favors the South.
C. The government in London is irresolute.
D. The government in London is prejudiced against the North.

Description	Dimension
Text source	Newspapers
Authenticity	Genuine
Discourse type	Mainly narrative, news reports
Domain	Public
Topic	Other: politics
Nature of content	Fairly abstract content
Text length in words	37 words
Vocabulary	Extended
Grammar	Wide range of complex structures
View CEFR scales	by level
Text likely to be comprehensible by learner at CEFR level	C1

Description	Dimension
Item type	Multiple choice
Operations	Recognise main idea from explicit information
Item Level Estimated	C1

Task level estimate: C1

4 Validation empirique des tâches de POSILANG

La version pilote du test POSILANG a été soumise à deux groupes d'étudiants en sciences de langage et en première année de licence d'anglais (L1). Ces deux groupes se distinguent par leur spécialisation et par leur degré d'avancement dans le cursus universitaire. Les linguistes du MASTER-RELAI accomplissent leur dernière année d'études en sciences de langage (SDL). Parvenus en M2, ils ont acquis un bon niveau de spécialité en linguistique générale et ont été régulièrement exposés à des écrits scientifiques en langue anglaise. Rien ne permet en revanche de garantir qu'ils ont pratiqué la langue courante durant toutes ces années.

Tout autre est la situation des « spécialistes » de première année. Les étudiants en première année de licence d'anglais sont des primo-entrants de niveau bac. Ils n'ont que quelques mois de spécialisation derrière eux et entament à peine leur exploration de la langue de spécialité en civilisation et en analyse littéraire. Officiellement engagés dans un cursus classique de langue, littérature et civilisation anglophones (LLCE), ils vivent encore sur les bases générales qu'ils ont acquises dans le secondaire. Il est important de signaler qu'un tiers environ de ces étudiants ne poursuivra pas en L2, par défaut d'orientation mais aussi par difficulté à atteindre le niveau B2 du Cadre, pourtant attendu en fin de terminale au lycée. Rappelons qu'en LLCE, l'accent est mis sur la maîtrise d'une seule langue vivante étrangère, en l'occurrence l'anglais. Bien qu'ils soient officiellement anglicistes, on ne peut s'attendre à ce que ces primo-entrants aient le même niveau de conscience linguistique et le même degré de compétence en anglais que leurs aînés en master de sciences du langage. Ces étudiants L1 ont néanmoins été choisis en priorité pour passer la version pilote du test POSILANG parce qu'ils correspondaient exactement à la population hétérogène ciblée par notre test. Comme nous l'avons dit dans le troisième chapitre de cette thèse, POSILANG est essentiellement conçu pour évaluer le niveau de compétences des bacheliers en anglais s'inscrivant pour la première fois dans une université.

Il est important de préciser que la passation du test pilote s'est déroulée à l'issue d'un semestre complet de cours (L1-S1 ; M2-S3), dans les mêmes conditions matérielles pour les deux groupes, à savoir dans un laboratoire

multimédia de langue, en mode papier enrichi de fichiers son. Avant de rendre compte des résultats de passation de la version pilote de POSILANG, il convient de préciser l'établissement des niveaux de référence des participants.

4.1 Etablissement des niveaux de référence

4.1.1 Groupe L1

La nécessité d'établir des niveaux de référence est d'autant plus forte qu'une grande hétérogénéité règne dans les deux groupes en matière de compétences langagières. En ce qui concerne le groupe de Licence d'anglais, ces étudiants sont de jeunes bacheliers censés avoir atteint le niveau B2 en anglais, à l'issue de leurs études secondaires. La réalité de terrain est cependant fort différente : certains ont choisi une spécialisation en langue parce qu'ils étaient bons dans cette matière. Ils ont donc un niveau d'entrée B2+, voire C1. D'autres, au contraire, se sont décidés pour une licence d'anglais pensant à tort qu'un enseignement entièrement spécialisé leur permettrait d'améliorer sensiblement leurs performances. Ces étudiants se situent souvent dans une zone intermédiaire entre B1 et B2 pour l'expression écrite ou orale. Malgré cette insuffisance, on peut estimer que, même chez ces étudiants, le niveau B2 est atteint en compréhension écrite ou orale.

Pour assigner un niveau global de référence à chaque étudiant dans notre *benchmark*, le procédé suivant a été adopté. Premièrement, un pré-niveau global a été attribué sur la base de l'épreuve terminale de lexicologie-grammaire passée par ces étudiants en fin de premier semestre (S1) dans le cadre de leur L1 d'anglais. Le barème suivant a été adopté : note inférieure ou égale à 8 (B1), note comprise entre 9 et 11 (B2), entre 12 et 14 (B2+), 15 et au-delà C1. Ensuite, ce pré-niveau a été confirmé ou modifié en observant le détail des performances écrites, tant en lexicologie qu'en grammaire. Pour évaluer la compétence lexicale, la formulation des définitions demandées et la restitution des proverbes courants ont été testées. Dans le but d'évaluer la compétence grammaticale, la façon de formuler les définitions a été analysée encore une fois, ainsi que la formation du *simple past* et du *present perfect* avec les verbes réguliers et irréguliers en contexte. La traduction d'une dizaine de phrases a en outre été

examinée. La correction des formes verbales du passé a été notée à l'aide de l'indicateur « verbes (ir)réguliers », alors que la traduction correcte des dix phrases a été évaluée au moyen de l'indicateur « thème grammatical ». Ces deux indicateurs sont extraits de l'épreuve finale, passée à la fin du premier semestre de l'année 2014/2015. Afin de consolider le niveau de référence, désigné *benchmark,* la performance en lexicologie et en grammaire a été combinée avec le pré-niveau global. Pour montrer l'établissement des niveaux de référence dans ce groupe, quelques fautes des candidats, situés aux différents niveaux, seront présentées et expliquées (Annexe 2).

Pour commencer, les niveaux de référence consolidés des étudiants dans ce groupe seront analysés (Annexe 1). Dans le cadre de cette analyse, la distribution des différents niveaux de référence au sein du groupe sera examinée. Ensuite, le lien entre le niveau final et les deux indicateurs de compétence grammaticale (« thème grammatical » et « verbes irréguliers ») sera étudié. La question est de savoir s'il existe dans tous les cas un lien fort entre la compétence grammaticale et le niveau de référence consolidé en grammaire-lexicologie. En outre, les fautes grammaticales et lexicales commises dans les formulations des définitions seront présentées et examinées. Pour conclure, le *benchmark* des mêmes étudiants en expression écrite sera exposé et comparé aux niveaux de compétence en grammaire-lexicologie.

En ce qui concerne la distribution des niveaux de référence en grammaire-lexicologie, on note qu'ils sont assez hétérogènes, car situés entre B1 et C1 (Annexe 1). Quatre étudiants ont obtenu le niveau B1 tandis que cinq personnes ont atteint C1. Les autres candidats ont acquis les niveaux intermédiaires, B1/B2, B2, B2+ et B2/C1, distribués de manière équilibrée. Il est frappant que les candidats qui ont obtenu C1 se distinguent par un bon niveau dans la compétence grammaticale, puisque tous les cinq ont un indicateur « thème grammatical » égal ou supérieur à 7 sur 10 (Annexe 1). Le deuxième indicateur de compétence grammaticale est également situé au-dessus de la moyenne dans ce groupe de candidats, à savoir, égal ou supérieur à 3 sur 6. Malgré leur bon niveau en grammaire, certains étudiants C1 commettent des fautes grammaticales et lexicales dans la formulation de définitions (Annexe 2). Ainsi des faiblesses sur le plan de la grammaire et du lexique peuvent être constatées

qui ne correspondent pas à ce niveau de compétence élevé. Les fautes grammaticales concernent le bon usage des verbes au temps du passé et au passif, l'emploi correct des articles ainsi que la congruence entre le sujet et le verbe (Annexe 2). Les fautes lexicales relèvent de l'orthographe, par exemple « an other » écrit en deux mots, et de l'usage des lexèmes qui ne conviennent pas dans le contexte donné.

Les étudiants qui se situent juste au-dessous, B2+, font également preuve d'une bonne habileté grammaticale parce qu'eux aussi ont tous obtenu un indicateur « thème grammatical » au-dessus de 7 sur 10. Le deuxième indicateur, « verbes (ir) réguliers » est pourtant, dans la plupart des cas, plus bas dans ce groupe d'étudiants que dans le précédent. Les fautes de grammaire et de lexique commises au niveau B2+ ressemblent à celles au C1. Sur le plan grammatical, il s'agit de l'usage des articles et de la congruence entre le sujet et le verbe (Annexe 2). Les nouveaux aspects grammaticaux qui posent des difficultés concernent l'usage des terminaisons et des pronoms relatifs : *« *de- in denominal stands for 'who comes from a noun'* » (Annexe 2). Quant au lexique, les fautes dans ce domaine de compétence n'apparaissent pas à ce niveau-là.

Pour ce qui est du niveau B2, les fautes sont plus nombreuses. Concernant la grammaire, il s'agit, en partie, des erreurs faites aux niveaux supérieurs, à savoir, de la non-congruence entre le sujet et le verbe, l'usage incorrect des articles et des pronoms relatifs ainsi que le choix incorrect des temps au passé. Les nouvelles fautes grammaticales au niveau B2 concernent notamment la malformation des formes du passé des verbes irréguliers, par exemple, «**it sweeped round the grassy curve…* » (Annexe 2). Deux autres erreurs non rencontrées aux niveaux supérieurs sont la formation incorrecte du gérondif - « **Compounding is when you create a new noun by added two others* » - ainsi que l'emploi incorrect des prépositions : « *[…] added affixes at a word » (Annexe 2). Sur le plan lexical, les expressions mal construites « *at the contrary » ainsi que l'emploi incorrect des pronoms interrogatifs surgissent : « *how he was going to the market ». En outre, on rencontre des problèmes d'orthographe (« adjectiv ») et l'usage de mots inexistants, comme « *basical » (Annexe 2).

Au niveau B1/B2, il y a un nombre considérable d'erreurs de grammaire. Certaines sont les mêmes qu'au B2, en l'occurrence, la non-congruence entre le sujet et le verbe, le choix inapproprié du temps au passé, la formation erronée des formes verbales au passé, l'emploi incorrect des prépositions, des articles et des pronoms relatifs. La construction incorrecte des phrases apparaît chez plusieurs étudiants pour la première fois: « *'Base' and 'stem' are identical *is case of there is* only one affixation to the base ». Sur le plan lexical, les défauts concernent l'emploi de lexèmes incorrects - « **which have their own built of the past* » qui ont un sens similaire à celui recherché ou alors l'utilisation de mots véhiculant un sens complètement différent : « **I was tall that this road is in summertime excellent* » (Annexe 2). Certains mots utilisés existent bien en anglais, mais ils sont ici écrits avec une orthographe française: « *remarquable ». Un autre défaut lexical au niveau B1/B2 est que les formulations de certains étudiants sont très peu rédigées.

Au niveau B1, les fautes de grammaire et de lexique se multiplient. En grammaire, un grand nombre d'erreurs se manifeste par des indicateurs encore plus bas qu'en B1/B2. L'indicateur « verbes (ir)-réguliers est à zéro chez presque tous les candidats, l'indicateur « thème grammatical » est plus bas chez les étudiants B1. Les fautes de grammaire ressemblent à celles rencontrées au B1/B2, à savoir, la non-congruence entre le sujet et le verbe, l'usage incorrect d'un temps du passé, la construction erronée des formes verbales au passé ainsi que l'emploi incorrect des pronoms relatifs. Les nouveaux phénomènes grammaticaux qui font l'objet d'erreurs en B1 concernent les déterminants possessifs : « *Each verb has his proper terminaison » - ainsi que l'usage erroné d'une préposition à la place d'une conjonction : « * like he's go to the market ».[1]

Dans le domaine lexical, les erreurs se concentrent notamment sur l'usage de mots dans une classe lexicale qui n'est pas la leur, par exemple « dinner » en tant que verbe- *« *Jack was dinnered when the ogre hit on the door »* - ou bien « *to gift »: « *It's very impolite to check at something you've been gifted ». Une faiblesse lexicale plus répandue en B1 concerne le calque des mots anglais sur le français : « **Stem : you don't change the word, you add an affix without making a modification oft he word* »; « **Each verb has his proper terminaison* ».

En ce qui concerne l'expression écrite, le niveau relevé dans cette compétence ne coïncide pas, la plupart du temps, avec celui de lexicologie-grammaire. La première raison à cela est que l'expression écrite est une compétence communicative productive qui se distingue dans une large mesure des compétences lexicale et grammaticale, qui sont des habiletés linguistiques. La deuxième raison est que la production textuelle était de nature essentiellement métalinguistique : il s'agissait d'analyser le fonctionnement de la langue. Le discours demandé était un discours de spécialité d'une relative abstraction et d'une incontestable technicité que les étudiants les plus sérieux avaient parfaitement intégrées mais que d'autres n'ont pas pris la peine d'assimiler. D'où l'abondance des calques opérés à partir du français. Pour toutes ces raisons, il n'est pas étonnant que peu d'étudiants aient obtenu le même niveau en lexicologie-grammaire (langue générale) et en expression écrite (métalangue). L'écart n'est cependant jamais marqué et dans seulement quelques cas très prononcé. Personne n'a par exemple atteint B1 dans une compétence et C1 dans une autre. Ceci montre que malgré la nature distincte de ces deux compétences elles sont liées puisque la maîtrise de grammaire et de lexicologie a un impact positif sur le niveau de compétence en expression écrite.

4.1.2 Groupe M2 (Sciences du Langage)

Le groupe réunit seize étudiants en sciences du langage, inscrits en deuxième année du Master ReLAI (sciences du langage). Il ne s'agit pas d'anglicistes. Tous ont suivi un enseignement hebdomadaire d'anglais de spécialité: 1h en présentiel, centrée sur la compréhension et l'expression orales, et 1h à distance, composée d'activités guidées de compréhension orale ou écrite ainsi que d'expression écrite. L'attribution d'un niveau de référence à chaque étudiant, avant la passation du test, repose sur les 4 activités exposées ci-dessous.

La première activité est constituée par la performance orale évaluée lors d'une prise de parole en continu.[2] Dans le cadre de cette tâche, l'étudiant-e présente un power point en anglais d'une durée de 5 à 10 minutes, portant sur un sous-domaine de la linguistique ou sur un centre d'intérêt personnel en sciences du langage. Pour préparer cette présentation, il est demandé de travailler à partir

des ressources en ligne de la *Linguistics Society of America*[3]. Les thèmes du menu « What is Linguistics ? » doivent être respectés et la phraséologie réutilisée. La prononciation peut être vérifiée à l'aide du dictionnaire en ligne *Collins*[4] et du système d'oralisation de l'écrit *Ivona.*[5] Les thèmes choisis par les étudiants pour leur présentation sont indiqués dans l'annexe 4. La situation de communication est similaire à celle d'un enseignant présentant une problématique lors d'un séminaire. Il s'agit donc d'une tâche authentique, utilisée pour évaluer une compétence orale spécifique, décrite dans le CECRL. Les critères retenus pour attribuer un niveau de référence sont formels: (i) correction grammaticale ; (ii) respect de la phraséologie du milieu professionnel ; (iii) clarté et précision articulatoires (réalisation des phonèmes, projection de la voix, accentuation); (iv) aisance et fluidité générales.[6] Un niveau global est attribué, par la doctorante et par l'instructeur du cours, en double aveugle. L'évaluation de la doctorante, qui ne connaît pas les étudiant-e-s, est étroitement liée aux descripteurs du CECRL (CECRL 2005 : 28). Les critères retenus par la doctorante sont les aspects qualitatifs de l'utilisation de la langue parlée : l'étendue, la correction, l'aisance, l'interaction et la cohérence (Annexe 5).[7] On constate qu'une partie des critères utilisés par la doctorante et par l'instructeur du cours sont les mêmes, à savoir, la correction grammaticale ainsi que l'aisance. En dehors du recours à certains critères différents, l'évaluation de l'instructeur qui connaît les étudiants, s'appuie davantage sur des éléments de connaissance et d'expérience antérieurs. Les différents critères respectés lors de l'évaluation des présentations orales expliquent le décalage des niveaux de référence attribués à la plupart des étudiants.[8] (Annexe 6). A quelques exceptions près, le niveau attribué par la doctorante est plus élevé que celui assigné par le professeur. Pour montrer l'usage des critères retenus pour l'évaluation, la manière d'attribuer les niveaux de référence sera décrite, de façon exemplaire. Le niveau A2 est assigné en raison des faiblesses par rapport à plusieurs critères, en l'occurrence, une prononciation et une intonation problématiques, gênant la compréhension. Les autres défauts à ce niveau-là se trouvent sur le plan de correction grammaticale, parce que les exposés ainsi évalués présentent un grand nombre d'énoncés souvent agrammaticaux et très influencés par le français (calques) : **I* *am* *in France since 5 years* (fr.: Je suis en France depuis 5 ans) ; **I study* *language science* (fr.: sciences du langage) at *Bordeaux* (fr.: à Bordeaux) ; **One*

years ago… (Annexe 7). Une autre insuffisance langagière relevée en A2 concerne l'aisance, parce que la construction des énoncés demande beaucoup d'efforts aux étudiants.

Au niveau B1, les points faibles des participants concernent aussi, dans certains cas, plusieurs critères. Un des exposés, par exemple, se distingue par une mauvaise articulation, par le manque d'autonomie par rapport aux notes aussi bien que par une mauvaise prononciation, par exemple, par des fautes de prononciation d'outils grammaticaux élémentaires comme *also*. Cependant, le niveau B1 est parfois assigné en raison d'un grand défaut dans un seul domaine, comme la présence de beaucoup de fautes de grammaire de première gravité : **I wasn't able to undersood* ; * *They didn't went to Japan* ; * *They had to listened to audio stimuli* ; * *When your boss ask you…* ; * *After listen to stimuli (Annexe 7).*

Au niveau B1/B2, deux cas de figure coexistent. Dans le premier, l'exposé présente des défauts par rapport à une seule qualité, par exemple, de nombreuses erreurs de prononciation, d'articulation de phonèmes ou de déplacements d'accents : *language, native, allow, areas.* En dehors de ces carences, l'exposé est globalement conforme aux autres critères d'évaluation. Le deuxième cas de figure se manifeste par la présence de faiblesses au regard de plusieurs critères, mais qui sont plutôt légères. Il arrive qu'un exposé soit parfois difficile à comprendre et présente des calques du français.

Au niveau B2, les points forts dans les exposés excédent les faiblesses. Pour donner un exemple, un manque de clarté dans l'articulation est contrebalancé par une aisance à communiquer et par une bonne fluidité. A partir du niveau B2+, on s'attache essentiellement aux points forts de la présentation orale. Ainsi, l'exposé B2+ se distingue par une bonne fluidité et par la richesse de son vocabulaire.[9] Au niveau C1, les présentations réunissent un grand nombre d'atouts, en l'occurrence, une accentuation correcte, une très bonne fluidité, une articulation claire ainsi qu'un lexique étendu. Au niveau C2, la quantité de points forts dans l'exposé s'élargit encore et concerne des aspects très divers: l'articulation, la prononciation, la correction grammaticale aussi bien que l'étendue du vocabulaire.

La deuxième activité effectuée par les étudiants est l'interaction orale. Au même titre que la présentation, celle-ci est également menée dans le but d'évaluer la performance orale.[10] Dans le cadre de cette tâche, les étudiant-e-s se mettent en binôme ou en trinôme, en situation d'interview fictive dans le studio d'une radio universitaire. Cette interview doit durer dix minutes environ. La fiction est celle de l'accueil sur le campus d'un-e linguiste étranger-e. Les sujets à aborder lors de l'interview sont l'identité, les centres d'intérêt, les recherches en cours et les expériences du linguiste invité. Les rôles entre *interviewer* et *interviewee* sont échangés dès que la première interview est terminée. Il a été demandé aux étudiants de travailler en amont sur des exemples écrits et oraux authentiques (en sciences du langage), de préparer leurs grilles d'interview ainsi que les modes de présentation de soi lorsqu'on est invité, de répéter l'ensemble au moins une fois avant de se produire devant la classe. Les étudiants ont le droit d'utiliser leurs notes, notamment la personne qui interroge. Cependant, il est interdit de lire les questions et les réponses intégralement. Les critères d'évaluation sont identiques à ceux de l'activité précédente qui est l'expression orale en continu. On y ajoute seulement un cinquième critère: l'autonomie par rapport aux notes.[11] En ce qui concerne l'attribution des niveaux, celle-ci sera expliquée au travers d'exemples. On observe deux cas de figure. Certains candidats ont obtenu un même niveau de référence pour la présentation et pour l'interaction orale. D'autres, en revanche, ont atteint deux niveaux de référence différents pour ces deux performances. Il faut noter que les étudiants ayant un niveau de référence identique pour les deux performances orales ne montrent pas toujours les mêmes faiblesses dans les deux activités. Pour donner un exemple, l'étudiant qui, en présentation en continu, a obtenu B1/B2 à cause de ses nombreuses erreurs de prononciation, comme l'articulation de phonèmes et les déplacements d'accents, fait de multiples erreurs de grammaire et des calques en interaction orale : * *pass an exam* (pour dire *sit an exam*), **my researches* (pluralisé comme en français, « mes recherches ») ; * a *master degree*, * *postgraduate study* (oubli du pluriel) ; * *a really interesting works* ; prononciation défectueuse de * *answered* (le 'w' prononcé), **bomb* (le 'b' prononcé) (Annexe 7). En revanche, certains apprenants qui ont obtenu le même niveau pour les deux performances, montrent le même genre de défauts dans les deux cas. A titre d'exemple, ceux qui commettent des fautes de prononciation

dans la présentation, montrent systématiquement cette faiblesse dans l'intervie*w*, quel que soit le niveau de référence. [12] Ce constat est valable également pour les qualités langagières démontrées puisque beaucoup de candidats révèlent les mêmes points forts langagiers dans les deux performances orales. Un étudiant qui communique avec une bonne accentuation et de la fluidité, une articulation claire et une variété de vocabulaire dans sa présentation en continu, déploie ces mêmes qualités dans l'interaction orale (Annexe 7).

Il n'est pas rare cependant que des candidats ayant des niveaux de référence distincts dans les deux activités exhibent des qualités et défauts similaires. Pour donner un exemple, un manque de clarté dans l'articulation mais une bonne fluidité et une bonne manière de communiquer se manifestent à la fois dans l'exposé et dans l'interview d'un étudiant. Tandis que ces caractéristiques lui valent l'attribution d'un niveau B2 pour sa performance en continu, elles donnent accès au niveau C1 pour l'interaction¸ parce qu'une qualité supplémentaire est décelée dans cette activité, comme une richesse et une précision particulières de vocabulaire (Annexe 7). Cet exemple montre que les faiblesses aussi bien que les points forts des candidats se reproduisent dans leurs diverses performances, mais que souvent de nouveaux défauts ou qualités se manifestent à l'occasion d'une nouvelle activité langagière, ce qui a une influence sur l'attribution du niveau de référence.

La troisième activité qui sert à attribuer un niveau de référence à chaque étudiant avant la passation de la version pilote du POSILANG est la production écrite en classe, lors de la prise de notes, ainsi que la rédaction de courtes réponses de compréhension orale. Cette activité est complexe car composée du visionnage de trois vidéos: (1) *Five reasons for doing linguistics*[13] (David Crystal, 8'51) avec, en consigne finale, « *Is linguistics relevant to today's world? Practical usefulness and profitability »;* (2) *How schools kills creativity*[14] (Sir Ken Robinson, 19'25) et (3) *The Stanford Prison Experiment*[15] (Philip Zimbardo, 13'40). Il s'agit de tâches intégrées parce que le niveau de deux compétences langagières est évalué: la compréhension orale et l'expression écrite. La compréhension orale est sollicitée lors du visionnage des vidéos tandis que l'expression écrite est nécessaire lors de la réponse aux questions. Le niveau maximal à atteindre lors de cette tâche est le niveau B2 parce que pour l'accomplir, il est requis d'avoir ce

niveau en compréhension orale et en expression écrite. Les trois tâches de production écrite, accomplies à la suite du visionnage des vidéos, résultent en l'attribution de niveaux relativement uniformes (Annexe 8). La première tâche, qui inclut deux questions portant sur la vidéo *Five reasons for doing linguistics,* montre une prévalence du niveau B1. Le niveau inférieur*,* A1*,* a été atteint par deux personnes uniquement et A2 par un seul candidat (Annexe 8). Il faut souligner que le contenu aussi bien que la correction langagière sont évalués. Pour donner un exemple de la performance de l'étudiant catégorisé A1, on note une absence totale de réponse à la première question (*Five reasons for doing linguistics).* En outre, des fautes de grammaire et de lexique se trouvent dans la réponse à la deuxième question, par exemple, la non-congruence entre le sujet et le verbe *(*every aspect <u>need</u>*) et l'emploi de mots inventés *(* specially).* Une autre performance évaluée par le niveau A1, ne contient pas de réponse à la première question non plus et une réponse incomplète à la deuxième question. De plus, plusieurs fautes grammaticales et lexicales ont été commises. La production écrite au niveau A2 se distingue par un grand manque sur le plan du contenu puisque la deuxième question est traitée de manière trop générale et ne contient donc pas l'information essentielle. En revanche, les défauts langagiers sont minimes (Annexe 9). En ce qui concerne le niveau B1, la plupart des faiblesses se situent au niveau du contenu et non pas au niveau de la langue. Ainsi, les réponses données sont incomplètes et concernent presque toujours la deuxième question. Dans certaines copies situées en B1, des fautes grammaticales et lexicales apparaissent (Annexe 9). En ce qui concerne le niveau B2, les réponses fournies sont relativement complètes, c'est-à-dire, couvrent l'information importante. Les erreurs langagières sont moins nombreuses qu'aux niveaux précédents (Annexe 9).

La deuxième tâche prévoit le visionnage de la deuxième vidéo, *How schools kill creativity,*[14] et la réponse aux quatre questions suivantes: *i) What are social attitudes to educationists, according to Ken Robinson? ii) Every education system has a hierarchy of subjects. Explain briefly. iii) Who designed the education system and how does it show? iv) What is found missing in today's education systems?* Les niveaux de référence attribués pour cette tâche s'étendent de A1 à B1 et sont distribués de manière relativement uniforme,

comme dans la première vidéo. La plus grande partie des étudiants obtient le niveau B1, comme pour la première tâche (ANNEXE 8). A1 est attribué à une production écrite qui fait l'impasse totale sur la première question et ne fournit que des réponses très lacunaires aux trois autres questions. Au niveau A2, deux réponses sur quatre manquent d'éléments d'information importants, tandis qu'une autre au moins est incomplète. Les performances au niveau B1 relèvent de deux cas de figure : ou bien la moitié des réponses ne contient pas les informations attendues (car contenues dans la vidéo) tandis que l'autre moitié les fournit toutes, ou alors une réponse est partielle et les autres sont presque complètes (Annexe 9).

En dehors des faiblesses constatées sur le plan des contenus, un nombre de fautes langagières se trouve dans les réponses fournies dans le cadre de cette tâche. Concernant le lexique, au niveau A1, plusieurs lexèmes sont utilisés qui relèvent d'une classe lexicale inadéquate, par exemple **of the difference intelligence* (au lieu de : *of the different intelligence*). Au niveau de la grammaire, deux erreurs surgissent. La première consiste en une omission de la particule infinitive, * *than learn mathematics* (Annexe 9). La deuxième erreur qui apparait également aux niveaux A2 et B1 est l'usage de noms au singulier au lieu du pluriel: * *above creative subject* (Annexe 9). Ce n'est pas le seul type d'erreur de grammaire en A2. Les autres sont les suivantes: i) la malformation du pluriel des noms irréguliers: * the childrens ; ii) l'emploi incorrect de l'article: * the same hierarchy of <u>the</u> subject ; iii) la non-congruence entre le sujet et le verbe: * it have ; iv) l'emploi des temps incorrects * *you can see […] and said …* (au lieu de : *say).* Sur le plan lexical, au niveau A2, certains mots qui n'existent pas sont employés : * *prior* (au lieu de : priority) et des fautes d'orthographe apparaissent, par exemple, * *statu,* * *proces* (Annexe 9). Au niveau B1, des fautes d'orthographe sont toujours commises, par exemple, * *appriciate,* * *speciall* (Annexe 9). En outre, certaines locutions inexistantes sont utilisées: * *to complete studies*. Sur le plan de la grammaire, une grande partie des erreurs sont identiques à celles commises en A2, voire en A1 (Annexe 9). Ce sont les suivantes : i) l'usage des noms au singulier au lieu du pluriel * *industrial <u>need</u>;* ii) l'emploi erroné des articles : * *in <u>the</u> society* ; iii) la non-congruence entre le sujet et le verbe: * *every system <u>have</u>.* Les nouvelles erreurs grammaticales sont :

i) l'omission des articles: * process of University entrance ; ii) l'emploi de prépositions incorrectes : * respect to them ; iii) une erreur de forme au participe passé : * was design ; iv) l'usage du présent au lieu du gérondif : * It's focused on go to university (Annexe 9). Le nombre plus élevé d'erreurs grammaticales en B1 s'explique par le fait qu'à ce niveau-là, plus de réponses élaborées sont fournies. Plus les réponses sont développées, plus le nombre et la variété des fautes augmentent. Pourtant, le nombre de fautes absolu est moindre en B1 qu'aux niveaux inférieurs.

La troisième tâche de compréhension de l'oral comporte le visionnage de la vidéo *The Stanford prison Experiment* (Philip Zimbardo) et la réponse aux trois questions : *a) What was the Stanford prison experiment about ? b) Describe what happened. Who was involved in this experiment? c) What happened exactly and why did it eventually have to be stopped?* Les résultats s'avèrent ici meilleurs que pour les deux vidéos précédentes. La majorité des réponses valident le niveau B2 (11 sur 16). Le reste des candidats se répartit sur deux niveaux, A2 et B1, avec trois étudiants en B1 et deux en A2. Comme pour les deux autres tâches de ce type, le contenu autant que la correction langagière ont été pris en compte. Les copies évaluées comme A2 ne fournissent aucune réponse à une ou deux questions. Les performances situées en B1 ne donnent pas de réponse à l'une des trois questions.

Au niveau A2, des erreurs morpho-lexicales apparaissent. Il s'agit de fautes d'orthographe, de l'emploi de noms composés fantaisistes ainsi que de l'emploi d'un nom au lieu d'un adjectif (Annexe 9). En grammaire, les erreurs sont plus nombreuses et plus variées. La non-congruence entre le déterminant et le nom ou entre le sujet et le verbe apparaissent, par exemple, * *the number replace the name* (Annexe 9). En outre, il faut noter l'usage incorrect des catégories grammaticales et lexicales suivantes : l'article défini, les pronoms personnels et plusieurs formes verbales (Annexe 9). En B1, le même type de fautes lexicales et grammaticales qu'au niveau inférieur est commis, mais leur occurrence est moindre. Sur le plan lexical, les fautes relèvent uniquement de l'orthographe. Sur le plan grammatical, les erreurs sont aussi moins variées en B1 qu'au niveau inférieur, mais recouvrent néanmoins plusieurs phénomènes : l'usage incorrect des articles, l'emploi de formes verbales incorrectes et

l'utilisation de formes incorrectes des adjectifs irréguliers (Annexe 9). En B2, la plupart des erreurs décelées se retrouvent également dans les niveaux inférieurs. Concernant le lexique, ce constat vaut pour deux des trois types d'erreurs commises : les fautes d'orthographe et l'emploi de locutions non-existantes : * *to make harrassment.* Seul un type d'erreur se situe en B2 uniquement : l'usage d'un mot relevant d'une classe lexicale inadéquate, par exemple, l'emploi d'un adjectif au lieu d'un adverbe (Annexe 9). Sur le plan grammatical, la grande majorité des erreurs sont également présentes aux niveaux inférieurs, à savoir, la non-congruence entre le sujet et le verbe ou entre le sujet et le pronom possessif, l'emploi de formes verbales incorrectes, l'usage incorrect des articles, y compris l'omission de l'article défini. Quant aux autres fautes de grammaire, la plupart d'entre elles se retrouvent aux différents niveaux de compétence des deux autres tâches de production écrite. Il s'agit de l'emploi incorrect des prépositions, de l'utilisation inadéquate des pronoms relatifs ainsi que de l'emploi des noms au pluriel au lieu du singulier et vice versa (Annexe 9). Seul un petit nombre des fautes grammaticales est propre au niveau B2 de la tâche en question : une structure incorrecte de phrase et l'omission du sujet (Annexe 9).

La description effectuée montre que la plupart des fautes commises ne sont pas propres à un niveau de référence, mais se répètent aux différents niveaux de compétence. Pourtant, même les types d'erreurs qui ne sont pas commis aux différents niveaux de compétence dans une tâche donnée se retrouvent à un niveau inférieur ou supérieur dans une autre tâche focalisée sur la même compétence. Ce constat illustre et corrobore le résultat de la recherche sur l'acquisition d'une langue seconde. Comme indiqué dans notre deuxième chapitre, l'existence d'un lien entre les items spécifiques de grammaire et les niveaux communs de référence du CECRL n'est pas validé par la recherche, bien que la plupart des manuels d'apprentissage des langues vivantes l'affirment (Westhoff 2007 : 678). Outre l'absence de lien avéré entre les phénomènes grammaticaux et les niveaux de référence, la répétition de fautes lexicales à différents niveaux de compétence dans les trois tâches de production écrite permet également de conclure à la non-existence d'un lien entre les aspects lexicaux et les niveaux de référence distingués dans le CECRL.

4.2 Analyse du test POSILANG

4.2.1 Conception de l'échelle d'évaluation

L'échelle d'évaluation de POSILANG est répartie sur cinq niveaux de référence: A1, A2, B1, B2 et C1. Ces niveaux de référence sont à atteindre dans chacun des trois domaines de compétence langagière : en compréhension orale, en compréhension écrite et en expression écrite.[16] Pour valider un niveau de compétence, il faut, au minimum, acquérir 60 % du score total. Ce dernier s'élève à 6 points pour chaque domaine de compétence aux cinq niveaux évoqués, à l'exception du domaine de compréhension de l'oral aux niveaux B2 et C1 où le score maximal égale cinq points (Annexe 11). Le pourcentage évoqué implique qu'il est requis d'acquérir 3,5 ou 3 points dans chaque compétence fonctionnelle à chaque niveau, en fonction du score maximal[17] (Annexe 11). Pour valider un niveau de compétence dans notre test, il faut obtenir ce score minimal dans chacun des trois domaines de compétences à tous les niveaux. Afin de valider chacun des niveaux A1, A2 et B1, il est obligatoire d'accumuler, au minimum, 11 points sur 18, le score maximal, tandis qu'il faut en acquérir au moins 10 sur 17 pour valider B2 et C1 (Annexe 11). Les points acquis aux niveaux inférieurs sont additionnés au score gagné au niveau à valider pour obtenir le score total d'un candidat. Il faut souligner que les types de niveaux attribués dans un certain domaine de compétences se distinguent de ceux assignés en tant que niveau global. Ces derniers peuvent être les niveaux dits « critériés » aussi bien que ceux dénommés « avancés », à savoir: A1, A2, A2+, B1, B1+, B1/B2, B2, B2+, B2/C1 et C1. En revanche, les niveaux à valider dans un domaine de compétences sont uniquement « critériés »: A1, A2, B1, B2 et C1 (Annexes 10, 11). Cette différence de types de niveaux s'explique par le fait que, pour obtenir un certain niveau global, il faut acquérir le même niveau en trois domaines de compétences. Dans le cas contraire, les niveaux acquis dans chacun des trois domaines de compétences sont combinés lors du calcul du niveau global.

La distribution des points dans notre test de positionnement est corrélée au nombre de tâches dans chaque domaine de compétences. Puisque la réponse correcte à chaque tâche correspond à l'attribution d'un point, on obtient le score maximal dans chaque domaine de compétences pour le traitement

correct de toutes les tâches. Lorsque les tâches sont complexes, chaque tâche correspond à un certain pourcentage de point (Annexe 12). Par exemple, la tâche 2 du domaine « compréhension de l'écrit » au niveau A1 se compose de deux tâches au format QCM. Le score attribué pour chaque bonne réponse est 0,5. La tâche 6 du domaine « expression écrite », au même niveau, est composée de quatre tâches dont chacune correspond à 0,25 points (Annexe 12).

4.2.2 Répartition des candidats selon les niveaux

La plupart des candidats se répartissent sur quelques niveaux de référence de l'échelle d'évaluation, tandis que d'autres niveaux ne sont attribués à personne. En ce qui concerne les étudiants de première année (L1 LLCE anglais), les performances de la grande majorité d'entre eux se situent aux niveaux B2+ (7), B2/C1 (7) et C1 (8). Le reste des candidats se répartit sur trois niveaux: B1 (1), B1/B2 (1) et B2 (4). On note que dans ce groupe, aucun étudiant ne se voit attribuer le niveau A1, A2, A2+ ou B1+. Parmi ceux-ci figurent deux types de niveaux évoqués ci-dessus, « critériés » et « avancés » (Annexe 10).

En ce qui concerne le groupe de master (M2 ReLAI, sciences du langage), la répartition des candidats sur les différents niveaux de compétence y est plus fluide que dans le cours de Licence. Cinq niveaux, A2+, B1, B1+, B1/B2 et B2, ont été validés une fois chacun. B2+ a été validé trois fois et B2/C1 deux fois. Le niveau C1 est atteint par cinq étudiants. Comme dans l'autre groupe, tous les étudiants dépassent A1 et A2. (Annexe 10)

4.2.3 Comparaison avec le niveau attribué avant la passation du test

Pour établir que notre test est fiable, il faut mettre en regard les résultats obtenus par les candidats avec les niveaux de référence. Pour ce faire, il faut déterminer le nombre d'étudiants dont le niveau établi en amont coïncide avec celui attribué via POSILANG et, inversement, le nombre d'étudiants dont les résultats divergent. Lorsqu'un écart est constaté, il est nécessaire d'en déterminer l'amplitude. Un écart entre B2+ et C1, par exemple, est minime. Il ne remet pas vraiment en cause les résultats, alors qu'un décalage plus important demande une explication.

Dans le groupe des étudiants L1, 6 personnes sur 25 se voient attribuer exactement le même niveau lors des deux procédures d'évaluation, ce qui représente 24%.[18] Plus de la moitié des étudiants (60%) atteignent des niveaux très proches, par exemple B2+ (référence) et C1 (POSILANG). Au total, 84% des candidats se positionnent à des niveaux très proches ou identiques. Seules 4 personnes sur 25 (16%) se situent à un niveau nettement décalé (POSILANG) par rapport à leur niveau de référence initial. Ce bilan est un indice de la validité de notre test de positionnement (Annexe 13).

Concernant le groupe M2 en sciences du langage, la comparaison des résultats obtenus trace une image différente. Un seul candidat valide exactement le même niveau de référence, ce qui représente un taux de 7%. Six étudiants sur quinze atteignent cependant des niveaux très proches (40 %). Huit candidats sur quinze (53 %) valident deux niveaux qui divergent plus que minimalement l'un de l'autre, par exemple B2 et C1. Le pourcentage beaucoup plus élevé des candidats atteignant deux niveaux divergents dans ce groupe peut s'expliquer par les conditions plus difficiles de passation du test.[19] Néanmoins, le fait que la moitié environ des étudiants valide deux niveaux limitrophes lors de deux procédures d'évaluation est un indice non négligeable de la validité de POSILANG**.**

4.2.4 Analyse des tâches

Il est d'abord essentiel de déterminer la valeur de facilité pour chaque tâche, c'est-à-dire le pourcentage de candidats ayant répondu correctement (Alderson, Clapham & Wall 1995 : 82). En second lieu, le nombre de tâches non fiables, dans tout le test et par niveaux de compétence doit être présenté. Troisièmement, il importe de calculer l'index de discrimination pour chaque tâche, afin de déterminer la capacité de ladite tâche à différencier des candidats ne possédant pas le même niveau de compétence (Alderson, Clapham & Wall 1995 : 81). Enfin, il convient de se pencher sur la fiabilité de chaque tâche. Cette analyse se fera en comparant les résultats obtenus dans les deux groupes (Annexe 15).

Le test se compose de 88 tâches au total. Les trois niveaux inférieurs en comportent 18 et les deux niveaux supérieurs 17. Aux niveaux B2 et C1, se

trouvent 5 tâches centrées sur l'évaluation de la compréhension orale. Il y en a 6 dans ce domaine de compétence aux trois niveaux les plus bas. En ce qui concerne les deux autres domaines de compétence, la compréhension et l'expression écrites, 10 tâches sont proposées à tous les niveaux (Annexe 14).

4.2.4.1 Valeur de facilité

Pour ce qui est de la valeur de facilité (VF) des tâches contenues dans notre test, celle-ci est élevée aux niveaux A1 et A2, puisqu'elle dépasse 70 % dans les deux groupes (Annexe 15). Dans un grand nombre de tâches, la VF égale 1.0 ce qui implique que tous les étudiants l'ont réussi. Dans le groupe L 1, la quantité de tâches avec une VF maximale est plus élevée que dans le groupe M2: 10 contre 7. Au niveau A2, la VF est légèrement inférieure à celle du niveau A1. Par exemple, les valeurs 0,68 et 0,67 sont attribuées pour deux tâches centrées sur la compréhension de l'oral. Dans les deux groupes, les résultats obtenus sont très proches. Un écart important est décelé dans la première tâche de compréhension orale et la quatrième tâche de compréhension écrite (Annexe 15). Au niveau B1, la VF est aussi élevée qu'au niveau A2. Le niveau B1 se distingue du niveau A2 par le nombre de tâches que la totalité des candidats a réussi. Seules trois tâches ont une VF de 1.0 en B1, contrairement aux 14 tâches du niveau inférieur. Les différences entre les deux cours, L1 et M2 sont ici minimes. La VF des tâches en B2 connaît des variations importantes. Celle-ci s'étend de 0,57 à 1,0. La distribution des tâches avec différentes VF est relativement équilibrée. Il y en a trois avec une VF de 0,57 et deux avec une VF maximale. Les autres prennent des valeurs de facilité différentes. La divergence entre les deux cours est plus importante à ce niveau-là qu'en B1, mais celle-ci n'est pas considérable. La première et la troisième tâche relevant du domaine « expression de l'écrit » illustrent ce cas de figure (Annexe 15). Quant au niveau C1, les valeurs de facilité sont mixtes : pour certaines tâches, des valeurs élevées sont attribuées tandis que pour d'autres, elles sont basses. La VF maximale de 1.0 concerne une seule tâche effectuée par les étudiants de M2 : il s'agit de la troisième située dans le domaine la compréhension écrite (Annexe 15). Les différences entre les deux groupes sont plus marquées qu'aux niveaux inférieurs. Le décalage considérable en matière de valeur de facilité se situe

notamment dans le domaine « compréhension de l'oral », mais est également pertinent pour la dernière tâche centrée sur l'expression de l'écrit.

4.2.4.2 Tâches non fiables

Les tâches non fiables concernent surtout les étudiants du groupe L1. 12 tâches sont réussies par moins de 60 % des étudiants de ce dernier groupe, par rapport à seulement 9 tâches pour les personnes du groupe M2.[20] Dans les exemplaires du test remplis par les étudiants de L1, les niveaux A1, B1 et C1 comportent davantage de tâches non fiables. Il y en a moins en B2. Au niveau A1, une seule tâche est réussie par moins de 60 % des étudiants de L1 tandis qu'en B1, deux fois plus de tâches non fiables apparaissent chez les candidats de ce groupe, à savoir quatre par rapport à deux. En C1, cinq tâches sont réussies par moins de 60 % des candidats de L1, contre quatre en M2. En revanche, en B2, deux tâches non fiables se trouvent dans les copies des étudiants de L1, contre trois en M2. Les trois niveaux, A1, B1 et C1, confirment notre observation concernant une plus grande quantité de tâches non fiables dans les performances des étudiants de licence, tandis que B2 la remet en question.

La convergence entre les deux groupes de candidats est haute en matière de tâches non fiables. Sept des neuf tâches réussies par moins de 60% de candidats de M2 ne sont pas réalisées correctement par les étudiants de L1 non plus (Annexe 15). En B1, deux tâches sont concernées par ce fait : la deuxième évaluant la compréhension de l'oral et la troisième ciblant l'expression de l'écrit. En B2 également, une tâche n'est pas réussie par suffisamment de candidats de L1 ou de M2. Il s'agit de la cinquième tâche de compréhension écrite. En C1, la coïncidence en termes de tâches non fiables entre les deux groupes est particulièrement élevée : les quatre tâches non réussies par les étudiants de L1 ne sont pas correctement traitées non plus par un nombre adéquat d'étudiants M2 (Annexe 15).

4.2.4.3 Index de discrimination

En ce qui concerne l'index de discrimination (ID), on constate une répartition parallèle sur les cinq niveaux de référence dans les deux groupes d'étudiants, L1 et M2. Dans le groupe L1, aux niveaux A1 et A2, l'ID est relativement bas et

distribué de manière similaire. En A1, cette valeur est située entre 0,11 et 0,34. En A2, les montants ressemblent au niveau ci-dessous, à l'exception d'une ID de 0,56 pour la première tâche dans le domaine de la « compréhension de l'oral ». En outre, le nombre de « zéro » est plus élevé au A1, où cet ID est attribué pour douze tâches, contrairement à sept en A2. L'ID « zéro » apparaît avec une haute fréquence à deux niveaux inférieurs parce qu'il n'est pas possible de différencier suffisamment le tiers le plus fort du plus faible à ces niveaux-là. La raison à cela est que même les candidats les plus faibles ont atteint A1 et A2 (Annexe 16).

En B1, la quantité de « zéro » est la même qu'en A2, en raison de la validation de ce niveau même par les étudiants les plus faibles. Néanmoins, une discrimination plus élevée peut être observée entre ces deux tiers d'étudiants. Celle-ci se manifeste par la présence des ID plus hauts à ce niveau-là qu'en A2. Pour donner un exemple, le montant de 0,45 apparaît deux fois et celui de 0,34 trois fois (Annexe 16). Le niveau B2 différencie mieux les candidats car on y rencontre beaucoup moins d'ID à « zéro » qu'aux niveaux inférieurs (Annexe 16 a). Ceci s'explique par le fait que certains étudiants du tiers le plus faible n'ont pas validé le niveau B2, ayant atteint B1 ou B1/B2 (Annexe 10). Une meilleure discrimination se révèle également par une plus grande quantité de tâches avec des ID relativement élevés : les valeurs 0,56, 0,45 et 0,34 sont chacune attribuées à deux tâches. Au niveau C1, cette tendance se confirme. Celle-ci se manifeste par la présence d'une seule tâche ne permettant pas d'établir de différence entre les deux tiers des candidats. L'ID est relativement élevé dans un grand nombre de tâches : le montant de 0,34 apparaît trois fois, ceux de 0,45 et même de 0, 78 sont attribués (Annexe 16 a).

Les résultats obtenus en M2 sont identiques, même s'ils sont distribués de manière différente (Annexe 16 b). Les montants sont bas et le nombre de « zéro » élevé, puisque, aux niveaux A1 et A2 il est impossible de différencier nettement le tiers le plus fort du tiers le plus faible. La raison à cela est que, comme en L1, les étudiants ont tous validé ces deux niveaux inférieurs. En B1, les ID sont plus élevés: les montants de 0,33 et de 0,4 apparaissent deux fois et celui de 0,5 trois fois. En B2, les tâches ont un meilleur pouvoir de discrimination entre les deux tiers. Le nombre de zéro est bien moindre à ce niveau-là qu'au-

dessous, avec seulement deux tâches concernées (Annexe 16 b). En outre, le montant de 0,3 est attribué deux fois, ceux de 0,5 et de 0,55 une fois. Au niveau C1, la discrimination est plus marquée entre le tiers le plus fort et le plus faible. La raison à cela est que les étudiants du tiers le plus faible ne valident pas ce niveau-là. Comme en L1, les ID sont beaucoup plus élevés en C1 qu`à tous les autres niveaux inférieurs: 2 x 0.8, 3 x 0.67 et 3 x 0.6. Et comme en L1 encore, une seule tâche ne permet pas d'établir de distinction entre les deux tiers concernés.

4.2.4.4 Coefficient de fiabilité

Il est possible de quantifier la fiabilité d'un test au moyen d'un coefficient de fiabilité. Ce dernier peut prendre une valeur comprise entre zéro et 1. Un coefficient égal à zéro implique qu'un test donne deux ensembles de résultats complètement déconnectés l'un de l'autre. En revanche, le montant 1 est un coefficient optimal parce qu'il suggère l'obtention de deux ensembles de résultats exactement identiques (Hughes 2003 : 39). Un coefficient de fiabilité qui s'établit à 1,0 n'apparaît que très rarement. Une fiabilité maximale n'en demeure pas moins un idéal à atteindre.

En ce qui concerne le groupe L1, le coefficient de fiabilité idéal est atteint par un seul candidat (Annexe 17). La précondition pour un tel montant est un niveau de compétence très élevé. Cependant, même le niveau de compétence C1 ne permet pas, la plupart du temps, d'atteindre le coefficient de fiabilité maximal. D'autres étudiants dans ce groupe qui valident le C1 ont un coefficient de fiabilité au-dessous de 1.0 parce qu'ils obtiennent un score plus élevé dans la première partie de POSILANG que dans la seconde. Ceci s'explique par une plus grande facilité même pour les candidats les plus forts de répondre correctement aux tâches de la première partie et de gagner plus de points pour celles-ci. Quant à la distribution des montants pour la passation de notre test parmi les étudiants de licence, la plupart obtiennent des montants élevés, entre 0,8 et 0,95. La raison à cela est que la grande majorité des étudiants obtiennent deux scores similaires pour les deux parties du test, mais toujours plus élevé pour la première partie. Le coefficient de fiabilité est inférieur à 0,75 seulement chez deux personnes. Le montant minimal est situé à 0,56.

Dans le groupe M2, les coefficients de fiabilité acquis par les candidats lors de la passation du POSILANG sont plus bas que ceux obtenus par les étudiant-e-s en L1. Les candidats qui obtiennent des montants au-dessous de 0,75 sont beaucoup plus nombreux que dans le groupe L1, même si la moyenne s'élève à 0,75. Certains étudiants obtiennent des montants proches de 1,0 également dans ce groupe, mais uniquement ceux ayant validé les niveaux B2+ et C1. Le coefficient de fiabilité idéal n'est acquis par aucun candidat issu du groupe M2, y compris ceux ayant validé le niveau C1 (Annexe 18). Le montant minimal est nettement inférieur à celui dans l'autre groupe, à savoir 0.46 (Annexe 18).

4.2.4.5 Conclusion

L'évaluation des compétences en langues est l'objet de l'intérêt croissant des chercheurs, des responsables de cursus, des enseignants, des étudiants et de la société en général (Brown 2010 : 1x). Cet intérêt a suscité l'apparition d'un nombre exponentiel de tests de langue. En fonction de leurs objectifs, ces tests se répartissent en sept types : en tests de positionnement, d'acquisition et de progrès, de compétence, de certification, diagnostiques et enfin tests d'aptitude. Malgré la multiplication des dispositifs disponibles sur le marché, un test comme POSILANG, développé dans le cadre d'un projet régional, présente des avantages.

Comme son prestigieux aîné DIALANG, POSILANG est à la fois gratuit et adossé au *Cadre européen commun de référence pour les langues* (CECRL). Cet adossement est réel: il ne s'agit pas d'un rattachement stratégique, comme cela a pu être le cas pour le test américain TOEIC développé par le géant ETS GLOBAL. L'enjeu n'est pas de procéder à un « adossement » de circonstance, de se prévaloir d'un cadrage institutionnel de pure forme, ou encore d'effectuer la conversion opportuniste d'un score en niveau CECRL. Il s'agit plutôt d'intégrer *à la source* un maximum de paramètres du Cadre, de respecter, autant que faire se peut, l'esprit de l'approche communicative-actionnelle prônée, alors même que le format du test est contraint et que des réductions doivent être opérées dans les compétences testées. Idéalement, un dispositif comme POSILANG doit pouvoir se distinguer de l'offre existante par sa praticité et sa fiabilité, par son accessibilité et sa gratuité.

L'adossement de POSILANG au CECRL se traduit de façon très explicite par le respect des niveaux de référence: A1, A2, B1, B2 et C1. (Seul le niveau C2 n'est pas proposé). L'attribution d'un niveau global à chaque candidat sur l'échelle du CECRL, à partir des descripteurs institués, repose sur l'évaluation de trois domaines de compétence langagière: la compréhension orale, la compréhension écrite et, en mode parcellaire, l'expression écrite. Par ailleurs, un test comme POSILANG doit respecter le critère d'authenticité. Ce respect se manifeste par l'élaboration d'items se rapprochant autant que possible de la forme, des fonctions et des usages situés de la langue. Afin de vérifier l'adéquation formelle et fonctionnelle des items, il est utile de faire appel à l'expérience et à l'intuition des natifs en langue anglaise. Il faut également aspirer à l'interrogation de corpus, à l'observation des pratiques. Cela est essentiel lorsque le dispositif est monté en milieu exogène. Enfin, un test construit sur le schéma que nous avons retenu pour POSILANG doit s'efforcer de respecter le principe d'interactivité dans la mesure où il est élaboré localement, en tenant compte d'une situation d'évaluation particulière.

Ainsi peuvent être satisfaites les exigences de la recherche dans la discipline, généralement ignorées par des concepteurs de tests, pas toujours experts. Le résultat de cette manière de construire des tests est le plus souvent un dispositif simplifié, focalisé sur la seule correction formelle, qui devient emblématique de la maîtrise de la langue. A part le caractère lacunaire de l'évaluation ainsi opérée, on note le paradoxe d'une démarche décontextualisée, en rupture avec les stratégies d'acquisition développées en milieu scolaire. On sait en effet, que l'enseignement scolaire des langues est aujourd'hui moins fixé sur la grammaire et la perfection formelle qu'il ne l'était à l'époque structuraliste, qu'on s'est distancé de la norme imposée par le locuteur autochtone et qu'un renoncement à la fiction du bilinguisme parfait a été opéré. Or, à l'occasion d'évaluations sommatives ou de la passation de tests de positionnement, le candidat doit soudain se comporter en détenteur expert de la norme grammaticale, sans pouvoir témoigner d'autres capacités. Certes, le respect de cette norme peut faire partie des exigences pertinentes, comme c'est le cas dans un département de spécialistes de langue par exemple. Cependant, même en pareil cas, la seule compétence grammaticale ne saurait servir à accréditer

l'intégralité de la compétence en langue. Cela est plus vrai encore pour les spécialistes d'autres disciplines, qui peuvent avoir d'autres besoins : comprendre des textes dans leur spécialité, rédiger de courtes notes, faire un exposé. D'où la nécessité de construire localement des dispositifs de positionnement, en intégrant au maximum la situation d'évaluation spécifique (Hughes 2003 : 17). D'où l'utilité aussi de diversifier les compétences mesurées afin de pouvoir placer de façon plus précise les candidats sur plusieurs échelles de compétence.

Par ailleurs, il nous semble essentiel de fournir au candidat des critères d'attribution des scores précis et transparents. C'est ce que nous avons tenté d'opérer pour POSILANG, dans la situation d'évaluation précise que constitue le site universitaire bordelais.

De la conception d'un test de positionnement à sa diffusion, les étapes sont nombreuses et exigeantes, dès lors qu'on renonce au bricolage. Nous savons par expérience que les équipes pédagogiques confient souvent la réalisation de tests de langues à des individus isolés, qui se dévouent dans l'urgence, sans formation particulière en psychométrie, sans moyens logistiques avancés, pour répondre à des injonctions administratives ou pédagogiques de positionnement à des fins de pré-orientation, de répartition dans des groupes de niveaux, d'aiguillage vers des activités de soutien ou de remédiation. Un tel dispositif peut dépanner et même fonctionner temporairement, mais en aucune manière cela ne devrait suffire à le pérenniser. Les langues et les étudiants qui les pratiquent méritent mieux.

Une étape essentielle dans la mise au point d'un dispositif d'évaluation est celle du pré-test. Nous ne saurions trop insister sur ce point. Pour POSILANG, nous avons proposé une passation manuelle à deux groupes distincts d'étudiants. Nous avons soumis à ces derniers une version pilote au format papier + enregistrement sonore de POSILANG. Ont participé des étudiants de master 2 en sciences de langage et des étudiants en première année de licence d'anglais (L1). Les étudiants de master sont certes des spécialistes d'une autre discipline mais étudient la linguistique à un niveau avancé depuis au moins quatre ans, lisent des écrits scientifiques en anglais, parlent souvent plusieurs langues et de ce fait, ont des compétences plutôt élevées (à quelques exceptions

près). En revanche, les étudiants en première année de licence d'anglais, tout spécialistes qu'ils soient, viennent de passer le baccalauréat (qui se situe officiellement au niveau B2 pour l'anglais LV1). Les étudiants L1 jouissent néanmoins d'un statut privilégié parce qu'ils représentent la population cible de ce dispositif. POSILANG est en effet conçu pour évaluer le niveau de compétences des bacheliers en anglais, primo entrants dans l'Enseignement Supérieur. Toutes les universités, on le sait, sont confrontées à une très grande hétérogénéité de niveaux. Elles doivent, en quelques heures, constituer des groupes L1 aussi homogènes que possibles pour les enseignements obligatoires d'anglais.

L'analyse des résultats du test de pilotage montre que la plupart des candidats se rassemblent sur quelques niveaux de compétence uniquement, tandis que d'autres niveaux ne sont attribués à personne. En ce qui concerne les étudiants de master en sciences du langage (M2-RELAI), la répartition des candidats sur les différents niveaux de compétence y est plus homogène que dans le cours de Licence (L1- ANGLAIS).

La comparaison entre le niveau de compétence langagière des candidats préalablement déterminé en cours avec le niveau de compétences atteint lors de la passation de POSILANG en version pilote a démontré la validité du test de positionnement. Dans le cours L1, plus de 80 % des candidats se positionnent sur deux niveaux (B1 et B2) qui sont soit les mêmes que les niveaux estimés au contrôle continu, soit minimalement déviants. Dans le groupe M2, la moitié environ des étudiants valident deux niveaux qui sont soit identiques soit très proches.

L'évaluation détaillée des résultats du pilotage nous permet de déterminer le pourcentage de candidats ayant réussi chacune des tâches. On constate avec satisfaction qu'aux deux niveaux inférieurs, la grande majorité des participants répondent correctement à presque toutes les tâches. A partir du niveau B1, le pourcentage de candidats qui réussissent baisse nettement dans les deux groupes sur plusieurs tâches. Les différences observées dans la valeur de facilité, entre les deux cours L1 et M2-RELAI, sont plus larges à partir du niveau

B1 qu'aux niveaux inférieurs, A1 et A2. Ceci est dû au fait que tous les candidats ont acquis au moins le niveau A2+.

Concernant l'index de discrimination des tâches contenues dans POSILANG, on constate une répartition parallèle sur les cinq niveaux de référence dans les deux groupes d'étudiants, L1 et M2-ReLAI. Chez les L1, aux deux niveaux inférieurs, il n'est guère possible de départager nettement le tiers le plus fort du tiers le plus faible. Cependant, à partir du niveau B1, la distinction se fait plus nette. Il y a une plus grande quantité de tâches permettant de séparer les deux groupes. Ce fait se manifeste par des indices de discrimination beaucoup plus élevés qu`aux niveaux inférieurs. Dans le cours M2-ReLAI, les tâches discriminent de la même manière entre les deux tiers des participants, mais la distribution des tâches est différente de l'autre groupe**.**

Il faut également souligner l'intérêt de procéder à une analyse des options fausses, dénommées « distracteurs. » Celles-ci sont autant de pièges tendus qui permettent d'identifier les lacunes langagières des candidats. Or déceler des propensions à tel ou tel type d'erreur relève bien de la composante diagnostique qui a été intégrée à notre test de positionnement. Dans les deux cours, la compréhension orale est l'activité communicative la plus concernée par les erreurs. Dans le domaine des compétences linguistiques, c'est la compétence grammaticale qui rassemble la plus grande quantité d'erreurs. Ces-dernières concernent tous les candidats de L1 et de M2-ReLAI. Enfin, pour ce qui est de la compétence lexicale, les candidats des deux groupes se trompent sur les tâches portant sur la synonymie.

En résumé, POSILANG permet de déterminer non seulement le niveau en langue global de candidats, mais aussi d'évaluer celui-ci par domaine de compétence. Le test remplit la fonction diagnostique qui lui a été assignée, en pointant les lacunes des candidats. Ainsi peuvent être créés différents groupes de niveaux à l'université pour remédier aux faiblesses repérées. Nous formulons donc le souhait qu'une version expérimentale (V1) puisse désormais être installée sur un serveur mutualisé des universités d'Aquitaine. Nous espérons aussi que notre travail, éminemment perfectible, pourra motiver et aider d'autres concepteurs à construire leurs propres dispositifs.

Notes:

[1]Au lieu de: “As he was going to the market…”

[2] La performance orale a été réalisée en semaines 9 et 10 du premier semestre 2014/2015.

[3] http://www.linguisticsociety.org/

[4] http://www.collinsdictionary.com/

[5] http://www.ivona.com/

[6] Ces quatre critères sont adoptés par le responsable du cours. (Annexe 5).

[7] Le critère « interaction » n’est pas pris en compte lors de l’évaluation de la présentation orale. Il le sera lors de l’évaluation de l’activité suivante, effectuée par les étudiants (interview). Les autres critères évoqués serviront d’appui à la doctorante lors de l’évaluation de ces deux activités.

[8] Quatre étudiants n’ont pu être évalués par la doctorante parce qu’ils n’ont pas présenté leur exposé en cours, mais en semaines 11-12 dans le bureau de l’instructeur.

[9] « La richesse de vocabulaire » n’apparait pas comme un critère séparé, mais cette qualité est pourtant prise en compte lors de l’évaluation. Elle est couverte par les critères « respect de la phraséologie du milieu professionnel » et « étendue », provenant du CECRL (CECRL 2005 : 28)

[10] L’interaction orale est menée dans la semaine 11 du premier semestre de l’année 2014/2015.

[11] Il s’agit ici des critères adoptés par l’enseignant responsable du cours. L’autonomie par rapport aux notes apparaît en tant que critère formel lors de l’évaluation de l’interaction orale, mais il a été déjà pris en compte par endroits lors de l’évaluation de la présentation orale. En ce qui concerne les critères de la doctorante, ce sont les mêmes que lors de l’évaluation des exposés. Le critère « interaction » qui n’a pas été pris en compte lors de l’évaluation de la première tâche, sera consulté dans le but d’évaluer les interactions.

[12] Les fautes de prononciation surgissent le plus fréquemment aux niveaux B1 et B2+, dans les deux types de performance orale.

[13] https://www.youtube.com/watch?v=r4VILAGHVCs

[14] http://www.ted.com/talks/ken_robinson_says_schools_kill_creativity

[15] https://www.youtube.com/watch?v=sZwfNs1pqG0

[16] Ces compétences sont désignées fonctionnelles pour les délimiter des compétences linguistiques

[17] Il est nécessaire d’acquérir 3,5 points lorsque le score maximal est de 6, et 3 points lorsque le score s’élève à 5 points.

[18] Bien que le groupe se compose de 28 étudiants, trois ne sont pas pris en compte lors de la comparaison des niveaux. La raison à cela est que l’établissement préalable des niveaux de compétences n’a pu être réalisé que pour 25 personnes.

[19] Dans le groupe M2 en sciences du langage, la passation du POSILANG s’est déroulée en deux fois à cause du manque de temps la première fois. Certains candidats, absents à l’une ou l’autre séance, n’ont pas pu terminer les tâches d’un ou de plusieurs niveaux.

[20] Le seuil limite de fiabilité se situe à 60% dans le groupe. Il s’agit de tâches auxquelles moins de 60% des étudiants ont répondu correctement.

Conclusion

L'évaluation des compétences en langues est l'objet de l'intérêt croissant des chercheurs, des responsables de cursus, des enseignants, des étudiants et de la société en général (Brown 2010 : 1x). Cet intérêt a suscité l'apparition d'un nombre exponentiel de tests de langue. En fonction de leurs objectifs, ces tests se répartissent en sept types : en tests de positionnement, d'acquisition et de progrès, de compétence, de certification, diagnostiques et enfin tests d'aptitude. Malgré la multiplication des dispositifs disponibles sur le marché, un test comme POSILANG, développé dans le cadre d'un projet régional, présente des avantages.

Comme son prestigieux aîné DIALANG, POSILANG est à la fois gratuit et adossé au *Cadre européen commun de référence pour les langues* (CECRL). Cet adossement est réel: il ne s'agit pas d'un rattachement stratégique, comme cela a pu être le cas pour le test américain TOEIC développé par le géant ETS GLOBAL. L'enjeu n'est pas de procéder à un « adossement » de circonstance, de se prévaloir d'un cadrage institutionnel de pure forme, ou encore d'effectuer la conversion opportuniste d'un score en niveau CECRL. Il s'agit plutôt d'intégrer *à la source* un maximum de paramètres du Cadre, de respecter, autant que faire se peut, l'esprit de l'approche communicative-actionnelle prônée, alors même que le format du test est contraint et que des réductions doivent être opérées dans les compétences testées. Idéalement, un dispositif comme POSILANG doit pouvoir se distinguer de l'offre existante par sa praticité et sa fiabilité, par son accessibilité et sa gratuité.

L'adossement de POSILANG au CECRL se traduit de façon très explicite par le respect des niveaux de référence: A1, A2, B1, B2 et C1. (Seul le niveau C2 n'est pas proposé). L'attribution d'un niveau global à chaque candidat sur l'échelle du CECRL, à partir des descripteurs institués, repose sur l'évaluation de trois domaines de compétence langagière: la compréhension orale, la compréhension écrite et, en mode parcellaire, l'expression écrite. Par ailleurs, un test comme POSILANG doit respecter le critère d'authenticité. Ce respect se manifeste par l'élaboration d'items se rapprochant autant que possible de la forme, des fonctions et des usages situés de la langue. Afin de vérifier

l'adéquation formelle et fonctionnelle des items, il est utile de faire appel à l'expérience et à l'intuition des natifs en langue anglaise. Il faut également aspirer à l'interrogation de corpus, à l'observation des pratiques. Cela est essentiel lorsque le dispositif est monté en milieu exogène. Enfin, un test construit sur le schéma que nous avons retenu pour POSILANG doit s'efforcer de respecter le principe d'interactivité dans la mesure où il est élaboré localement, en tenant compte d'une situation d'évaluation particulière.

Ainsi peuvent être satisfaites les exigences de la recherche dans la discipline, généralement ignorées par des concepteurs de tests, pas toujours experts. Le résultat de cette manière de construire des tests est le plus souvent un dispositif simplifié, focalisé sur la seule correction formelle, qui devient emblématique de la maîtrise de la langue. A part le caractère lacunaire de l'évaluation ainsi opérée, on note le paradoxe d'une démarche décontextualisée, en rupture avec les stratégies d'acquisition développées en milieu scolaire. On sait en effet, que l'enseignement scolaire des langues est aujourd'hui moins fixé sur la grammaire et la perfection formelle qu'il ne l'était à l'époque structuraliste, qu'on s'est distancé de la norme imposée par locuteur autochtone et qu'un renoncement à la fiction du bilinguisme parfait a été opéré. Or, à l'occasion d'évaluations sommatives ou de la passation de tests de positionnement, le candidat doit soudain se comporter en détenteur expert de la norme grammaticale, sans pouvoir témoigner d'autres capacités. Certes, le respect de cette norme peut faire partie des exigences pertinentes, comme c'est le cas dans un département de spécialistes de langue par exemple. Cependant, même en pareil cas, la seule compétence grammaticale ne saurait servir à accréditer l'intégralité de la compétence en langue. Cela est plus vrai encore pour les spécialistes d'autres disciplines, qui peuvent avoir d'autres besoins : comprendre des textes dans leur spécialité, rédiger de courtes notes, faire un exposé. D'où la nécessité de construire localement des dispositifs de positionnement, en intégrant au maximum la situation d'évaluation spécifique (Hughes 2003 : 17). D'où l'utilité aussi de diversifier les compétences mesurées afin de pouvoir placer de façon plus précise les candidats sur plusieurs échelles de compétence.

Par ailleurs, il nous semble essentiel de fournir au candidat des critères d'attribution des scores précis et transparents. C'est ce que nous avons tenté

d'opérer pour POSILANG, dans la situation d'évaluation précise que constitue le site universitaire bordelais.

De la conception d'un test de positionnement à sa diffusion, les étapes sont nombreuses et exigeantes, dès lors qu'on renonce au bricolage. Nous savons par expérience que les équipes pédagogiques confient souvent la réalisation de tests de langues à des individus isolés, qui se dévouent dans l'urgence, sans formation particulière en psychométrie, sans moyens logistiques avancés, pour répondre à des injonctions administratives ou pédagogiques de positionnement à des fins de pré-orientation, de répartition dans des groupes de niveaux, d'aiguillage vers des activités de soutien ou de remédiation. Comme tout bricolage, le dispositif peut dépanner et même fonctionner temporairement, mais en aucune manière cela ne devrait suffire à le pérenniser. Les langues et les étudiants qui les pratiquent méritent mieux.

Une étape essentielle dans la mise au point d'un dispositif d'évaluation est celle du pré-test. Nous ne saurions trop insister sur ce point. Pour POSILANG, nous avons proposé une passation manuelle à deux groupes distincts d'étudiants. Nous avons soumis à ces derniers une version pilote au format papier + enregistrement sonore de POSILANG. Ont participé des étudiants de master 2 en sciences de langage et des étudiants en première année de licence d'anglais (L1). Les étudiants de master sont certes des spécialistes d'une autre discipline mais étudient la linguistique à un niveau avancé depuis au moins quatre ans, lisent des écrits scientifiques en anglais, parlent souvent plusieurs langues et de ce fait, ont des compétences plutôt élevées (à quelques exceptions près). En revanche, les étudiants en première année de licence d'anglais, tout spécialistes qu'ils soient, viennent de passer le baccalauréat (qui se situe officiellement au niveau B2 pour l'anglais LV1). Les étudiants L1 jouissent néanmoins d'un statut privilégié parce qu'ils représentent la population cible de ce dispositif. POSILANG est en effet conçu pour évaluer le niveau de compétences des bacheliers en anglais, primo entrants dans l'Enseignement Supérieur. Toutes les universités, on le sait, sont confrontées à une très grande hétérogénéité de niveaux. Elles doivent, en quelques heures, constituer des groupes L1 aussi homogènes que possibles pour les enseignements obligatoires d'anglais.

L’analyse des résultats du test de pilotage montre que la plupart des candidats se rassemblent sur quelques niveaux de compétence uniquement, tandis que d’autres niveaux ne sont attribués à personne. En ce qui concerne les étudiants de master en sciences du langage (M2-RELAI), la répartition des candidats sur les différents niveaux de compétence y est plus homogène que dans le cours de Licence (L1- ANGLAIS).

La comparaison entre le niveau de compétence langagière des candidats préalablement déterminé en cours avec le niveau de compétences atteint lors de la passation de POSILANG en version pilote a démontré la validité du test de positionnement. Dans le cours L1, plus de 80 % des candidats se positionnent sur deux niveaux (B1 et B2) qui sont soit les mêmes que les niveaux estimés au contrôle continu, soit minimalement déviants. Dans le groupe M1, la moitié environ des étudiants valident deux niveaux qui sont soit identiques soit très proches.

L’évaluation détaillée des résultats du pilotage nous permet de déterminer le pourcentage de candidats ayant réussi chacune des tâches. On constate avec satisfaction qu’aux deux niveaux inférieurs, la grande majorité des participants répondent correctement à presque toutes les tâches. A partir du niveau B1, le pourcentage de candidats qui réussissent baisse nettement dans les deux groupes sur plusieurs tâches. Les différences observées dans la valeur de facilité, entre les deux cours L1 et M2-RELAI, sont plus larges à partir du niveau B1 qu’aux niveaux inférieurs, A1 et A2. Ceci est dû au fait que tous les candidats ont acquis au moins le niveau A2+.

Concernant l’index de discrimination des tâches contenues dans POSILANG, on constate une répartition parallèle sur les cinq niveaux de référence dans les deux groupes d’étudiants, L1 et M2-ReLAI. Chez les L1, aux deux niveaux inférieurs, il n’est guère possible de départager nettement le tiers le plus fort du tiers le plus faible. Cependant, à partir du niveau B1, la distinction se fait plus nette. Il y a une plus grande quantité de tâches permettant de séparer les deux groupes. Ce fait se manifeste par des indices de discrimination beaucoup plus élevés qu`aux niveaux inférieurs. Dans le cours M2-ReLAI, les

tâches discriminent de la même manière entre les deux tiers des participants, mais la distribution des tâches est différente de l'autre groupe**.**

Il faut également souligner l'intérêt de procéder à une analyse des options fausses, dénommées « distracteurs. » Celles-ci autant de pièges tendus qui permettent d'identifier les lacunes langagières des candidats. Or déceler des propensions à tel ou tel type d'erreur relève bien de la composante diagnostique qui a été intégrée à notre test de positionnement. Dans les deux cours, la compréhension orale est l'activité communicative la plus concernée par les erreurs. Dans le domaine des compétences linguistiques, c'est la compétence grammaticale qui rassemble la plus grande quantité d'erreurs. Ces-dernières concernent tous les candidats de L1 et de M2-ReLAI. Enfin, pour ce qui est de la compétence lexicale, les candidats des deux groupes achoppent sur les tâches portant sur la synonymie.

En résumé, POSILANG permet de déterminer non seulement le niveau en langue global de candidats, mais aussi d'évaluer celui-ci par domaine de compétence. Le test remplit la fonction diagnostique qui lui a été assignée, en pointant les lacunes des candidats. Ainsi peuvent être créés différents groupes de niveaux à l'université pour remédier aux faiblesses repérées. Nous formulons donc le souhait qu'une version expérimentale (V1) puisse désormais être installée sur un serveur mutualisé des universités d'Aquitaine. Nous espérons aussi que notre travail, éminemment perfectible, pourra motiver et aider d'autres concepteurs à construire leurs propres dispositifs.

Bibliographie

1 Références

1.1 Ouvrages cités

Alderson, J.-Ch., Clapham, C. & Wall D. (1995). *Language Test, Construction and Evaluation*. Cambridge University Press

http://books.google.fr/books/about/Language_Test_Construction_and_Evaluatio.html?id=VQ3XpZvA5eAC&redir_esc=y

Alderson, J. Ch. (2000). *Assessing Reading*. Cambridge : CUP.

http://books.google.de/books?id=XrhcbBr9gLwC&printsec=frontcover&dq=Alderson,+J.+Ch.+(2000).+Assessing+Reading.+Cambridge+:+CUP

Alderson, J. et al. (2004). *The Development of Specifications for Item Development and Classification within The Common European Framework of Reference for languages: Learning, Teaching, Assessment. Reading and Listening.* Final Report of the Dutch CEF Construct Project

http://eprints.lancs.ac.uk/44/1/final_report.pdf

Alderson, J. C. (2005). *Diagnosing Foreign Language Proficiency: the Interface between Learning and Assessment*. Paperback edition.

http://books.google.de/books?id=Y3xsr7g3aIC&printsec=frontcover&dq=Alderson,+J.+Ch.+(2005).+Diagnosing+Foreign+Language+Proficiency:+the+Interface+between+Learning+and+Assessment.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

http://books.google.de/books?id=5_KJCfkWgqcC&printsec=frontcover&dq=Bachman,+L.F.+(1990).+Fundamental+considerations+in+language+testing.+Oxford:+Oxford+University+Press.

Bachman, L.F. & Palmer, A.S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

http://books.google.de/books?id=E0yH0NdySrQC&printsec=frontcover&dq=Language+testing+in+practice:+Designing+and+developing+useful+language+tests.

Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Techniques*. New York: Marcel Dekker.

http://books.google.fr/books?id=yQ_Q7pasJ0C&pg=PA466&lpg=PA466&dq=BAKER,+F.+B.+(1992).+Item+Response+Theory

Beacco, J.-C., de Ferrari, M., & Lhote, G. (Eds.). (2006). Niveau Al.1 pour le français: Référentiel et certification (DILF) pour les premiers acquis en français. Paris: Didier.

Bourguignon, C. (2010). *Pour enseigner les langues avec le CECRL : Clés et conseil*. Paris : Delagrave

http://babordplus.univbordeaux.fr/notice.php?q=fulltext%3A%28Bourguignon

Brown, H. D. (2010). *Principles and classroom practices*. Pearson Education.

http://books.google.de/books?id=4EyVQAAACAAJ&dq=Principles+and+classroom+practices.

Chapelle, C. & D. Douglas (2006). *Assessing Language through Computer Technology*. Cambridge: CUP.

http://books.google.de/books?id=CkkDXmI4dJMC&printsec=frontcover&dq=Assessing+Language+through+Computer+Technology

Chardenet, P. (1999). *De l'activité évaluative à l'acte d'évaluation*, Paris : L´Harmattan

file:///C:/Users/User/Downloads/extrait_activite__de_l___evaluative_a_l_acte_d_eval.pdf

Crocker, L.M. & J. Algina (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston,

http://books.google.ca/books/about/Introduction_to_classical_and_modern_tes.html?id=tfgkAQAAMAAJ

Davies, A. (2007). *An Introduction to Applied Linguistics: From Practice to Theory*, Edinburgh University press

http://books.google.fr/books?id=GvdsrHHJQ4C&printsec=frontcover&hl=de&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

De Jong, J.H.A.L. (2004). Comparing the psycholinguistic and the communicative paradigm of language proficiency. Presentation given at the international workshop "Psycholinguistic and psychometric aspects of language assessment in the Common European Framework of Reference for Languages". University of Amsterdam, 13-14 February, 2004.

Dervin, F. & E. Suomela-Salmi (Eds.) (2007): *Evaluer les compétences langagières et interculturelles dans l'enseignement supérieur.* Département d'Etudes françaises. Université de Turku.

http://doria32kk.lib.helsinki.fi/bitstream/handle/10024/69212/assessment.pdf?sequence=1

Douglas, D. (2010). *Understanding Language Testing*. London: Hiroshi Higuchi.

http://books.google.de/books?id=S4V2PgAACAAJ&dq=Understanding+Language+Testing.&hl=de&sa=X&ei=i9AwUdTTEueM4ATdnoBY&ved=0CDUQ6AEwAA

Firth, J. E: (1957). Papers in Linguistics 1934-1951. London: Oxford University Press.

http://books.google.fr/books/about/Papers_in_linguistics_1934_1951.html?id=yxZZAAAAMAAJ&redir_esc=y

Fulcher, G. & F. Davidson (2007). *Language Testing and Assessment*. New York: Routledge.

http://www.amazon.com/Language-Testing-Assessment-Routledge-Linguistics/dp/0415339472

Fulcher, G. (2010). *Practical Language Testing*. London.

http://www.amazon.com/Practical-Language-Testing-Glenn-Fulcher/dp/0340984481

Gronlund (1998). *Assessment of student achievement.* Boston: Allyn & Bacon.

http://books.google.de/books/about/Assessing_Language_for_Specific_Purposes.html?id=exfxgbsRn1gC&redir_esc=y

Goullier, F. (2005). *Les outils du Conseil de l'Europe en classe de langue. Cadre européen commun et Portfolios*. Paris : Didier.

http://www.coe.int/t/dg4/linguistic/Source/Goullier_Outils_1.FR.pdf

Gronlund, N.E & Waugh, C.K. (2008). *Assessment of student achievement* (9th ed.), Boston: Allyn & Bacon.

Hasselgreen, A. (2003.) *The Bergen-Cando Project* (book and CD). Graz: ECML/Council of Europe.

http://www.hib.no/senter/suf/AngelaHasselgreen.asp

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation, Research.* Cambridge, MA: Newbury House.

Hughes, A. (1989): *Testing for language teachers*, 1. Edition, Cambridge University Press,

http://catdir.loc.gov/catdir/samples/cam034/2003268576.pdf

Hughes, A. (2003): *Testing for language teachers*, Second Edition, Cambridge University Press

http://books.google.de/books/about/Testing_for_Language_Teachers.html?id=C66hOKjHWGAC

Huver, E. & Springer, C. (2011). *L'évaluation en langues*. Paris: Didier.

http://www.amazon.fr/L%C3%A9valuation-en-langues-Emmanuelle-Huver/dp/2278064002

Lallement, B., Pierret-Lallement, N. (2007). *L´essentiel du CECR pour les langues : le cadre européen commun de référence pour les langues. Ecole, collège, lycée.* Paris, Hachette éducation

http://babordplus.univbordeaux.fr/notice.php?q=fulltext%3A%28Lallement%2C%20Brigitte%29&spec_expand=1&sort_define=score&sort_order=1&rows=10&start=20

Madsen, H.S. (1983). *Techniques in testing.* Oxford: Oxford University Press

http://eric.ed.gov/?id=ED242211

Martuccelli, D. (2010). *La societé singulariste*. Paris : Armand Colin

http://sociologies.revues.org/3344

Mc Namara, T. (2000). *Language testing.* Oxford university press

http://de.scribd.com/doc/102946131/What-is-a-Language-Test-Chapter-1-McNamara

Mousavi, S.A. (2009). An encyclopedic dictionary of language testing (4.th ed.), Tehran: Rahnama Publications

Nichols, S. & Berliner, D. (2007). *Collateral damage: How high stakes testing corrupts America´s schools.* Cambridge, MA: Harvard Education Press

http://books.google.fr/books?id=f7eAAAAMAAJ&q=Nichols:+Collateral+damage:+How+high+stakes+testing+corrupts.++America

Noel- Jothy, F. & Sampsonis B. (2006). *Certifications et outils d´évaluation en FLE* : Hachette, Paris

http://books.google.fr/books/about/Certifications_et_outils_d_%C3%A9valuation_e.html?id=e2JLHQAACAAJ&redir_esc=y

Purpura, J. E. (2004). *Assessing Grammar.* Cambridge: CUP.

http://www.upbo.com/servlet/file/store6/item2457252/version1/item_9780521003445_frontmatter.pdf

Shohamy, E. (2001). *The Power of Tests: A Critical Perspective on the Uses of Language Tests*. London: Longman.

http://books.google.fr/books?hl=de&lr=&id=c5rgnGYIhwcC&oi=fnd&pg=PR11&dq=Shohamy+:The+power+of+tests

Sisamakis, E.M. (2006). *The European Language Portfolio in Irish Post-Primary Education*. A longitudinal empirical evaluation. PhD thesis.

http://archive.ecml.at/mtp2/Elp_tt/Results/DM_layout/Reference%20Materials/English/Manolis%20Sisamakis%20PhD%20Thesis.pdf

Tagliante, Ch. (1991). *L´Evaluation.* Paris : CLE International

Tagliante, Ch. (2005). *L'évaluation et le Cadre Européen Commun*. Paris: CLE International.

http://www.amazon.fr/L%C3%A9valuation-cadre-europ%C3%A9en-Christine-Tagliante/dp/2090331194

Thornbury, Scott (1999). *How to teach Grammar.* Pearson/Longman.

http://books.google.fr/books?id=eKK9mgEACAAJ&dq=Thornbury:++How+to+teach+Grammar.+Pearson/Longman

Vernon, P.E. (1956). *The Measurement of Abilities.* 2nd ed. London: University of London Press.

Vygotski, L. (1934). *Pensée et langage*. 3ème édition parue en 1997. Paris: La Dispute.

http://www.amazon.fr/Pens%C3%A9e-langage-Lev-Semenovitch Vygotski/dp/2843030048

Waugh N.E, Gronlund, C.K. (2008). *Assessment of Student Achievement*. Pearson.

http://www.amazon.co.uk/Assessment-Student-Achievement-ninth-Paperback/dp/B00BSZWPP4

Widdowson, H.G (1983). *Learning purpose and language use.* Oxford University http://books.google.fr/books/about/Learning_purpose_and_language_use.html?id=YlkFAQAAIAAJ&redir_esc=yess

1.2 Articles universitaires cités

Alderson, J.C. & D.Wall (1993). Does washback exist? *Applied Linguistics*, 14 (1993), 115-129.

http://span676testingnassessment.wikispaces.com/file/view/Anderson+%26+Wall+1993+Backwash+Fact+of+Fiction

Alderson, J.C. (2004). Foreword. In: Cheng, L., Watanabe, Y. and Curtis, A. (Eds.) *Washback in language Testing: Research Contexts and Methods*. Mahwah, NJ: Erlbaum.

http://books.google.fr/books?id=jicFacTeHs0C&hl=de&source=gbs_ViewAPI&redir_esc=y

Alderson, J. C. (2004). The shape of things to come: will it be the normal distribution? In: *European Language testing in a global context: Proceedings of the ALTE Barcelona Conference*. July 2001. Cambridge University press, 1-26. http://assets.cambridge.org/97805215/35878/sample/9780521535878ws.pdf

Alderson, J.C.& Banerjee, J. (2002). Language testing and assessment (Part 2). *Language teaching*, 35, 79-113.

http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=108751&fileId=S0261444802001751

Alderson, J.C. & Huhta A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing* 22, 301-320.

http://www.research.lancs.ac.uk/portal/en/publications/the-development-of-a-suite-of-computerbased-diagnostic-tests-based-on-the-common-european-framework. html

Bachman L.F. & Palmer, A.S. (1982a) The Construct Validation of Some Components of Communicative Proficiency. *TESOL Quarterly*, Vol. 16, No. 4, 449-465.

http://www.jstor.org/stable/3586464.

Bachman, L. & Purpura, J. (2008). Language assessments: Gate-keepers or door-openers? In B. M. Spolsky & F. M. Hult (Eds.), *Blackwell handbook of educational linguistics*. Oxford, UK: Blackwell, 521-532.

http://www.blackwellreference.com/public/tocnode?id=g9781405154109_chunk_g978140515410933

Bélanger, J. (2002). Construction des items. Introduction à la psychométrie. Université du Québec à Montréal.

http://www.er.uqam.ca/nobel/r30034/PSY4130/doc/items.html

Boud, D. & N. Falchikov (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education*. Vol. 31, No. 4, August 2006, 399–413.

http://www.jhsph.edu/departments/population-family-and-reproductive-health/_docs/teaching-resources/cla-01-aligning-assessment-with-long-term-learning.pdf

Brown, J. D. & T. Hudson (1998). The alternatives in language assessment. *TESOL Quarterly*, 32(4): 653–675.

http://onlinelibrary.wiley.com/doi/10.2307/3587999/abstract

Byrnes, H. (2007). Introduction to perspectives. The Modern Language Journal, Vol. 91, No.4, 641-645.

http://www.jstor.org/stable/4626090

Canale, M (1987). The measurement of communicative competence. In: R.B. Kaplan, et al. (eds.) *Annual Review of Applied Linguistics*, 8. New York: Cambridge University Press, 67-84.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics,* 1, 1-47.

http://ibatefl.com/wp-content/uploads/2012/08/CLT-Canale-Swain.pdf

Carroll, J.B. (1968). Contrastive Analysis and Interference Theory. In J. Alatis (Ed.) *Contrastive Linguistics and its Pedagogical Implications.* Washington, DC: Georgetown University Press, 113-122.

http://eric.ed.gov/?id=ED022159

Cenoz, J. & D. Gorter (2011). A Holistic Approach in Multilingual Education: Introduction. *The Modern Language Journal.* Special Issue: The Special Issue: Toward a Multilingual Approach in the Study of Multilingualism in School Contexts, 95/3, 339-343.

http://onlinelibrary.wiley.com/doi/10.1111/j.1540-4781.2011.01204.x/abstract

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.) *Second language acquisition and language testing interfaces.* Cambridge: Cambridge University Press, pp. 32-70.

http://ebooks.cambridge.org

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254-272.

http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=168271&fileId=S0267190599190135

Cohen, A. D. (1998). Strategies and processes in test taking and SLA. In: Bachman, L.F. & A.D. Cohen (Eds.) *Interfaces between second language acquisition and language testing research*. Cambridge: CUP, 90-111.

file:///C:/Users/User/Downloads/1998%20-%20Strats%20&%20Procs%20in%20Tst-Tkg%20&%20SLA%20in%20Bachman%20&%20Cohen%20(1).pdf

Cronbach L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. Psychological bulletin, Vol. 52 No. 4, 281-302.

http://marces.org/EDMS623/Cronbach%20LJ%20&%20Meehl%20PE%20(1955)%20Construct%20validity%20in%20psychological%20tests.pdf

Davies, A. (2007). Ethics, professionalism, rights and codes. In E. Shohamy & N.H. Hornberger (Eds.) *Encycolopedia of language and education* (2ndEd.), Volume 7: Language Testing and Assessment. Springer Science and Business Media, 419-443.

http://www.tc.umn.edu/~lazaratn/images/publications/Lazaraton-2008.pdf

Davies, A. (1997b). Demands of being professional in language testing. *Language Testing* 14, 3, 328-339.

http://ltj.sagepub.com/content/14/3/328.short

Demaizière, F. & J.-P. Narcy-Combes: Du positionnement épistémologique aux données de terrain. *Cahiers de l'Acedle*, numéro 4, juin 2007.

http://acedle.org/IMG/pdf/Demaiziere-Narcy_cah4.pdf

Doucet, Patrick (2001). Pour un test utile, *ASp* [En ligne], 34 |(2001), 13-33.

http://asp.revues.org/1696

Douglas, D. & V. Hegelheimer (2008). Assessing language using computer technology. *Annual Review of Applied Linguistics* (2007) 27, 115-132.

http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=1820476&fileId=S0267190508070062

Davidson, F. & G. Fulcher (2007). The Common European Framework of Reference (CEFR) and the Design of Language Tests: A matter of Effect. *Language Teaching: The International Research Resource for Language Professionals*, 40 (3), 231-241.

http://languagetesting.info/articles/store/CEFR%20A%20Matter%20of%20Effect.pdf

Frederiksen, J.R. & Collins, A. (1989). A Systems approach to educational testing. *Educational Researcher* 18, 9, 27-32.

http://edr.sagepub.com/content/18/9/27.short

Glabionat, M., et al. (2002). Profile deutsch. Gemeinsamer Europaischer Referenzrahmen. Lernzielbestimmungen, Kannbeschreibungen, kommunikative Mittel. Muenchen: Langenscheidt.

Grice, H. P. (1975). Logic and conversation. In: Cole, P. & J.L. Morgan, (eds). *Speech Acts*. New York: Academic Press, 41–58

http://www.cog.brown.edu/courses/cg45/lecture%20slides/Gricean%20Maxims.pdf

Hulstijn, J.H. (2007). The Shaky Ground beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency. *Modern Language Journal*, 91 (4), 663-667.

http://www.jstor.org/stable/4626094

Huhta, A. et al. (2002). Dialang- A diagnostic language assessment system for adult learners. In: Alderson, C. (Ed.). *Common European Framework of Reference for Languages*. *Learning, Teaching, Assessment.* Case Studies. Strasbourg, 130-145.

http://www.coe.int/t/dg4/linguistic/publications_fr.asp

Kaftandjieva, F. & S. Takala (2002). Council of Europe scales of language proficiency: a validation study. In: Alderson, C. (Ed.). *Common European Framework of Reference for Languages. Learning, Teaching, Assessment.* Case Studies. Strasbourg, 106-129.

http://www.coe.int/t/dg4/linguistic/publications_fr.asp

Ketele De., J.M. (2006). La notion émergente de compétence dans la construction des apprentissages. In: Figari, G. & L. Mottier Lopez (Eds.) *Recherche sur l'évaluation en éducation – Problématiques, méthodologies et épistémologie*. Paris : L'Harmattan.

http://books.google.fr/books?id=mTKfrBrj1EC&pg=PA17&dq=Ketele+:+La+notion+%C3%A9mergente+de+comp%C3%A9tence+dans+la+construction+des+apprentissages.

Kingsbury, G. C. & Weiss D. J. (1980). An Alternative-Forms Reliability and Concurrent Validity Comparison of Bayesian Adaptive and Conventional Ability

Tests. Research Report 80-5. Minneapolis, MN: University of Minnesota, Department of Psychology.

https://www.psych.umn.edu/psylabs/catcentral/pdf%20files/ki80-05.pdf

Krumm, H.-J. (2007). Profiles Instead of Levels: The CEFR and Its (Ab) Uses in the Context of Migration. *Modern Language Journal*, 91, 667-669.

http://onlinelibrary.wiley.com/doi/10.1111/j.15404781.2007.00627_6.x/abstract

Kunnan, A.J. (2000). Fairness and justice for all. In: Kunnan, A.J. (Ed.). *Fairness and Validation in Language Assessment. Studies in Langage Testing* 9. Cambridge: Cambridge University Press, 1-14.

http://cambridgelearning.net/servlet/file/store6/item2350537/version1/item_9780521658744_excerpt.pdf

Laurier, M. (1998). Méthodologie d'évaluation dans des contextes d'apprentissage des langues assistés par les environnements informatiques multimédias. *Etudes de linguistique appliquée*, Klincksieck 1998, Hypermédia et apprentissage des langues, 247-255.

https://halshs.archives-ouvertes.fr/edutice-00000234/document

Laurier, M. (2006). Les tests adaptatifs en langue : quel est leur avenir? Colloque ACFAS. Solutions apportées et problèmes engendrés par l'évaluation assistée par les technologies de l'information et de la communication en éducation.

www.camri.uqam.ca/.../2006_Laurier_ACFAS.ppt

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *Modern Language Journal*, 91, 645-655.

http://www.jstor.org/stable/4626091

Luoma, S (2004). Self-assesment in DIALANG. An account of test development. In: Alderson, J. C. (2004). *European Language testing in a global context: Proceedings of the ALTE Barcelona Conference July 2001*. Cambridge: Cambridge University press, 143-156.

http://assets.cambridge.org/97805215/35878/sample/9780521535878ws.pdf

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational measurement* (3rd ed.) New York: American Council on Education and Macmillan, 13-103.

Messick, S. (1996). Validity and washback in language testing. *Language testing* 13, 241-256.

http://ltj.sagepub.com/content/13/3/241.short

Miller, Marie (2007). Les Instructions Officielles au Collège. Palier 2.

http://esperessources.unistra.fr/web.ressources/web/ressources_pedagogiques/productions_pedagogiques_iufm/anglais/2nddegre/palier2.pdf

Morrow, K. (1986). The Evaluation of tests of communicative performance. In: Portal, M. (Ed). *Innovations in Language Testing*. London. NFER/Nelson, 1-13.

http://scholar.google.fr/scholar?q=+Innovations+in+Language+Testing%2C+&btnG=&hl=de&as_sdt=0%2C5

Narcy-Combes, J.-P. et al. (2007). Apport des savoirs savants en didactique des langues : modélisation ou transposition ? *Le Français dans le monde: recherches et applications*, janvier 2014, n° 55, 153-167

http://www.ciep.fr/veille-editoriale/avril-2014/politiques-linguistiques-didactique-langues

North B., Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, vol. 15 n° 2, 217-262.

http://ltj.sagepub.com/content/15/2/217.short

North, B. (2007). The CEFR Illustrative Descriptor Scales. *Modern Language Journal*, 91 (4), 656-663.

http://www.jstor.org/stable/4626092

Purpura, J.E. (2001). Evaluating selected-response items: guidelines. Teachers college, Columbia University.

Shohamy, E. (1983) The stability of oral proficiency in the oral interview procedure. *Language Learning* 33, 527-540.

http://onlinelibrary.wiley.com/doi/10.1111/j.1467-1770.1983.tb00947.x/abstract

Shohamy, E. (2000). Fairness in language testing. In: Kunnan, A.J (ed). *Fairness and Validation in Language Assessment. Studies in Language Testing* 9. Cambridge: Cambridge University Press, 15-19.

https://books.google.fr/books?hl=de&lr=&id=x83rx-lq8k8C&oi=fnd&pg=PA15&dq=Shohamy,+E.+(2000).+Fairness+in+language+testing.

Westhoff G. (2007) Challenges and opportunities of CEFR for reimagining FL Pedagogy. *Modern language journal* 91, (4), 675-8.

http://www.jstor.org/discover/10.2307/4626098?uid=2&uid=4&sid=21104037447261

1.3 Ouvrages et articles consultés mais non cités

Alderson, J.C. (1991). Bands and scores. In: J.C. Alderson& B. North. *Language Testing in the 1990s*. London: British Council/ Macmillan, Developments in ELT, 71-86.

http://books.google.fr/books/about/Language_Testing_in_the_1990s.html?id=d1NfAQAACAAJ&redir_esc=y

Alderson, J.C & Banerjee, J. (2001). Language testing and assessment (Part 1). *Language teaching*, 34 (4), 213-36.

http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=2768484&fileId=S0261444800014464

Bachman, L.F. (2004). Statistical Analysis for Language Assessment. Cambridge: CUP.

http://books.google.de/books?id=P4NkTF_rhIC&printsec=frontcover&dq=Bachman,+L.F.+(2004).+Statistical+Analysis+for+Language+Assessment.+Cambridge:+CUP.+between+Learning+and+Assessment.

Bachman, L.F & Clark, J.L.D. (1987). The measurement of Foreign/Second Language Proficiency. *Annals of the American Academy of Political and Social Science,* Vol. 490, Foreign Language Instruction: A National Agenda, 20-33.

http://www.jstor.org/stable/1045233

Bailey, C.H.N. (1996). Essays on Time-based Linguistic Analysis. Clarendon Press, Oxford.

http://www.questia.com/library/7958492/essays-on-time-based-linguistic-analysis

Beacco, J.-C. & Byram, M. (2007). From linguistic diversity to plurilingual education: Guide for the development of language education policies in Europe. Strasbourg, France: Council of Europe, Language Policy Division.

http://www.coe.int/t/dg4/Linguistic/Source/FullGuide_En.pdf

Candlin, C. (1987). Towards task-based language learning. In C. Candlin & D. Murphy (Eds.) *Language learning tasks*. Englewood Cliffs, New York: Prentice Hall, 5-22.

Chapelle, C. (1999a). From reading theory to testing practice. In: Chalhoub-Deville, M. (Ed). *Issues in Computer-adaptive Testing of Reading*. Cambridge: Cambridge University Press, 150-166.

http://books.google.fr/books?hl=de&lr=&id=xBCDHYrMgKgC&oi=fnd&pg=PA150&dq=Chapelle,+C.+(1999a).+From+reading+theory+to+testing+practice

Coste, D. (1976). Un niveau-seuil. Conseil de la coopération culturelle du Conseil de l'Europe. Conseil de l'Europe. Paris : Hatier

http://www.coe.int/t/dg4/linguistic/publications_fr.asp

Hymes, D. (1972). On communicative competence. In: Pride, J.B. and Holmes, J. (Eds). *Sociolinguistics: Selected Readings.* Harmonsworth: Penguin, 269-293.

http://people.reed.edu/~lalzimman/LING212/PDFs/Hymes1972.pdf

Ockey, Gary J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal*. Vol. 93, Focus Issue: Technology in the Service of Language Learning: Update on Garrett (1991) Trends and Issues (2009), 836-847.

http://www.jstor.org/stable/25612278

Trim, J. et al. (2001). Cadre européen commun de référence pour les langues: apprendre, enseigner, évaluer. Guide pour les utilisateurs. Strasbourg : Division des Politiques linguistiques.

http://media.leidenuniv.nl/legacy/common-european-framework-of-reference-for-languages,-a-guide-for-users.pdf

Van Ek, J. A. & J.L.M. Trim (1990). Waystage. Cambridge: Cambridge University Press (CUP).

http://www.coe.int/t/dg4/linguistic/publications_fr.asp

Van Ek, J. A. & J.L.M Trim (2001). Vantage level. Cambridge: Cambridge University Press (CUP).

http://www.coe.int/t/dg4/linguistic/publications_fr.asp

Vongpumivitch, V & N. (2001). An Interview with J. Charles Alderson. Regents of the University of California

2. Autres documents

2.1 Tests de langue pour l'anglais

Cutting Edge Placement Test. New York: Longman 2007

http://englishtips.org/1150791650-cutting_edge__placement_tests.html

DIALANG. 2001

http://www.lancaster.ac.uk/researchenterprise/dialang/about

Energy Placement Test. Pearson Education 2004

http://www.pearson.pl/pub/angielski/uploaddocs/placements_tests/Energy_Placement_Test.pdf

Oxford Quick Placement Test. Version 1. 2001

https://www.vhs-aschaffenburg.de/documents/5000/Oxford_Test.pdf

Oxford Placement Test 2. Dave Allan (ed.) 1992

http://www.amazon.com/Oxford-Placement-Tests-Test-Pack/dp/019432804X

Success Placement Test. Pearson Longman 2007

http://www.scribd.com/doc/54165945/Placement-Tests

Upstream Placement Test. Egis 2010

http://de.scribd.com/doc/27535540/Placement-Test-Upstream-Enterprise

2.2 Manuels scolaires

Christie, D. (2006). (Hrsg.) Gateway: Englisch für berufliche Schulen: Stuttgart: Klett

East, P. & McCredie, B. (2007). Englisch 9/10. Klasse, Köln: Komet

Knop, B. C. et al. (1998). (Hrsg.) Words in context: thematischer Oberstufenwortschatz, Taschenbuch. Stuttgart: Klett

Otte, M.D. (2004). Englisch komplett, 5/8.Schuljahr, Auflage 1. Stuttgart: Klett

Schwarz et al. (1985). (Hrsg.) English G. Band A1 für das 5.Schuljahr an Gymnasien, 1. Auflage Berlin: Cornelsen

Schwarz, H. (2001). (Hrsg.) English G. 2000. A5. Berlin: Cornelsen

Schwarz, H. (2008). (Hrsg.) English G 21. Workbook, Berlin: Cornelsen

Thaler, E. (2007). (Hrsg.) The new Summit: Text and Methods, Braunschweig: Bildungshaus

Weisshaar, H. (2006). (Hrsg.) Green Line 2. Textbook and workbook. Stuttgart/ Leipzig: Klett

Weisshaar, H. (2007). (Hrsg.) Green Line 3. Textbook and workbook. Stuttgart/ Leipzig: Klett

Weisshaar, H. (2010). (Hrsg.). Green Line Oberstufe Klasse 10, Stuttgart/ Leipzig: Klett

Weisshaar, H. (2009). (Hrsg.). Green Line Oberstufe, Stuttgart/ Leipzig : Klett

Weisshaar, H. (2009). (Hrsg.) Green Line: Oberstufe Skill and Examtrainer. Nordrhein- Westfalen. Stuttgart/Leipzig: Klett

2.3 Rapports et référentiels institutionnels

ALTE (2002). The ALTE CAN DO Project. Articles and Can Do statements produced by the members of ALTE 1992-2002

http://www.alte.org/attachments/files/alte_cando.pdf

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) (1999). Standards for Educational and Psychological Testing, Washington, DC: AERA.

www.aera.net

Breakthrough (2001). An objective at Level A1 of the Common European Framework of Reference for Languages, Learning, Teaching, Assessment (CEFR)

http://www.coe.int/t/dg4/linguistic/publications_fr.asp

Conseil de l'Europe (2001). Un cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer. Paris : Les Editions Didier.

http://www.coe.int/t/dg4/linguistic/Source/Framework_fr.pdf

Alderson, Ch. (2002). Common European Framework of Reference for Languages. Learning, Teaching, Assessment. Case Studies. Strasbourg

http://www.coe.int/t/dg4/linguistic/publications_fr.asp

Conseil de l'Europe (2005). Un cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer. Paris : Les Editions Didier.

http://www.coe.int/t/dg4/linguistic/source/framework_fr.pdf

Conseil de l'Europe (2007). DE LA DIVERSITE LINGUISTIQUE A L'EDUCATION PLURILINGUE : GUIDE POUR L'ELABORATION DES POLITIQUES LINGUISTIQUES EDUCATIVES EN EUROPE. STRASBOURG

http://www.coe.int/t/dg4/linguistic/publications_fr.asp

Conseil de l'Europe (2008) : Division des Politiques Linguistiques. Paris : Les Editions Didier.

http://www.coe.int/t/dg4/linguistic/Source/leaflet_LPD_%20Aug08_FR.pdf

Conseil de l´Europe (2009). Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual. Strasbourg: Language Policy Division

www.coe.int/lang

Haahr, J. et al. (2004). Defining a strategy for the direct assessment of Skills: final report. Danish Technological Institute

http://www.ec.europa.eu/education.pdr/doc288_en.pdf

Haahr, J. & M.E. Hansen (2006). Adult Skills Assessment in Europe. Feasibility Study. Policy and Business Analysis. Final Report

http://www.pedz.uni-mannheim.de/daten/edz b/omb/06/adult_skills_assessment.pdf

ILTA (2000). Code of ethics. International Language Testing Association. Available on-line.

http://www.iltaonline.com/images/pdfs/ILTA_Code.pdf

ILTA: Draft code of practice: Version 3. (2005, June 21). Retrieved November 12, 2006.

http://www.iltaonline.com/images/pdfs/ILTA_Guidelines.pdf

Schneider G. & Lenz P. (2004) European Language Portfolio: Guide for Developers

http://www.coe.int/t/dg4/education/elp/elpreg/Source/Publications/Developers_guide_EN.pdf

CNDP (2002). Classe de Seconde Générale et Technologique. Langue vivante 1&2. Le B.O. N°7, HORS-SÉRIE. Futuroscope

http://www2.cndp.fr/lesScripts/bandeau/bandeau.asp?bas=http://www2.cndp.fr/doc_administrative/programmes/accueil.htm

CNDP (2003). Programmes des Lycées Anglais. Classes de Première Générales et Technologiques Le B.O. 11, N°, HORS-SÉRIE. Futuroscope

http://www2.cndp.fr/archivage/valid/154066/154066-22617-28669.pdf

CNDP (2006). PROGRAMMES DE L'ENSEIGNEMENT DE LANGUES VIVANTES ÉTRANGÈRES AU COLLÈGE. PRÉAMBULE COMMUN. Anglais palier 1. Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche. Direction générale de l'enseignement scolaire. Futuroscope

http://cache.media.eduscol.education.fr/file/LV/72/1/Programme_anglais_palier1_123721.pdf

CNDP (2008). PROGRAMMES DE L'ENSEIGNEMENT DE LANGUES VIVANTES ÉTRANGÈRES AU COLLÈGE. PRÉAMBULE COMMUN. Anglais palier 2. Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche. Direction générale de l'enseignement scolaire. Futuroscope

http://cache.media.eduscol.education.fr/file/LV/19/1/Programme_anglais_palier2_120191.pdf

Van Ek, J. A/ & Trim, J. L. M. (1975). Threshold Level. Cambridge University Press (CUP).

http://www.coe.int/t/dg4/linguistic/publications_fr.asp

Van Ek, J.A. (2001). Objectifs de l'apprentissage des langues vivantes. Strasbourg : Conseil de l'Europe

https://www.google.fr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0CCsQFjAB&url

2.4 Articles de presse

Begos, K. (17.08. 2012). Carbon dioxide surprise sees decreased levels being released. Associated press

http://articles.philly.com/2012-08-17/news/33249254_1_natural-gas-fall-in-coal-prices-energy-department

Gutterman, S. & de Carbonnel, A. (20.08.2012). UPDATE 1-Russian police pursuing other members of Pussy Riot. Reuters

http://in.reuters.com/article/2012/08/20/russia-pussyriot-idINL6E8JKATO20120820

Karon, T. (22.08. 2012). As South Africa Reels from Mine Shootings, Social Inequality Threatens to Undo the Post-Apartheid ‘Miracle’. Time

http://world.time.com/2012/08/22/as-south-africa-reels-from-mine-shootings-social-inequality-threatens-to-undo-the-post-apartheid-miracle/

Park, A. (23.08.2012). Older Fathers Linked to Kids’ Autism and Schizophrenia Risk. Time

http://healthland.time.com/2012/08/23/older-fathers-linked-to-kids-autism-and-schizophrenia-risk/

Wearden, G. (Wednesday 22.08.2012). Eurozone crisis: Greek hopes for leniency over austerity set back. The Guardian

http://www.theguardian.com/business/2012/aug/22/eurozone-crisis-greek-leniency-austerit

Résumé

Il existe un nombre croissant de tests de langues sur le marché. Ces derniers se répartissent en plusieurs catégories selon leur forme et leur fonction: tests de positionnement, d'acquisition et de progrès, de compétence, de certification, tests diagnostiques et enfin tests d'aptitude. La conception du test de positionnement POSILANG, entièrement adossé au *Cadre européen commun de référence pour les langues* (CECRL), force à s'interroger sur le processus général d'évaluation des compétences en LVE, à définir ce que sont les bonnes pratiques, à explorer la compatibilité de l'approche communicative-actionnelle prônée par le Cadre avec le format automatisé. Les atouts visés par le dispositif sont la gratuité, la praticité, la fiabilité, l'accessibilité, l'interactivité ainsi que l'authenticité. L'élaboration des items doit tenir compte des usages situés de la langue dans des situations de communication réelles. Conçu localement, en intégrant la situation particulière du site universitaire bordelais, le test est prévu pour évaluer le niveau de compétences des bacheliers en anglais, accédant à l'Enseignement Supérieur.

L'évaluation de la version pilote de POSILANG, menée dans le cadre de la présente recherche, montre que POSILANG permet bien de déterminer à la fois le niveau en langue global de candidats et leur niveau par domaine de compétence. Le test remplit aussi une fonction diagnostique, en pointant les lacunes des candidats. Le positionnement par domaines de compétences permet de créer différents groupes de niveaux aussi homogènes que possible pour les enseignements obligatoires d'anglais et pour remédier aux difficultés langagières repérées.

Mots clés:

Test de langue – Positionnement – Evaluation – Fonction diagnostique - Compétence langagière - CECRL

Summary

In education today, a great number of language tests are available. These tests can be divided into different categories according to their features and function: placement tests, acquisition and progress tests, certification tests and aptitude tests.

The conception of the POSILANG placement test has required an analysis of the overall process of assessing language competences in order to determine what can be considered as good practice. As the test is conceived in accordance with the Common European framework of reference for languages (CEFR), the compatibility between the automated format of the test and the action-oriented approach advocated by the framework also needed to be explored. POSILANG is conceived to be free of charge, readily available, reliable, interactive and authentic. The latter is achieved by using language taken from real life communication situations for the conception of items. The test is designed for local use on the campus of Bordeaux University and takes into account the specificities of this environment. Its aim is to assess the competence level of secondary school leavers as they enter the higher education system.

Through testing of the pilot version of POSILANG, this study has found that the test is able to assess both the overall language level of participants and their level in three separate areas of competence. It has a diagnostic function centred on pinpointing the weaknesses of students. The placement of students according to the different areas of competence enables teachers to form homogeneous student groups geared at addressing the specific instructional needs identified by the test.

Key words:

Language testing - Placement test – Language skills assessment – Language ability- CEFRL