Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus

Karen Fort

UNIVERSITÉ PARIS 13 — SORBONNE PARIS CITÉ
LIPN — UMR CNRS 7030

T H È S E

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS 13

Discipline : Informatique

présentée et soutenue publiquement par

Karën FORT

le 7 décembre 2012

Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus

Composition du jury

Frédéric Béchet	Professeur, Université de la Méditerranée	Rapporteur
Claire François	Ingénieure de Recherche, INIST-CNRS, Nancy	Examinatrice
Benoît Habert	Professeur, ENS de Lyon	Rapporteur
Lori Lamel	Directrice de Recherche, LIMSI-CNRS, Orsay	Présidente
François Lévy	Professeur émérite, Université Paris 13	Examinateur
Adeline Nazarenko	Professeure, Université Paris 13	Directrice
Eric Villemonte de la Clergerie	Chargé de Recherche, INRIA, Rocquencourt	Examinateur

Remerciements

Je tiens à remercier ici Adeline Nazarenko, qui a accepté de m'encadrer en tant que doctorante et d'accompagner la personne que je suis à travers ces deux aventures fondamentales que sont la thèse et la maternité. Cette interaction m'a apporté plus que je ne saurais l'écrire ici.

Je remercie également mes rapporteurs, Benoît Habert et Frédéric Béchet, pour la qualité et la richesse de leurs retours sur ce travail. Merci aux autres membres du jury, Lori Lamel, François Lévy et Éric de la Clergerie d'avoir accepté d'en faire partie.

Sans certaines personnes, rien de tout cela ne serait arrivé.
Merci à Bruno Guillaume et Claire François d'avoir fait le deuil de leur ingénieure et de m'avoir offert les moyens de cette thèse.
Merci à Christian Boitet d'y avoir cru, bien avant moi.

Je dois également beaucoup à Benoît Sagot et Sophie Rosset, avec qui il fait si bon faire de la recherche, et qui m'ont apporté énormément. Yann Mathet et Thibault Mondary ont toujours répondu présent, je les en remercie.

Plus généralement, je remercie tous mes co-auteurs pour leur énergie, leur intelligence, leur envie. Cette thèse leur doit beaucoup.

Merci aux annotateurs de l'INIST-CNRS, Françoise Tisserand, Bernard Taliercio, Claire Ris, Alain Zerouki et Flora Badin, dont le travail a formé le terreau de cette recherche. Le cadre du programme Quæro a été particulièrement riche de rencontres et d'interactions, j'en remercie les organisateurs et les participants.

Enfin, les derniers mois de cette thèse ont été effectués au sein de l'équipe Sémagramme[1]*, du Loria, que je remercie ici dans son ensemble, avec une pensée particulière pour Guy Perrier et Bruno Guillaume, mes hébergeurs de toujours.*

Mes amis Jean-Christophe Bach, Paul Bédaride, Florent Pompigne et Sylvain Raybaud ont su être là au cours de ces années et leur soutien m'a porté jusqu'au bout. Merci.
Merci également à Alain Couillault qui m'a aidée à franchir les derniers mètres, en remettant tout en questions.

Je remercie également ma mère, Michèle Gabert, d'avoir répondu présente à un moment difficile et de m'avoir rendu la monnaie de ma pièce en relisant ce document.

1. http://semagramme.loria.fr/

Surtout, merci à mon compagnon Michel Ancé, qui m'a soutenue et supportée à travers mes hauts et mes bas, mes disparitions et mes doutes,
et à Léonard, qui a tout changé.

Suzanne, tu es toujours là.
Merci.

Sommaire

Chapitre 1

Introduction

L'annotation manuelle de corpus occupe aujourd'hui une place importante en Traitement Automatique des Langues (TAL). S'agissant d'une activité humaine, il est tentant de s'en remettre aux méthodes de gestion de projet et au simple bon sens pour résoudre les problèmes qu'elle pose. Nous allons cependant montrer dans cette thèse que cette activité peut être, au moins en partie, formalisée, puis outillée, ce qui permet d'en réduire les aléas et d'en assurer la qualité.

1.1 Contexte

Cette thèse a été effectuée à l'INIST-CNRS [1] (Institut de l'Information Scientifique et Technique), à Nancy et au LIPN (Laboratoire d'Informatique de Paris Nord), à Villetaneuse. Elle a été financée dans le cadre de Quæro [2], un programme collaboratif franco-allemand privé-public portant sur le traitement automatique de contenus numériques multimédias multilingues. Elle a été encadrée scientifiquement par Adeline Nazarenko, au LIPN, et administrativement, à l'INIST-CNRS, par Claire François.

Cette thèse s'inscrit, au sein de Quæro, dans le projet *Corpus*, dont l'existence montre l'importance donnée aux corpus (en particulier annotés) dans le programme. Une partie de notre travail à l'INIST-CNRS a consisté à gérer des campagnes d'annotation manuelle décidées dans le cadre de ce projet. Nous en avons ainsi organisé trois : une concernant les noms d'espèces, de gènes et de protéines (*Microbiologie*), la deuxième les relations de renommage de gènes (*Renommage*) et la dernière les entités nommées, actions et relations en football (*Football*). Nous avons également participé à des campagnes d'annotation dans le cadre de Quæro sans en être la gestionnaire. Ces campagnes ont consisté à annoter des entités nommées, termes et relations en pharmacologie (*Pharmacologie*) et des entités nommées structurées (*EN1* et *EN2*). Toutes les campagnes dans lesquelles nous avons été impliquée sont présentées dans la figure 1.1 et sont détaillées dans le chapitre 5.

Ce rôle de gestionnaire de campagne nous a permis d'acquérir de l'expérience dans le domaine de l'annotation manuelle de corpus et cette thèse en est le fruit.

Cette thèse est également le reflet de la richesse des échanges encouragés par Quæro. Nous avons en effet pu initier de nombreuses collaborations autour de cette recherche, aussi bien en interne au programme, qu'à l'extérieur de celui-ci.

1. http://www.inist.fr/
2. http://www.quaero.org

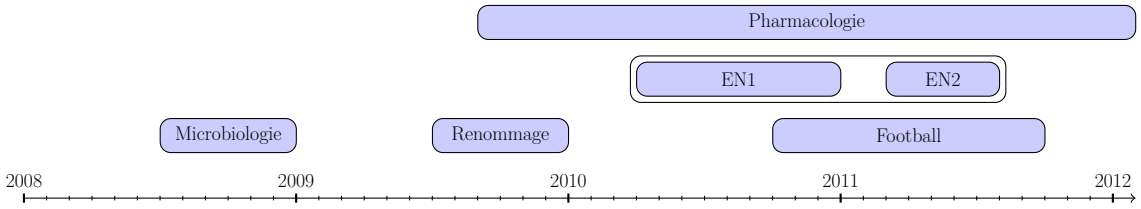


FIGURE 1.1: Répartition temporelle des campagnes d'annotation auxquelles nous avons participé.

1.2 Enjeux de l'annotation manuelle de corpus

Cette recherche se situe dans le domaine du TAL et plus précisément de l'analyse de contenu textuel.

On assiste en effet depuis les années 90 à un regain d'intérêt pour les corpus dans le domaine du TAL, en particulier les corpus annotés. Ce renouveau est poussé par la pression des besoins applicatifs, et est dû à la fois aux progrès réalisés en termes de matériel (capacité de stockage et de traitement), au développement des techniques d'apprentissage (aussi bien symboliques que statistiques) qui utilisent des corpus d'entrainement, et au succès des campagnes d'évaluation (dont celles du programme Quæro), campagnes qui nécessitent des corpus de référence.

De la qualité des corpus annotés manuellement dépend assez directement la qualité des outils créés à partir de ces corpus ou de l'évaluation qui les utilise. Ces corpus annotés doivent donc offrir la meilleure qualité d'annotation possible, ce qui implique de faire intervenir des experts humains dans le processus d'annotation, que ce soit pour annoter directement le corpus ou pour corriger une annotation réalisée automatiquement. Cette phase manuelle est extrêmement fastidieuse et nécessite un travail de longue haleine, de qualité si possible constante. En outre, le coût de développement manuel de ressources linguistiques en général, et de corpus annotés en particulier, est notoirement élevé. En fonction d'un besoin applicatif donné, il faut donc trouver un équilibre entre la qualité attendue, le coût de l'annotation et le volume à annoter.

Nous présentons ici quelques exemples de corpus annotés considérés comme des références, afin d'illustrer cet état de fait.

1.2.1 Le *Penn Treebank*

Le corpus arboré de l'anglais de l'Université de Pennsylvanie, plus connu sous le nom de *Penn Treebank*, a été créé entre 1989 et 1992, pour la partie morpho-syntaxique (*part-of-speech tagging*), et jusqu'en 1994 pour la partie syntaxe (*bracketing*) de la première version.

Le *Penn Treebank* comprend, en novembre 1992, 4,8 millions de tokens[3] en anglais américain provenant de neuf sources différentes, dont le Brown corpus ré-annoté. Tous ces tokens ont été annotés en morpho-syntaxe, et une partie en a été annotée en syntaxe (presque trois millions de tokens). Tous les tokens de cette dernière partie sont donc annotés par une catégorie morpho-syntaxique et compris dans un ensemble plus large, qui est lui-même annoté en catégories syntaxiques. Dans les deux cas, une annotation automatique (pré-annotation) a précédé une correction manuelle. Les annotateurs ont utilisé pour cela l'éditeur `Emacs` enrichi d'un *package* spécifique à chacune des tâches.

Le jeu de catégories morpho-syntaxiques utilisé comprend 36 catégories principales et 12 catégories pour la ponctuation et les symboles. A la différence du corpus arboré du français (Abeillé *et al.*, 2003), les catégories du *Penn Treebank* ne sont pas explicitement hiérarchisées. La phase d'apprentissage des annotateurs pour la partie morpho-syntaxique a été de moins d'un mois, à raison de 15 heures par semaine. Après un mois, leur vitesse de correction a dépassé les 3 000 mots à l'heure.

Le jeu de catégories syntaxiques comprend lui une quinzaine de catégories. La phase d'apprentissage a été sensiblement plus longue pour la partie syntaxe (environ deux mois), la vitesse d'annotation des annotateurs passant de 375 mots par heure après trois semaines à 475 mots par heure après six semaines. Ces performances ont encore été améliorées en réduisant la structure syntaxique à une structure plus plate, puis en permettant aux annotateurs de ne pas distinguer entre arguments et circonstants dans les cas ambigus. Au final, le plus rapide d'entre eux annotait plus de 1 500 mots par heure.

1.2.2 Le *Prague Dependency Treebank* (PDT)

Le corpus arboré du tchèque, ou *Prague Dependency Treebank* (PDT) a été créé entre 1996 et 2004 (Böhmová *et al.*, 2001). Construit à partir du corpus national du tchèque (*Czech National Corpus*), il présente une structure à trois niveaux : morphologique, analytique (syntaxe de dépendance), et ce que ses créateurs appellent tectogrammatical (sens linguistique annoté à l'aide de la description fonctionnelle générative ou *Functional Generative Description*).

La Version 1.0 du corpus arboré du tchèque inclut l'annotation manuelle des niveaux morphologique (1,8 millions de tokens) et analytique. Fait rare, nous disposons, pour ce corpus, d'informations détaillées sur son coût exact. Il a nécessité cinq ans de travail

3. Le terme « token » est généralement employé dans ce document au sens d' « une séquence de caractères présent[e] dans le corpus et séparé[e] de ses voisins par des espaces ou par certaines autres marques typographiques (ponctuation,...) qui varient selon la langue. » (Sagot et Boullier, 2008).

et a impliqué 22 personnes (dont au maximum 17 en parallèle) [4]. Le coût final de ce corpus a été évalué à environ 600 000 dollars.

1.2.3 GENIA

Le corpus GENIA (Kim *et al.*, 2003), aujourd'hui en version 3.0, comprend 2 000 titres et résumés de la base MEDLINE [5] (soit plus de 400 000 mots) annotés sémantiquement en biologie (près de 100 000 annotations). Ce corpus est disponible [6] et se présente sous format XML. Le corpus GENIA a été créé explicitement pour la fouille de texte et est présenté comme une référence (*gold standard*) dans le domaine de la biologie.

L'annotation du corpus a été réalisée manuellement par deux experts du domaine, qui ont utilisé pour cela les descripteurs de l'ontologie GENIA. Au final, GENIA contient 9 372 phrases et son annotation a nécessité cinq annotateurs à temps partiel, un coordinateur sénior et un coordinateur junior pendant un an et demi (Kim *et al.*, 2008).

1.2.4 Redéfinir le coût

Le coût est généralement évalué en fonction du temps nécessaire pour obtenir l'annotation et du nombre de personnes impliquées, voire en fonction du nombre d'interactions nécessaires avec le système d'aide à l'annotation (Felt *et al.*, 2010). Cependant, une telle définition est limitative, car elle ne tient pas compte de la ré-utilisabilité dudit corpus. Or, cette ré-utilisabilité dépend de la qualité, en termes de cohérence et de fiabilité, du corpus produit. Ainsi, Cohen *et al.* (2005) ont montré, dans le domaine biomédical, que la ré-utilisation d'un corpus dépend avant tout de la cohérence de l'annotation par rapport à une documentation (qui doit être disponible), de la maintenance du corpus et de sa disponibilité dans un format « standard ».

La maintenance, en l'occurrence, peut représenter un effort lourd, car il faut parfois reprendre toute l'annotation. Ce n'est pas anormal, ni même inhabituel, mais il est possible de limiter les déviations dès le début de la campagne.

Nous sommes convaincue qu'un niveau de qualité élevé ne peut être obtenu que par l'application d'une méthodologie stricte, qui permet de penser en amont l'arbitrage entre investissement et retour sur investissement.

4. Le temps passé par chacune de ces personnes sur l'annotation n'est pas disponible, il nous est donc impossible de donner une équivalence en homme-mois.
5. http://www.ncbi.nlm.nih.gov/pubmed
6. http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA

1.3 Une maîtrise insuffisante du processus d'annotation

Nous avons constaté, lors des campagnes d'annotation auxquelles nous avons participé, que le processus d'annotation manuelle de corpus est encore mal connu, peu documenté et, finalement, mal maîtrisé.

1.3.1 Des outils utiles, dont l'influence est mal évaluée

De nombreux outils d'aide à l'annotation existent aujourd'hui, permettant d'améliorer les conditions et la qualité de l'annotation réalisée. Ces outils sont très utilisés, leur influence sur les campagnes d'annotation est donc importante. Or, aucun ne permet de préparer l'annotation, peu d'entre eux proposent d'évaluer l'annotation (par le biais du calcul d'un accord inter-annotateurs, par exemple) et seuls deux d'entre eux séparent clairement les rôles des différents utilisateurs. Ces étapes fondamentales d'une campagne d'annotation sont donc généralement ignorées ou réalisées à la hâte, souvent de manière implicite et sans méthode.

Par ailleurs, si dans certains cas comme l'annotation morpho-syntaxique, une pré-annotation automatique paraît indispensable tant elle fait gagner en temps et en qualité (Dandapat *et al.*, 2009), dans quelle mesure l'outil de pré-annotation doit être de bonne qualité ? surtout, quels sont les biais introduits par cette pré-annotation ? Ces questions sont fondamentales si l'on veut connaître et contrôler les effets de la pré-annotation automatique, afin de profiter au mieux des bénéfices qu'elle apporte.

Enfin, les méthodes de création de ressources langagières par des « non-expert » via le *crowdsourcing* (travail réalisé sur Internet par une foule d'intervenants) sont aujourd'hui très à la mode et semblent prometteuses en terme de coût, puisque les annotateurs ne sont pas ou peu rémunérés. Il est cependant très difficile d'obtenir une évaluation précise du coût réel de l'annotation par ces moyens. Les calculs effectués prennent en effet rarement en compte le développement nécessaire en amont (notamment la mise au point de l'interface, voir un exemple dans (Hong et Baker, 2011)) et en aval (la correction du travail réalisé, voir par exemple (Xu et Klakow, 2010), ou la publicité pour attirer les participants (Chamberlain *et al.*, 2009b)). Quant à la qualité produite sur ces outils, elle est extrêmement variable.

1.3.2 Une complexité peu étudiée

Il n'existe à ce jour aucune grille permettant d'évaluer précisément la complexité de l'annotation envisagée. Les difficultés que présente l'étiquetage morpho-syntaxique du hindi (Dandapat *et al.*, 2009) ne sont pas les même que celles de l'annotation de renommages de gènes, mais on peine à les définir, donc à les réduire.

Très peu de publications traitent de ces difficultés. Dans le domaine des sciences cognitives, Tomanek *et al.* (2010) ont utilisé un dispositif de suivi du regard (*eye-tracking device*) pour analyser et modéliser la complexité cognitive de l'annotation. Ils ont conduit leurs expériences sur une tâche simple d'annotation d'entités nommées (avec les catégories *Personnes*, *Lieux* et *Organisations*), avec pré-identification de groupes nominaux complexes contenant au moins une entité nommée potentielle. Ils ont mesuré l'influence des complexités syntaxique et sémantique [7], ainsi que la taille du contexte utilisé par les annotateurs. Les résultats montrent que les performances d'annotation tendent en moyenne à être corrélées « avec la complexité sémantique des syntagme annotés » [8] et moins avec leur complexité syntaxique. Par ailleurs, la taille du contexte utilisé dépend elle-aussi de la complexité des syntagme annotés.

Aussi intéressantes qu'elles puissent être, leurs conclusions ne s'appliquent qu'à une tâche simple d'annotation en entités nommées et ne couvrent pas tout le processus d'identification des unités à annoter puisqu'une pré-identification a été appliquée. Il nous semble en outre que le domaine du TAL lui-même peut apporter des éléments d'analyse sur le sujet.

1.3.3 Des mesures d'évaluation souvent peu adaptées

Enfin, si l'évaluation des ressources annotées est aujourd'hui considérée comme fondamentale et si la plupart des publications concernant des corpus annotés fournissent aujourd'hui une mesure d'accord inter-annotateurs, l'adéquation entre la mesure choisie et le type d'annotation est encore peu considérée.

En particulier, les mesures de la famille des Kappas, dont le très utilisé Kappa de Cohen (Cohen, 1960), qui permettent de prendre en compte le hasard, ne sont pas adaptées à toutes les annotations. Il est indispensable de préciser leur usage et de mieux les caractériser.

1.4 Objectifs et plan

Étant donné le coût élevé de l'annotation manuelle de corpus, le niveau de qualité attendu est généralement élevé. Or, l'activité d'annotation reste mal connue et les outils utilisés peu adaptés. Par ailleurs, l'évaluation de la qualité obtenue est souvent réalisée avec des mesures inappropriées ou présentées de manière incomplète.

7. Respectivement, en mesurant le nombre de nœuds dans l'arbre syntaxique et la fréquence inverse de document dans le syntagme, en fonction d'un corpus de référence.

8. *« [To] correlate with the [semantic] complexity of the annotation phrase »*

Cette thèse propose donc une étude systématique des différentes dimensions de l'annotation manuelle de corpus et en particulier des complexités qu'elle présente, à travers les expériences menées dans le cadre du programme Quæro. Les solutions que nous proposons ont vocation à s'intégrer dans un ensemble cohérent formant une méthodologie globale pour l'annotation manuelle de corpus.

La première partie de notre état de l'art (partie I) va consister, après avoir brossé un rapide aperçu historique de l'annotation manuelle de corpus (section 2.1), et après avoir donné les définitions nécessaires à notre propos (section 2.2), à montrer que l'annotation manuelle de corpus en TAL est aujourd'hui faite d'expériences disparates sans vision synthétique et qu'elle tire peu profit des expériences précédentes (sections 2.3 et 2.4). Dans un second temps, nous présentons rapidement les principaux formats existants pour l'annotation (section 3.1) et plus longuement les outils (section 3.2) d'aide à l'annotation, ainsi que les mesures d'évaluation les plus utilisées (section 3.3). Enfin, nous décrivons certaines solutions (section 3.4) permettant d'alléger le coût et d'améliorer la qualité de l'annotation manuelle.

La deuxième partie de la thèse (partie II) détaille les éléments de méthodologie que nous proposons pour organiser une campagne (chapitre 4), assis sur les campagnes auxquelles nous avons participé (chapitre 5). Nous réalisons dans le chapitre 6 une analyse des dimensions de complexité d'une campagne d'annotation.

Pour finir, nous présentons en partie III les recherches que nous avons menées pour outiller le gestionnaire de campagne d'annotation. Cela concerne la pré-annotation automatique (chapitre 7) et les mesures d'évaluation (chapitre 8). Nous présentons également dans un dernier chapitre (chapitre 9) les processus en œuvre et les outils nécessaires dans une campagne d'annotation.

Première partie

État de l'art

Chapitre 2

Formes et types d'annotations

2.1 Historique

Nous présentons ici un bref historique du terme et de l'activité d'annotation, afin d'en montrer les différentes facettes, mais également les constantes.

2.1.1 Qu'est-ce que l'annotation ?

Origine

Le *Dictionnaire de l'Académie*[1] définit une annotation comme étant « une note un peu longue que l'on fait sur un livre pour en éclaircir quelques passages ». Le *TLFi*[2] la définit plutôt au pluriel, *les* annotations étant des « remarques manuscrites notées en marge d'un texte » (voir figure 2.1).

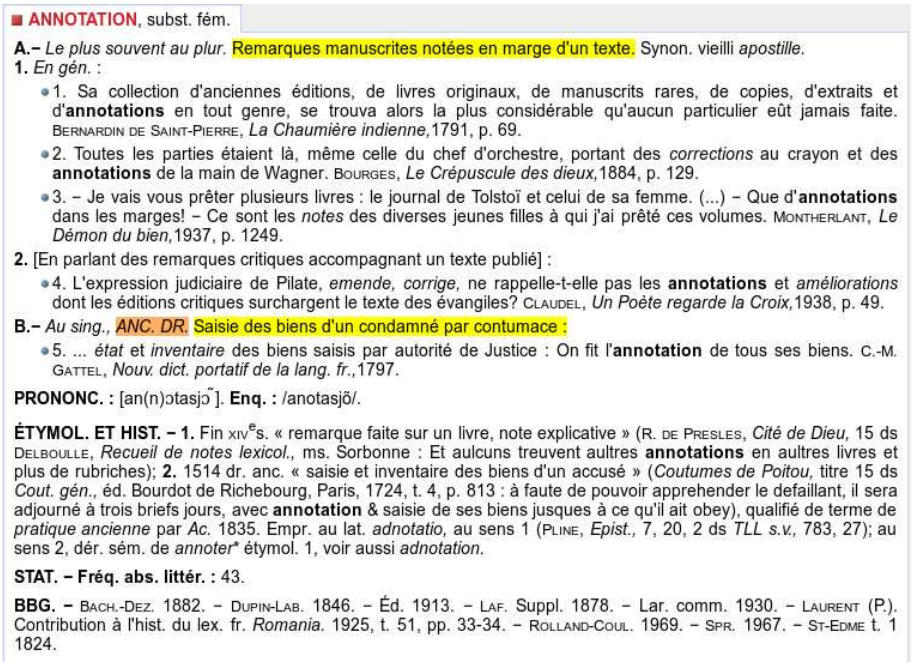
La partie étymologie de cette définition (voir figure 2.1) montre que le sens de ce mot a peu évolué, puisqu'une annotation est à l'origine « [une] remarque faite sur un livre, [une] note explicative ». La première attestation du mot, telle que citée dans le *TLFi*, daterait de la fin du XIVe siècle. Cette affirmation est cependant erronée (B. Stumpf, ATILF-CNRS, communication personnelle, le 15 mars 2012), le mot apparaît en effet pour la première fois dans un incunable de 1486, une traduction par Raoul de Presles de la Cité de Dieu de Saint Augustin.

Il apparaît également (voir figure 2.2) dans le *Complément du dictionnaire de l'ancienne langue française et de tous ses dialectes du IXème au XVème siècle* (Godefroy, 1895-1902), dans lequel il est défini comme « [une] note explicative, [une] chose à remarquer ».

La perspective explicative traverse donc toutes les définitions : une annotation *explique*, apporte des *remarques*, pour *éclaircir*. Dans le *Complément du dictionnaire de l'ancienne langue française et de tous ses dialectes du IXème au XVème siècle* apparaît également une perspective d'indexation : « chose à remarquer ».

Les données annotées considérées dans ces définitions sont surtout textuelles : on annote un *ouvrage*, un *livre*, pour le *Dictionnaire de l'Académie* et le *TLFi*.

1. http://www.cnrtl.fr/definition/academie4/annotation
2. http://www.cnrtl.fr/definition/annotation



ANNOTATION, subst. fém.

A.– *Le plus souvent au plur.* Remarques manuscrites notées en marge d'un texte. Synon. vieilli *apostille.*
1. *En gén.* :

- 1. Sa collection d'anciennes éditions, de livres originaux, de manuscrits rares, de copies, d'extraits et d'**annotations** en tout genre, se trouva alors la plus considérable qu'aucun particulier eût jamais faite. Bernardin de Saint-Pierre, *La Chaumière indienne,*1791, p. 69.
- 2. Toutes les parties étaient là, même celle du chef d'orchestre, portant des *corrections* au crayon et des **annotations** de la main de Wagner. Bourges, *Le Crépuscule des dieux,*1884, p. 129.
- 3. – Je vais vous prêter plusieurs livres : le journal de Tolstoï et celui de sa femme. (...) – Que d'**annotations** dans les marges! – Ce sont les *notes* des diverses jeunes filles à qui j'ai prêté ces volumes. Montherlant, *Le Démon du bien,*1937, p. 1249.

2. [En parlant des remarques critiques accompagnant un texte publié] :

- 4. L'expression judiciaire de Pilate, *emende, corrige,* ne rappelle-t-elle pas les **annotations** et *améliorations* dont les éditions critiques surchargent le texte des évangiles? Claudel, *Un Poète regarde la Croix,*1938, p. 49.

B.– *Au sing., ANC. DR.* Saisie des biens d'un condamné par contumace :

- 5. ... *état* et *inventaire* des biens saisis par autorité de Justice : On fit l'**annotation** de tous ses biens. C.-M. Gattel, *Nouv. dict. portatif de la lang. fr.,*1797.

PRONONC. : [an(n)ɔtasjɔ̃]. **Enq. :** /anotasjõ/.

ÉTYMOL. ET HIST. – 1. Fin xiv^e^s. « remarque faite sur un livre, note explicative » (R. de Presles, *Cité de Dieu,* 15 ds Delboulle, *Recueil de notes lexicol.,* ms. Sorbonne : Et aulcuns treuvent aultres **annotations** en aultres livres et plus de rubriches); **2.** 1514 dr. anc. « saisie et inventaire des biens d'un accusé » (*Coutumes de Poitou,* titre 15 ds *Cout. gén.,* éd. Bourdot de Richebourg, Paris, 1724, t. 4, p. 813 : à faute de pouvoir apprehender le defaillant, il sera adjourné à trois briefs jours, avec **annotation** & saisie de ses biens jusques à ce qu'il ait obey), qualifié de terme de *pratique ancienne* par *Ac.* 1835. Empr. au lat. *adnotatio,* au sens 1 (Pline, *Epist.,* 7, 20, 2 ds *TLL s.v.,* 783, 27); au sens 2, dér. sém. de *annoter** étymol. 1, voir aussi *adnotation.*

STAT. – Fréq. abs. littér. : 43.

BBG. – Bach.-Dez. 1882. – Dupin-Lab. 1846. – Éd. 1913. – Laf. Suppl. 1878. – Lar. comm. 1930. – Laurent (P.). Contribution à l'hist. du lex. fr. *Romania.* 1925, t. 51, pp. 33-34. – Rolland-Coul. 1969. – Spr. 1967. – St-Edme t. 1 1824.

FIGURE 2.1: Définition du mot « annotation » dans le *TLFi*

Espace sémantique

Dans le dictionnaire des synonymes du CRISCO de Caen [3], on trouve treize synonymes pour « annotation » : « apostille », « commentaire », « critique », « explication », « glose », « indication », « notation », « note », « observation », « réflexion », « remarque », « renvoi » et « scolie ». Le mot entre ainsi dans les dix cliques suivantes ce qui forme l'espace sémantique présenté dans la figure 2.3 et obtenu sur le site du CRISCO [4] :

1. Commentaire, explication, glose, note, remarque, scolie ;
2. Commentaire, critique, glose, note, remarque ;
3. Commentaire, critique, note, observation, remarque ;
4. Notation, note, observation, remarque ;
5. Note, observation, réflexion, remarque ;
6. Apostille, note, renvoi ;
7. Explication, indication, note ;
8. Indication, notation, note ;

3. http://www.crisco.unicaen.fr/des/synonymes/annotation
4. http://www.crisco.unicaen.fr/maitrise/visu.cgi?mot=annotation

ANNOTATEUR, s. m., celui qui fait des annotations :

(1552, Ch. Est., *Dict. lat.-goll.*, dans *Dict. gén.*)

ANNOTATION, s. f., note explicative, chose à remarquer :

Annotation. (Paré, VIII, 35.)

— Inventaire des biens saisis par autorité de justice :

Et en cas que non, que du moings il plaise ordonner audit fisque et au receppveur des *annotations* des biens confisquez qu'ils ayent a payer et furnir prestement a ladicte suppliante les deux mille florins cy dessus mentionnez. (13 nov. 1589, *Requête présentée à son Altesse*, A. mun. Mortagne.)

Richelet dit : Prononcez *anotacion*, et il ajoute : *Annotation* est un peu plus usité qu'*annotateur* et a la mine d'avoir été plus tôt introduit.

FIGURE 2.2: Définition du mot « annotation » dans le *Complément du dictionnaire de l'ancienne langue française et de tous ses dialectes*

9. Indication, note, renvoi ;
10. Note, réflexion, renvoi.

Là encore, la dimension explicative (*commentaire*, *glose*) apparaît clairement, ainsi qu'une dimension d'indexation (*indication*, *renvoi*).

A l'heure du numérique, l'annotation est omniprésente et se définit différemment selon les disciplines et donc l'application considérée. Ainsi, si les applications liées à l'herméneutique, le travail coopératif assisté par ordinateur (Lortal *et al.*, 2006) ou la rédaction assistée par ordinateur conservent cette dimension explicative, les travaux sur les langages de représentation, comme ceux du Web sémantique, la mettent en avant plutôt comme un index permettant d'accéder à l'information.

L'activité d'annotation n'est cependant pas nouvelle et il nous semble important d'en connaître les origines, afin de mieux l'appréhender.

2.1.2 Qu'est-ce qu'annoter ?

Origine

Il est impossible de dater avec précision les premiers usages d'annotations, mais ils remontent sans doute aux premiers écrits dont le support physique permettait d'amender

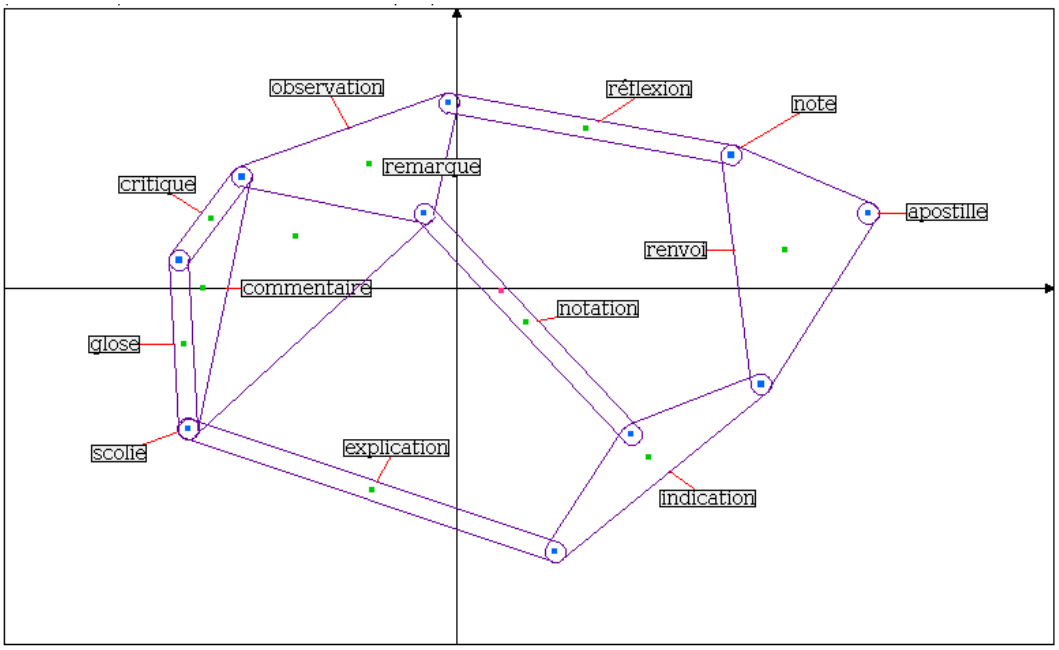


FIGURE 2.3: Visualisation de l'espace sémantique du mot « annotation »

le texte sans trop de difficultés.

Les annotations sont alors d'usage privé (commentaires) ou public (visée explicative), elles permettent également la communication entre rédacteurs (auteurs ou copistes) (Bakhouche *et al.*, 2010).

Dès les premiers manuscrits, on peut ainsi voir apparaître des *gloses*, c'est-à-dire, selon le *TLFi*[5] des « Annotation[s] brève[s] portée[s] sur la même page que le texte, destinée[s] à expliquer le sens d'un mot inintelligible ou difficile ou d'un passage obscur, et rédigée[s] dans la même langue que le texte. » Les *gloses* sont donc utilisées pour informer, former le lecteur. Les *apostilles* sont généralement des annotations portées par l'auteur lui-même ou correspondent à une mise à jour du texte. Les *scolies* sont quant à elles dues à des auteurs antiques.

Formes de l'annotation

La forme des *gloses* elle-même est très variable, Muzerelle (1985, p 134) en distingue ainsi neuf types différents, avec des formes variées. Elles peuvent être ajoutées entre les lignes du manuscrit (*gloses interlinéaires*), dans la marge du texte (*gloses marginales*), sur le pourtour de la page (*gloses encadrantes*) ou entre les paragraphes expliqués (*gloses intercalaires*). Elles sont plus ou moins fusionnées dans le texte, de

5. http://www.cnrtl.fr/definition/glose

la *glose organique* qui peut être considérée comme une partie intégrante du texte, à la *glose formelle* qui constitue en elle-même un texte, transmis de copie en copie (se rapprochant ainsi d'une annotation dite « déportée »).

Des marqueurs, par exemple des *crochets alinéaires*, peuvent également être présents dans le texte (voir un exemple dans un texte de Virgile[6] figure 2.4) afin d'indiquer les premiers mots commentés. Cet ancrage primitif n'indique cependant pas la fin de l'empan glosé.

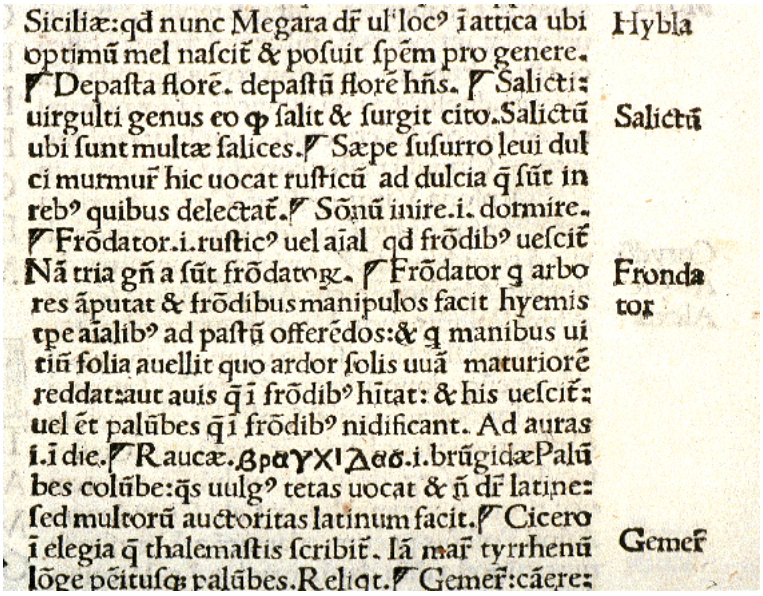


FIGURE 2.4: Crochets alinéaires dans Virgile

Cette limitation se retrouve dans les *auctoritates*, qui apparaissent à partir du VIII^e^ siècle afin de citer les auteurs, considérés comme des autorités, d'une citation. L'ancrage s'effectue par deux points posés au dessus du premier mot de la citation, sans en définir la fin (voir figure 2.5, extraite de (Frunzeanu et Pons, 2012)).

Ce problème de frontières est amplifié par les erreurs des copistes, qui déplacent les *auctoritates* et leurs ancres. Pour remédier à cela, et en l'absence de guillemets (qui n'apparaîtront que beaucoup plus tard), des introducteurs textuels antéposés et périmétriques sont inventés.

Du point de vue du contenu, on est donc passé de la *glose explicative* (texte libre) à la citation d'auteurs (nom de l'auteur, dans une liste d'autorités limitée) précisément identifiés dans le texte. L'ancrage s'est quant à lui amélioré peu à peu, pour finalement ressembler à un balisage textuel.

6. Opera, com. de Servius. Milan : Leonardo Pachel, 1509, in-fol., source : http://enssibal.enssib.fr/bibliotheque/documents/travaux/sordet/nav.liv.ancien.html

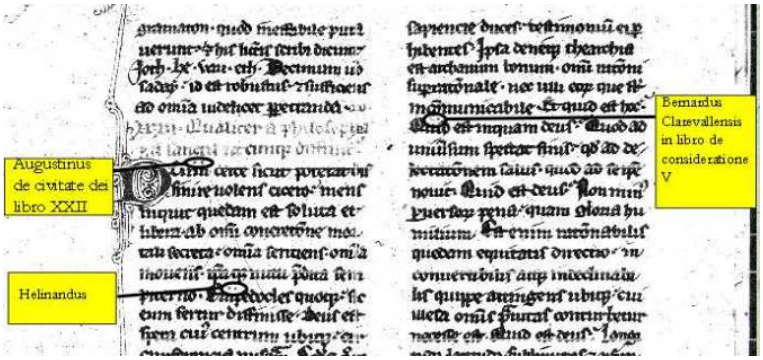


FIGURE 2.5: Ancrage des auctoritates dans *De sancta Trinitate*, Bâle, UB B.IX.5

Cet aperçu rapide montre que de nombreuses problématiques d'aujourd'hui – frontières à délimiter, annotations à laisser libres ou à faire choisir dans un référentiel, ancrage, méta-données à transmettre – étaient déjà présentes hier.

2.2 Définitions

Les définitions de la langue générale pour l'annotation sont bien entendu trop vagues pour notre recherche, il nous faut donc les préciser. D'autre termes seront également utiles pour décrire les concepts liés à l'annotation. Nous les définissons ici.

2.2.1 Annotation

Geoffrey Leech (Leech, 1997) définit l'annotation de corpus comme « la pratique consistant à ajouter des informations linguistiques interprétatives à un corpus de données langagières parlées et/ou écrites. L'"annotation" décrit également le produit final de ce processus. » [7] Si cette définition met l'accent sur la dimension interprétative de l'annotation, elle la limite à des informations linguistiques, et à des supports particuliers, sans citer l'objectif visé par celle-ci.

Habert (2005) précise la définition de Leech en ne limitant plus le type d'information ajouté : « l'annotation consiste à ajouter de l'information (une interprétation stabilisée) aux données langagières : sons, caractères et gestes. » Il ajoute qu'« [e]lle associe deux ou trois volets : (i) segmentation pour délimiter des fragments de données et/ou ajout de points singuliers ; (ii) regroupement de segments ou de points pour leur affecter une catégorie ; (iii) (éventuellement) mise en relation de fragments ou de points ».

7. *« [T]he practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. "Annotation" can also refer to the end-product of this process. »*

Nous définirons quant à nous plus largement l'annotation comme suit :

Définition 1 (Annotation) *L'annotation recouvre à la fois le processus consistant à apposer (*ad-*) une* note *sur un support, l'ensemble des notes ou chaque note particulière qui en résulte et ce, sans préjuger* a priori *de la nature du support considéré (texte, vidéo, images, etc.), du contenu sémantique de la note (note chiffrée, valeur choisie dans un référentiel fermé ou texte libre), de son positionnement global ou local, ni de son objectif (visée évaluative ou caractérisante, simple commentaire discursif).*

Nous nous intéressons ici plus spécifiquement à une pratique de l'annotation qui a cours en TAL et qui consiste à apposer des étiquettes (ou notes) de nature linguistique ou reflétant l'usage des technologies du TAL sur du discours oral ou écrit. Cela recouvre néanmoins une grande diversité de phénomènes puisque ces annotations peuvent varier dans leur nature (étiquettes phonétiques, morpho-syntaxiques, sémantiques ou de pertinence par rapport à une tâche particulière), par leur portée (elles concernent selon les cas quelques caractères, un mot, un paragraphe ou un texte considéré globalement), leur degré de couverture (tout le texte est annoté ou seulement une partie) et leur forme (de la valeur atomique à la structure de traits complexe et à la relation entre annotations, qui parfois peuvent relever de corpus différents qu'elles contribuent à aligner).

Une pratique courante en TAL consiste en effet à expliciter une interprétation sous la forme d'une séquence d'annotations. Une séquence d'étiquettes syntaxiques traduit ainsi une lecture syntaxique du discours qui est annoté, tandis qu'un étiquetage des entités nommées se contente de souligner les entités du domaine qui sont mentionnées dans le discours, ce qui en constitue une lecture sémantique assez fruste. Une étiquette particulière ne peut être considérée que relativement à la séquence à laquelle elle appartient [8] et à la perspective d'interprétation dans laquelle elle s'inscrit, c'est-à-dire à son auteur et à l'angle de lecture qui est le sien.

Il peut donc exister, pour un même discours, plusieurs annotations concurrentes, si plusieurs personnes ne l'interprètent pas de la même manière (une phrase ambigüe peut être lue de plusieurs façons), et plusieurs annotations complémentaires qui reflètent des niveaux de lectures différents.

En TAL, les annotations peuvent être posées soit manuellement par un interprète humain soit de manière automatique par un outil d'analyse. Dans le premier cas, l'interprétation peut refléter une part de la subjectivité de son auteur. Dans le second cas, l'interprétation est entièrement déterminée par les connaissances et l'algorithme incorporés dans l'outil d'analyse.

Nous nous intéressons ici à l'annotation manuelle en tant que tâche exécutée par des agents humains (les annotateurs).

8. Cette séquence peut éventuellement être très discontinue, comme dans le cas du renommage de noms de gènes (voir campagne 2, en section 5.2).

2.2.2 Termes associés

La différence entre *note* et *étiquette* est lié au fait que ce dernier terme est généralement utilisé pour désigner des annotations simples (traits atomiques ou couple de traits-valeurs) alors que nous ne souhaitons pas préjuger à ce stade de leur complexité (Ide et Suderman (2007) font par exemple référence à des structures de traits complexes).

Le principe de base d'une annotation consiste donc à apposer une *note* sur un *signal source*. L'annotation est alors la note en tant qu'elle est *ancrée* en un point ou en un *empan* du signal source (figure 2.6).

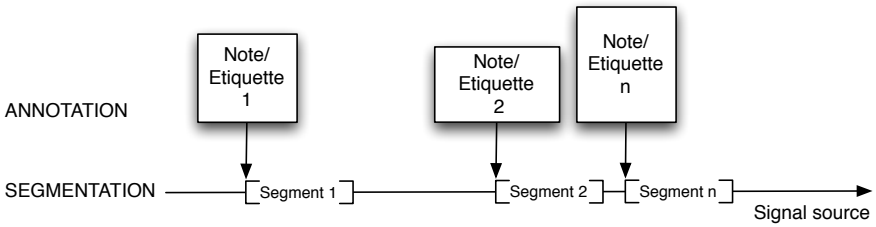


FIGURE 2.6: Ancrage des notes dans le signal source

2.2.3 Architecture générale d'un schéma d'annotation

L'architecture générale d'un schéma d'annotation se décompose, comme l'architecture d'un SGBD [9], en trois niveaux distincts qui assurent l'indépendance de la vue que l'utilisateur a des annotations et de leur encodage dans les fichiers qu'il manipule (Bird et Liberman, 2001) (figure 2.7) :

- au niveau le plus extérieur, le schéma externe est celui qui est visualisé et manipulé par l'utilisateur, généralement par le biais d'outils d'aide à l'annotation ;
- au niveau physique, différents schémas ont été proposés pour encoder et stocker les annotations. On distingue couramment, d'une part les annotations déportées et les annotations insérées dans le signal source qui est annoté, et d'autre part les balisages de type XML et les formats linéaires (voir section 3.1). C'est à ce niveau physique que se situent également les formats d'échange ;
- au niveau intermédiaire, le schéma logique (ou modèle de données) permet de spécifier et de contrôler le langage d'annotation : le formalisme ou structure logique détermine le pouvoir expressif des annotations (syntaxe du langage d'annotation) et la spécification du contenu indique le ou les jeu(x) d'étiquettes à utiliser (sémantique du langage d'annotation).

Les termes de *modèle, schéma, format* sont tour à tour employés pour désigner ce que nous appelons *structure logique* d'un schéma d'annotation, à la suite de (Bird et Liberman, 2001). Ide et Suderman (2007) parlent de *modèle* mais en un sens plus large.

9. L'analogie est reprise de (Bird et Liberman, 2001).

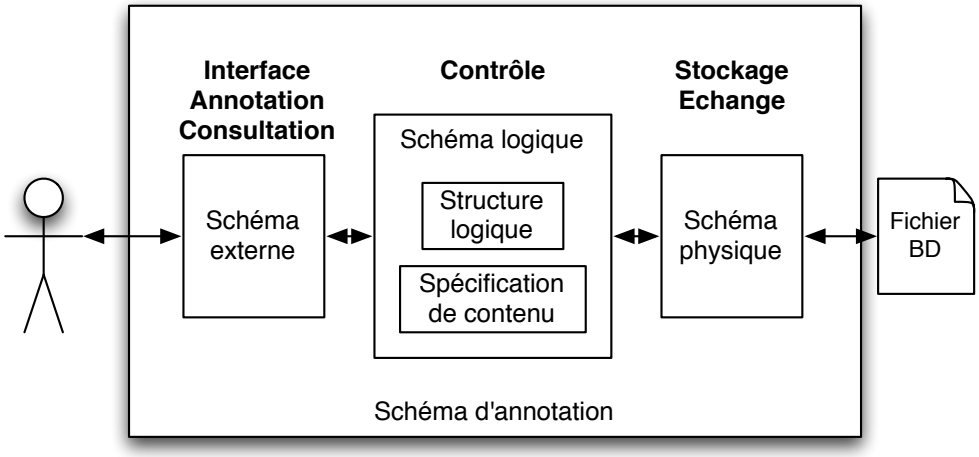


FIGURE 2.7: Architecture du schéma d'annotation inspirée de (Bird et Liberman, 2001)

Wittenburg (2000) propose le terme d'*architecture* mais nous préférons conserver ce dernier pour décrire l'organisation générale et tripartite du schéma d'annotation. Le format est pour nous la manière dont les annotations sont encodées dans le flux de données. Nous en présentons différents exemples dans la section 3.1.

2.2.4 Couches d'annotation

La structure logique du schéma d'annotation permet également d'organiser les annotations en couches :

- Les couches peuvent être définies comme des ensembles disjoints d'étiquettes apposables sur le texte indépendamment les unes des autres. C'est le cas des couches 1 et 2 dans le schéma de la figure 2.8.
- Les couches peuvent être définies comme la simple inclusion des segments de signal auxquels les annotations se rapportent, comme le proposent Bird et Liberman (2001) (cas des couches 2 et 4 dans le schéma de la figure 2.8).
- Enfin, on peut considérer que l'ensemble des annotations portant sur une couche d'annotation forme une couche distincte (cas des couches 1 et 3 dans le schéma de la figure 2.8).

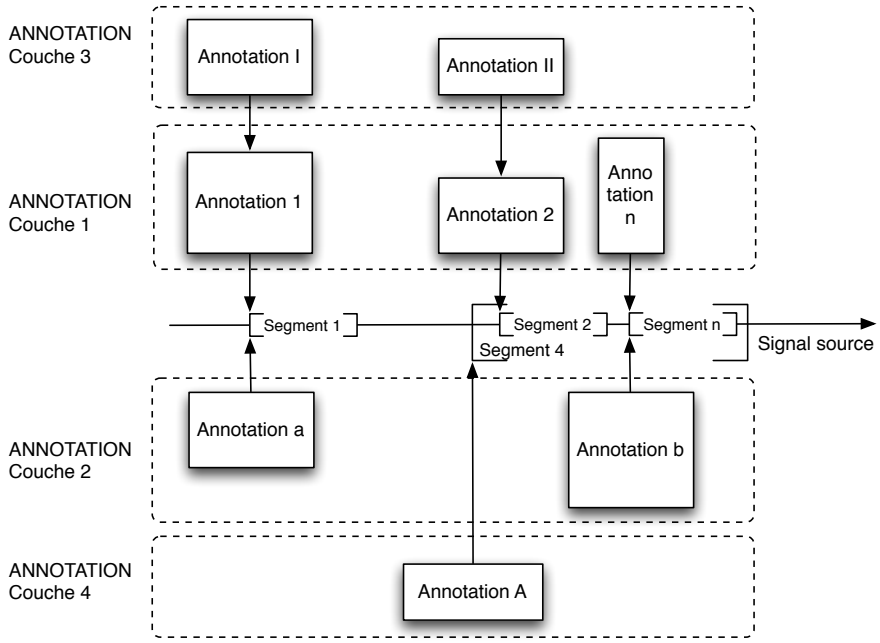


FIGURE 2.8: Structure logique et couches d'annotations

2.3 Typologie des campagnes d'annotation en fonction des usages

Les corpus annotés sont de plus en plus utilisés en TAL pour des usages variés. Nous les détaillons ici selon plusieurs angles de vue [10].

2.3.1 Acquisition *vs* évaluation

En TAL, les corpus annotés servent à entraîner des outils, qu'il s'agisse de systèmes par apprentissage supervisé ou à base de règles. Pour ne pas apprendre de modèles erronés, les systèmes à base de règles nécessitent une bonne cohérence de l'annotation manuelle. Il en va de même pour les systèmes par apprentissage supervisé (voir sous-section suivante).

Des corpus annotés sont également construits pour l'évaluation d'outils de TAL (y compris, dans ce cas, de systèmes par apprentissage non supervisé) et servent de référence par rapport à laquelle la sortie de ceux-ci est comparée. Ils doivent donc être cohérents et fiables par rapport au guide d'annotation, qui est utilisé comme documentation pour les systèmes évalués. Nous détaillons ce que signifie cette fiabilité

10. Cette réflexion a été menée en collaboration avec Adeline Nazarenko et Sophie Rosset (LIMSI-CNRS).

et les moyens à mettre en œuvre pour l'obtenir dans la section 3.3. Il est à noter que, dans le cas de l'évaluation, les corpus annotés peuvent être de taille plus réduite que pour l'apprentissage automatique.

Cela étant, les deux usages sont souvent combinés : une partie d'un corpus est utilisée pour l'entraînement et une autre pour l'évaluation de systèmes. C'est généralement le cas dans les campagnes d'évaluation, en particulier celles du programme Quæro (voir section 5.5 pour un exemple concernant les entités nommées).

2.3.2 Humain *vs* automatique

L'apprentissage réalisé à partir des corpus annotés peut concerner des systèmes automatiques comme des humains.

L'exemple de système automatique le plus connu est sans doute celui des étiqueteurs morpho-syntaxiques (comme, par exemple, le Brill Tagger (Brill, 1992) ou le TreeTagger (Schmid, 1997)), mais il existe aujourd'hui de très nombreuses applications de ce type, dans pratiquement tous les domaines du TAL, de l'extraction d'entités nommées (Nédellec *et al.*, 2006) à l'interprétation des noms composés (Tratz et Hovy, 2010), en passant par l'analyse syntaxique (Crabbé et Candito, 2008). La taille des corpus annotés doit dans ce cas être suffisante pour permettre l'apprentissage des systèmes (au moins plusieurs milliers d'annotations).

Par ailleurs, ces systèmes fonctionnant à partir de modèles numériques, on pourrait croire que les erreurs d'annotation sont en quelque sorte « lissées » par le nombre et que la qualité de l'annotation manuelle n'est pas fondamentale. Or, il a été démontré par Reidsma et Carletta (2008) que pour ce type d'utilisation, les erreurs d'annotation manuelle systématiques (dues, par exemple, à une incompréhension d'un point du guide), sont apprises par le système.

Les linguistes utilisent quant à eux les corpus annotés pour en extraire des informations, par exemple pour la création de nouvelles ressources lexicales (Kupsc, 2007). Dans ce cas, la qualité de l'annotation peut être moindre, l'humain intervenant dans le processus et pouvant corriger d'éventuelles erreurs. Les corpus servent également aux humains à tester des théories, des grammaires, des analyses. Les phases d'acquisition et de test sont alors mêlées dans le processus global de mise au point.

2.3.3 Application finale *vs* intermédiaire

Nous reprenons ici à notre compte une citation de Benoît Habert :

> « Quelle que soit sa richesse, une annotation est cependant toujours orientée par une tâche, même si cela est implicite » (Habert, 2000).

En effet, l'annotation d'un corpus peut différer en fonction de l'application visée. Nous en montrons des exemples dans (Fort *et al.*, 2009), dont nous reprenons la figure 2.9, qui montre l'impact de deux perspectives d'annotation (l'indexation et la modélisation du domaine) sur les résultats de l'annotation d'un petit exemple [11]. Dans cette figure, la première annotation est moins riche que la seconde, qui considère plus de types sémantiques (`taxon`) avec une granularité plus fine (`EuKVirus` est un sous-type de `gene`), ce qui introduit des enchâssements de balises. Par ailleurs, toutes les mentions sont annotées dans la seconde annotation alors que seuls les assimilés noms propres le sont dans la première.

ANNOTATION D'INDEXATION ; types `gene` et `protein`
We conclude that <gene>3CDproM</gene> can process both structural and nonstructural precursors of the <EukVirus uncertainty-type="too-generic">poliovirus polyprotein</EukVirus> and that it is active against a synthetic peptide substrate.

ANNOTATION DE MODÉLISATION ; types `taxon`, `gene` et `protein`
We conclude that <EukVirus>3CDproM</EukVirus> can process both structural and nonstructural precursors of the <EukVirus uncertainty-type="too-generic"><taxon>poliovirus</taxon> polyprotein</EukVirus> and that <EukVirus>it</EukVirus> is active against a synthetic peptide substrate.

FIGURE 2.9: Exemples d'annotations en biologie

Il est donc indispensable de garder cette application en vue lors de l'annotation et de la documenter clairement dans le guide d'annotation. Geoffrey Leech insiste d'ailleurs sur ce point :

> « les annotations sont d'autant plus utiles qu'elles ont été conçues pour une application particulière » [12] (Leech, 2005)

Cette application peut être finale, avec une visée applicative directe, comme le résumé de matchs pour notre campagne d'annotation de matchs de football (présentée en section 5.3), ou intermédiaire (ou interne au TAL), comme dans le cas de l'étiquetage morpho-syntaxique, qui ne représente en général qu'une étape dans une application finale.

2.3.4 Conclusion

Nous venons de montrer que les données annotées manuellement ont des usages très variés dans le domaine du TAL.

Par ailleurs, si la quantité de données annotées nécessaire peut varier en fonction du type d'application considéré, la qualité de l'annotation produite reste un élément fondamental pour la grande majorité des usages qui en sont faits.

11. http://www.ncbi.nlm.nih.gov/pubmed/1331532
12. *« [T]he annotations are more useful, the more they are designed to be specific to a particular application »*

2.4 Diversité des types d'annotations

Outre une diversité d'usage, nous constatons une importante diversité de types dans les annotations réalisées. Cette diversité s'exprime d'abord à travers l'expressivité des langages d'annotation utilisés, mais également par la complexification des annotations réalisées.

2.4.1 Expressivité des langages d'annotation

Le langage d'annotation est le vocabulaire utilisé pour annoter le flux de données. Dans la grande majorité des cas d'annotation pour le TAL, ce langage est contraint. Il peut être de différents types.

Le plus simple est le type booléen, qui couvre les cas d'annotation ne nécessitant qu'une seule catégorie. Un segment est alors annoté, avec cette catégorie (qui peut n'être qu'implicite), ou non annoté. Les expériences menées par Laignelet et Rioult (2009) sur l'annotation de segments d'obsolescence utilisent par exemple ce type de langage d'annotation.

Viennent ensuite les langages du premier ordre. Les langages de types sont par exemple utilisés pour l'annotation morpho-syntaxique sans traits (parties du discours) ou avec traits (morpho-syntaxe). Le premier cas est rare, car même si le jeu d'étiquettes paraît peu structuré, comme dans le cas du *Penn Treebank* (Santorini, 1990), on peut presque toujours en déduire des traits (par exemple, *NNP*, nom propre singulier et *NNPS*, nom propre pluriel, pourraient être traduits en *NNP+Sg* et *NNP+Pl*).

En ce qui concerne les relations, on en annote aujourd'hui une grande variété dans le domaine du TAL, des relations binaires orientées (par exemple, le renommage de gènes, présenté en section 5.2) aux relations n-aires non orientées (par exemple, les chaînes de co-référence (Poesio et Artstein, 2005)).

Enfin, des langages du second ordre pourraient être utilisés, par exemple, pour annoter des relations sur des relations. Ainsi, *interception(passe(j1, j2), j3)* représente une relation de passe entre deux joueurs, qui est interceptée. En pratique, on cherche à simplifier l'annotation et on se ramène à un langage du premier ordre en réifiant la première relation. Nous ne connaissons aucun exemple d'une annotation utilisant un langage du second ordre.

2.4.2 Complexification des annotations

Dans son état de la technologie d'annotation automatique, Véronis (2000) conclut sur un schéma synthétisant la situation en 2000, tout en notant la progression rapide des

techniques. Sur ce schéma, seuls l'annotation en parties du discours et l'alignement multilingue de phrases sont jugées « opérationnels ». Une majorité est à l'état de prototype (annotation de prosodie, syntaxe partielle, alignement multilingue de mots) et le reste ne donne pas encore lieu à des « implémentations utilisables en situation d'annotation réelle » (syntaxe pleine, sémantique du discours) ou est proche du prototype (transcription phonétique, sémantique des mots).

Le domaine s'est aujourd'hui largement développé et la période actuelle voit une complexification des annotations réalisées, que ce soit du point de vue des médias traités ou de celui des phénomènes annotés.

On voit ainsi se développer depuis quelques années les annotations de sources vidéo, notamment de langues des signes, annotations qui ont d'ailleurs fait l'objet d'un atelier dans le cadre de la conférence TALN en 2011 et 2012, DEGELS (Défi d'annotation de Gestes et de Langue des Signes) [13]. Une formation à l'annotation de corpus vidéo a également été organisée par l'ATALA (Association pour le Traitement Automatique des Langues [14]) en 2011 [15].

Par ailleurs, des annotations sémantiques de plus en plus complexes sont réalisées. On peut citer, par exemple, l'annotation d'opinions (Paroubek *et al.*, 2010), un sujet en ce moment très prisé, ou encore les annotations de noms de gènes et de protéines (Kim *et al.*, 2003), qui cèdent maintenant la place à des annotations plus complexes de relations comme le renommage de gènes (voir notre campagne 2, section 5.2) ou de relations entre entités, et qui donnent lieu à des *shared tasks* BioNLP [16]. Des annotations sémantiques sont également réalisées par rapport à un modèle formel (ontologie) (Cimiano et Handschuh, 2003).

En outre, des campagnes d'annotation sur des sujets devenus classiques (entités nommées, résolution d'anaphore) font aujourd'hui l'objet d'une complexification notable (voir notre campagne 5 d'annotation d'entités nommées structurées, section 5.5).

2.4.3 Conclusion

Il existe encore peu de corpus annotés (et disponibles) en différents niveaux, par exemple selon différentes théories. MASC (*Manually Annotated Sub-Corpus*) (Ide *et al.*, 2008) représente de ce point de vue une avancée intéressante, puisqu'il inclut, entre autres, des annotations de « frame » à la FrameNet (Baker *et al.*, 1998) et de « senses » WordNet (Fellbaum, 1998). Dans le même ordre d'idées, il n'existe pas encore, à notre connaissance, de corpus annoté en multimédia (avec alignement de chaque niveau

13. http://degels.limsi.fr/
14. http://www.atala.org/
15. http://tals.limsi.fr/jatala2011.html
16. http://sites.google.com/site/bionlpst/home

d'annotation sur le signal source) pour le français. Un tel corpus serait pourtant extrêmement utile à l'exploration des questions d'annotation du multimédia.

Enfin, si, comme nous allons le voir dans la section suivante, les formats des annotations ont largement évolué au fil des ans pour s'adapter à cette complexité grandissante, celle-ci n'est pas encore prise en compte dans la méthodologie et la préparation d'une campagne d'annotation.

Chapitre 3

Outils et techniques pour l'annotation

3.1 Formats d'annotation

Si le processus d'annotation manuel lui-même est encore peu étudié, de nombreuses propositions ont été faites concernant la manière d'« encoder » les annotations. Nous présentons ici les grandes tendances de ces dernières années et détaillons quelques normes parmi les plus utilisées.

3.1.1 Annotation insérée dans le signal source *vs* annotation déportée

La manière la plus immédiate d'annoter un signal source consiste à y insérer directement les notes. Ce principe, utilisé tout naturellement depuis les débuts de l'annotation, a été remis en cause dès 1993 par Geoffrey Leech (Leech, 1993) dans les deux premières de ses sept maximes pour l'annotation de corpus :

1. « Il doit toujours être possible de revenir aux données d'origine » [1] : le corpus d'origine, une fois annoté (et souvent modifié), est parfois difficile à retrouver dans son état initial.
2. « Les annotations doivent pouvoir être extraites du texte » [2] : il n'est pas toujours trivial d'extraire les annotations et les empans annotés du texte (par exemple, le marqueur d'annotation peut être ambigu et difficile à retrouver à coup sûr).

Ces principes ont par la suite été repris, notamment par Ide et Romary (2006), pour qui l'annotation déportée (ou *stand-off*), c'est-à-dire présentée à part, souvent dans un ou des fichiers séparés, doit devenir une norme.

L'annotation déportée présente en effet de nombreux avantages. Elle conserve le corpus d'origine en l'état, ce qui assure le respect des droits existant sur le corpus (qui peut être en lecture seule) et ne « pollue » pas celui-ci. Elle offre également une certaine souplesse dans l'annotation : d'autres niveaux d'annotation peuvent être ajoutés, sans perturbation des existants. Enfin, chaque niveau d'annotation peut être manipulé séparément, y compris dans des fichiers distincts, ce qui permet, entre autres, de comparer

1. *« It should always be possible to come back to initial data. »*
2. *« Annotations should be extractable from the text. »*

des annotations plus facilement. Ce type d'annotation permet surtout d'annoter des groupes discontinus, des chevauchements et des recouvrements, ainsi que des relations entre différentes couches de l'annotation.

Cependant, l'annotation déportée présente également des inconvénients. Le premier d'entre eux est que la taille de l'annotation y est plus importante, car il faut ajouter des éléments permettant d'identifier l'empan considéré (position de début et de fin, par exemple). Or, la taille de l'annotation peut faire grossir la taille d'un corpus de manière considérable, et si cette annotation déportée est réalisée dans le fichier même du corpus, cela peut avoir des conséquences. Ainsi, il a été montré dans (Nazarenko *et al.*, 2006) que sur 55 000 pages Web, soit 80 millions de mots, une segmentation en phrases, en tokens, puis une annotation morpho-syntaxique et une annotation des termes et entités nommées augmente la taille du corpus d'un facteur 16.

En outre, annoter de manière déportée implique l'assistance d'un outil, qui calcule les positions des empans annotés et facilite le travail de l'annotateur, comme `Glozz` par exemple.

Enfin, l'annotation déportée ne permet généralement pas de modifier le corpus d'origine (c'est d'ailleurs l'un de ses avantages). Toutefois, celui-ci comprend parfois des erreurs qu'il faut pouvoir corriger pour annoter correctement. Ces corrections peuvent bien entendu être réalisées au moyen de traits spécifiques dans l'annotation, mais cela alourdit considérablement l'annotation. Ainsi, dans notre deuxième campagne d'annotation d'entités nommées structurées (voir section 5.5), le corpus provenait d'une numérisation et comprenait donc des erreurs, notamment de segmentation en mots, qu'il était important de pouvoir corriger avant d'annoter et nous avons dû ajouter une balise de correction spécifique (Rosset *et al.*, 2012).

3.1.2 Formats linéaires *vs* formats balisés

Une manière simple d'encoder des annotations consiste à les ajouter directement dans le texte, associées à un séparateur simple, bien défini. Ce format linéaire a été utilisé par le passé dans de très nombreux projets, notamment le *Brown University Standard Corpus of Present-Day American English*, dit *Brown corpus* (Kucera et Francis, 1967) ou le *Penn Treebank* (Marcus *et al.*, 1993).

> Exemple extrait du *Brown corpus* :
> ('The', 'AT'), ('Fulton', 'NP-TL'), ('County', 'NN-TL'), ('Grand', 'JJ-TL'), ('Jury', 'NN-TL'), ('said', 'VBD')

> Exemple extrait de la partie morpho-syntaxique du *Penn Treebank* :
> It/PRP settled/VBD with/IN a/DT loss/NN of/IN 4.95/CD cents/NNS at/IN / 1.3210/CD a/DT pound/NN ./.

Bien plus tard, le projet CHILDES (MacWhinney, 2000) a également utilisé un format linéaire, le format CHAT, qui encode des transcriptions d'interactions conversationnelles. Ce format est aujourd'hui repris en version XML.

> Exemple extrait du corpus CHILDES :
> PAT : <boy [*] no>[//] girl [/] girl truck # girl +...

Ce type de format a également été utilisé dans le projet IST-AMITIÉS de systèmes de dialogues multilingues pour le routage d'appels (Hardy *et al.*, 2002) qui a débuté en 2001 et s'est terminé en 2005 pour sa partie annotation thématique. Dans le fichier d'annotation thématique, la ligne :

> #15h ;INFO_COURRIER_RECEPTION ;alors j'ai reçu une lettre euh mentionnant de payer 18 euros 95 sous 4 jours sinon etc etc ;C ;1 ;14

indique que pour le dialogue 15h, le thème INFO_COURRIER_RECEPTION est instancié par le client (C) au quatorzième tour de parole (indiqué sur la ligne) et représente le premier thème de la taxinomie. Les annotations sémantiques et dialogiques sont elles représentées en XML.

Chaque projet a donc choisi une manière différente d'encoder les informations nécessaires, plus ou moins complexe en fonction des annotations prévues : forme parenthésée pour le *Brown corpus*, ajout de barre oblique avant l'étiquette pour le *Penn Treebank*, formats complexes pour CHILDES et pour AMITIÉS.

Une fois clairement définis dans la documentation, ces formats linéaires sont simples à comprendre et à utiliser et sont très économes en espace. Cette simplicité s'accompagne cependant d'une expressivité limitée : une interprétation est nécessaire pour en comprendre la signification, en particulier lorsque l'annotation devient complexe, comme c'est le cas pour CHILDES et pour AMITIÉS.

On voit donc se généraliser, depuis une dizaine d'années, l'utilisation de formats plus structurés, même pour des projets dans lesquels l'annotation produite reste simple, comme par exemple dans l'identification de segments d'obsolescence (Laignelet et Rioult, 2009).

L'accès facilité aux ressources a permis de développer l'utilisation de corpus électroniques. Cependant, l'accès à ces ressources par des programmes informatiques nécessite une structuration accrue des données (Wittenburg, 2000). Cette structuration facilite également l'échange et la ré-utilisation des corpus annotés, si coûteux à produire. En outre, les langages de balises permettent une cohérence accrue, et dans le cas de XML, vérifiable par une validation automatique. Enfin, de nombreux outils de visualisation et de manipulation XML sont aujourd'hui disponibles.

3.1.3 Normes *de jure vs* normes *de facto*

Nous présentons ici les normes les plus courantes pour l'annotation en général[3], sachant que beaucoup de projets actuels (en particulier de projets de petite ampleur) utilisent le langage de balise XML (eXtensible Markup Language) et créent leur propre grammaire (dtd ou schéma) pour décrire leurs annotations. Nous distinguons les normes *de jure*, que tentent d'imposer des organisations comme ISO, des normes *de facto* (ou standards), qui proviennent de la pratique d'un groupe d'utilisateurs, par exemple la TEI.

La TEI (*Text Encoding Initiative*)

L'initiative pour l'encodage de textes (TEI) s'est mise en place dès 1987, à l'initiative de plusieurs associations savantes liées aux sciences humaines et à la linguistique informatique (*Association for Computers and the Humanities*, *Association for Computational Linguistics*, *Association for Literary and Linguistic Computing*), afin de développer une norme de représentation de textes sous forme électronique. Depuis ce noyau, un consortium s'est créé en 2000 pour maintenir et développer la norme TEI.

La TEI propose non seulement un format normalisé pour l'échange de données, mais surtout des guides très complets permettant d'utiliser ce format. Celui-ci se veut indépendant de toute application et doit permettre l'encodage de n'importe quel type de texte ou d'informations (annotations). De ce fait, il est d'une grande richesse et donc d'une grande complexité. Il est en outre aujourd'hui possible d'étendre la TEI pour ses besoins propres, ajoutant encore à la modularité du format, mais parfois aussi à sa complexité. Celui-ci propose heureusement une distinction entre *required practices* (pratiques obligatoires), *recommended practices* (pratiques recommandées) et *optional practices* (pratiques optionnelles). Une très large documentation est également disponible sur le site de la TEI[4], ainsi que des outils permettant, par exemple, de faciliter la création du modèle (outil `Roma`) ou l'annotation elle-même (outils `SIG`).

Du fait de la complexité du format TEI, le consortium a d'emblée opté pour un langage à balises, SGML (*Standard Generalized Markup Language*). Il est logiquement passé à son successeur plus contraint, XML (*eXtensible Markup Language*), en 2002. TEI est un format non déporté : les annotations sont ajoutées directement au corpus d'origine, dans le même fichier.

Le format TEI présente enfin une caractéristique qui ajoute encore à sa complexité : il est assez peu contraignant et permet d'annoter un même phénomène de plusieurs

3. Nous ne présenterons donc pas les normes telles que MAF (*Morpho-syntactic Annotation Framework*, ISO 24611), qui traitent d'un type d'annotation en particulier (en l'occurrence, l'annotation morpho-syntaxique).

4. http://www.tei-c.org/Guidelines/

manières. Ainsi, la balise <p> est une et il est possible d'utiliser cette dernière à la place d'une balise <p> , ce qui n'est pas sans poser problème pour l'interprétation des annotations.

Un exemple de texte encodé en TEI est donné ci-dessous :

```
<div1 type="storylist" org="composite">
 <head>News in brief</head>
 <div2 type="story">
  <head>Police deny <soCalled>losing</soCalled> bomb</head>
  <p>Scotland Yard yesterday denied claims in the Sunday Express
  that anti-terrorist officers trailing an IRA van loaded with
  explosives in north London had lost track of it 10 days ago.</p>
 </div2>
```

Malgré une phase d'apprentissage importante, le format TEI a su séduire nombre de chercheurs, en particulier en sciences humaines, et il est aujourd'hui très utilisé pour l'analyse littéraire ou historique de textes, par exemple (voir la longue liste des projets utilisant le format TEI sur le site de la TEI [5]).

(X)CES (*Corpus Encoding Standard)*

La norme pour l'encodage de corpus (*Corpus Encoding Standard*) a été créé par EAGLES/ISLE afin de compenser certains défauts du format TEI. En particulier, CES étend le format TEI pour interdire l'utilisation de plusieurs représentations pour un même type d'annotation. CES présente en outre la caractéristique d'être un format de type déporté : les annotations sont identifiées par leur position dans le corpus d'origine et stockées dans un fichier séparé.

Cependant, la spécification précise des catégories linguistiques est laissée à des projets tels qu'EAGLES/ISLE (dont CES faisait partie).

GrAF (*Graph Annotation Format)*

GrAF (Ide et Suderman, 2007) est une extension du *Linguistic Annotation Framework* (LAF) (Ide et Romary, 2006), développé dans le cadre d'ISO TC37 SC4 et qui a donné naissance en 2012 à la norme ISO 24612. Le format GrAF implémente donc les dernières « bonnes pratiques » en la matière :
– séparation entre les données (en lecture seule) et les annotations (annotation déportée),

5. http://www.tei-c.org/Activities/Projects/

– séparation entre le format d'annotation utilisateur et le format d'échange (avec passerelles),
– séparation entre la structure et le contenu dans le format d'échange.

GrAF est une sérialisation XML (en TEI) de la structure de graphe générique décrite dans LAF. Il permet d'encoder tous les types d'annotations linguistiques et fournit les moyens de représenter des informations éventuellement complexes.

3.1.4 Conclusion

L'évolution du domaine vers le multimédia et plus généralement vers une complexité grandissante a poussé à la définition de jeux d'étiquettes de plus en plus riches, mais aussi de plus en plus complexes et lourds à implémenter. Cette évolution s'est traduite dans les formats physiques proposés, eux aussi de plus en plus riches. Un important et salutaire effort de normalisation des formats a été mené, en particulier par Nancy Ide et Laurent Romary (Ide *et al.*, 2003; Ide et Romary, 2006), dans le cadre de l'ISO, dans le but d'améliorer les capacités d'échange et de ré-utilisation de corpus annotés.

Cela étant, la communauté semble encore réticente à utiliser ces normes, tant il est vrai que leur utilisation nécessite une phase d'apprentissage qui, malgré les nombreuses facilités proposées (outils, passerelles existantes, etc), peut être décourageante, en particulier dans le cadre de petits projets d'annotation. Cet état de fait est apparu clairement lors d'un débat sur le sujet organisé par Nancy Ide lors du troisième *Linguistic Annotation Workshop*, en 2009. Les formats simples (XML avec dtd « maison ») restent aujourd'hui encore très utilisés.

Il faut ajouter à cela l'influence prépondérante des outils d'aide à l'annotation : les gestionnaires des projets d'annotation ont logiquement tendance à rechercher en priorité un outil correspondant à la tâche prévue, sans se soucier du format physique utilisé. Le pragmatisme prend alors le pas sur les « bonnes pratiques ». Heureusement, ces outils d'aide à l'annotation évoluent eux aussi, en parallèle des formats, et les intègrent progressivement. Ainsi, il existe un *plugin* ANC de l'outil d'aide à l'annotation GATE (voir Annexe A.3.1) qui permet d'enregistrer les annotations au format GrAF. Par ailleurs, dans le cadre de MASC (*Manually Annotated Sub-Corpus*) (Ide *et al.*, 2008), les annotations des contributeurs sont distribuées dans leur format original et dans le format GrAF, et c'est le projet ANC (*American National Corpus*) lui-même qui assure la traduction vers GrAF [6].

6. Voir : http://www.anc.org/MASC/Contribute_Data_and_Annotations.html

3.2 Outils d'aide à l'annotation

Les outils d'aide à l'annotation permettent de grandement faciliter l'annotation manuelle de corpus, en particulier depuis l'apparition des langages de balises (voir section 3.1). Ils permettent en premier lieu d'éviter les fautes de frappe, sources de nombreuses erreurs dans les annotations manuelles réalisées sans outil (voir par exemple (Fort et Sagot, 2010)). Leur apport va cependant bien au-delà.

S'il existe pléthore d'articles présentant des outils d'annotation, très peu en font une analyse de haut niveau. A notre connaissance, seuls (Dipper *et al.*, 2004), (Reidsma *et al.*, 2004) et (Burghardt, 2012) proposent une comparaison poussée d'outils permettant leur évaluation. Ils ne considèrent cependant qu'un nombre limité d'outils (cinq pour (Dipper *et al.*, 2004), deux pour (Reidsma *et al.*, 2004) et trois pour (Burghardt, 2012)) et l'analyse proposée par les deux premiers est focalisée sur une tâche d'annotation particulière (annotation purement linguistique pour le premier et de vidéo pour le deuxième). Si cet état de l'art reprend certaines des informations que contiennent ces articles, il est plus majoritairement issu de notre expérience et de notre analyse des outils existants (dont un inventaire est présenté en annexe A).

3.2.1 Définition

Un outil d'aide à l'annotation, ou outil d'annotation [7], est une interface facilitant l'annotation manuelle de données. Il existe des outils permettant l'annotation manuelle de corpus non textuels, comme par exemple de vidéo (`Anvil` [8] ou `Advene` [9]), de parole (`Praat` [10]) ou encore de musique (`wavesurfer` [11]), mais un tel inventaire nous emmènerait trop loin et nous nous restreindrons ici à l'annotation de flux textuels (éventuellement, des transcriptions). Nous ne traiterons pas non plus des outils d'annotation de pages Web, dont les fonctionnalités sont très proches de celles des outils d'annotation présentés ici, sans en avoir la richesse. Enfin, si nous ne considérons pas non plus ici les éditeurs XML ou les éditeurs de texte traditionnels, il nous faut les distinguer des outils d'annotation.

Certaines campagnes d'annotation ont en effet été réalisées à l'aide d'éditeurs XML, par exemple `Morphon`, utilisé par T. Lebarbé pour l'annotation des manuscrits de Stendhal (Lebarbé, 2008) ou `Epic` d'Arbortext, qui nous a été imposé par le client dans le cadre de notre campagne d'annotation en pharmacologie (voir section 5.4). Un éditeur XML a pour but d'éditer des fichiers au format XML, pas d'annoter des

7. Nous utilisons dans ce document le terme simplifié d'« outil d'annotation » pour désigner les outils d'aide à l'annotation.
8. http://www.anvil-software.de/
9. http://liris.cnrs.fr/advene/
10. http://www.fon.hum.uva.nl/praat/
11. http://www.speech.kth.se/wavesurfer/

fichiers. Même si les fonctionnalités sont proches, la logique sous-jacente n'est pas la même et un éditeur XML sert avant tout à « modifier » (éditer) un fichier XML, pas à ajouter des annotations sur un corpus de textes. La première différence tient à la notion de corpus, qui n'existe pas dans les éditeurs XML, ce qui ne permet pas d'avoir une vision globale de l'avancement de la campagne d'annotation et ne permet donc aucune gestion de celle-ci. En outre, les éditeurs XML ne permettent généralement pas l'annotation déportée (voir sous-section 3.1.1) et interdisent donc l'annotation simple de recouvrements, de groupes discontinus et de relations (dont la visualisation n'est jamais prévue). Enfin, certains outils d'annotation (entre autres `Knowtator`, `Glozz`, `Slate` ou `GATE`) permettent de visualiser les désaccords entre annotateurs concurrents et/ou de calculer l'accord inter-annotateur, ce qui n'est jamais le cas dans un éditeur XML.

Les éditeurs de texte présentent les mêmes limites. Il faut y ajouter la non validation du XML et donc de possibles erreurs de chevauchement de balises (voir section 5.5). Cela étant, des outils simples peuvent se révéler très utiles dans le cadre d'expériences limitées (prototypage, par exemple) ou complétés par d'autres outils de normalisation (voir encore section 5.5).

Étant donné le foisonnement des campagnes d'annotation, il est impossible de connaître tous les outils d'aide à l'annotation utilisés aujourd'hui. Nous présentons néanmoins en annexe A une liste que nous pensons relativement complète des outils d'annotation récents et nous en détaillons les caractéristiques principales, ainsi que la logique sous-jacente correspondant au contexte de création. Cette liste comprend des outils plus ou moins disponibles, plus ou moins maintenus, open-source ou non. Cependant, de notre point de vue et à l'image de ce qui a été observé pour les corpus (Cohen *et al.*, 2005), un outil d'annotation, pour être pérenne (donc « rentable » en terme d'investissement) doit être open-source, gratuit, maintenu et bien documenté. On attend également de ce type d'outil qu'il soit facile à installer et à utiliser. Les critères ergonomiques sont en effet importants : on sait que les fonctionnalités difficiles d'accès ne sont pas utilisées. Dans le cas de notre expérience en microbiologie (voir section 5.1), les annotateurs ont souvent négligé de mentionner leur incertitude parce qu'ajouter l'attribut correspondant à leurs annotations était malaisé.

3.2.2 Des fonctionnalités qui se généralisent

Nous observons que les outils d'annotation, s'ils sont généralement spécifiés autour d'une ou plusieurs tâches d'annotation plutôt qu'autour des besoins des annotateurs, prennent de mieux en mieux en compte ceux-ci, à travers des interfaces de plus en plus conviviales et efficaces. Les formalismes utilisés sont quant à eux en voie de normalisation, avec la généralisation de XML et de l'annotation déportée.

Des interfaces de plus en plus efficaces

Les interfaces des outils d'annotation deviennent ainsi de plus en plus faciles d'utilisation, même s'il reste des progrès à faire (pas de *undo*[12] dans `GATE`, raccourcis clavier en cours d'ajout dans `Glozz`, multi-fenêtrage lourd à gérer dans `MMAX2`, etc.). Elles proposent par exemple des possibilités d'édition qui permettent à l'annotateur d'automatiser certains traitements (fonctionnalité *annotate all*[13] dans `GATE`, `Glozz`, `Djangology`, expressions régulières générées automatiquement et révisables par l'annotateur dans `SYNC3`, annotation rapide en une seule sélection dans `Knowtator`). Elles offrent souvent une fonctionnalité de masquage de certaines annotations (par niveau ou selon d'autres critères, comme dans `Glozz`) pour alléger la visualisation, et presque toutes permettent la personnalisation des couleurs associées aux catégories à annoter. Enfin, la recherche et la correction des annotations sont parfois facilitées par l'existence de moteurs de recherche puissants, qui accèdent à la fois aux annotations et au texte (langage GlozzQL pour `Glozz`, ou moteur de recherche d'`EasyRef`, recherche par expressions régulières dans `Dexter` et `UAM CorpusTool`). Une fois les annotateurs formés à la tâche et familiarisés avec l'outil, ces fonctionnalités de l'interface permettent d'améliorer la vitesse et le confort d'annotation (Dandapat *et al.*, 2009).

Même si cela ne semble pas une préoccupation également importante pour tous les créateurs d'outils d'aide à l'annotation et si le LDC est évidemment plus sensible à la question que l'INRA (`Cadixe`), la grande majorité des interfaces sont écrites en Java et permettent donc une prise en compte aisée de nombreuses langues[14].

Certaines fonctionnalités, bien qu'encore peu répandues, nous semblent d'un grand intérêt. Ainsi, `brat` associe à chaque annotation une adresse URL unique, ce qui permet d'y faire référence précisément dans la documentation. Quant à `Glozz`, il propose une vue ruban de tout le texte qui fournit une vue d'ensemble des annotations et du texte très utile pour de l'annotation de niveau macro, comme la structure du discours.

Certaines fonctionnalités spécifiques manquent cependant encore à l'appel. Ainsi, nous avons constaté dans le cadre de l'annotation de presse ancienne que le texte numérisé étant souvent erroné, les annotateurs avaient besoin de voir l'image d'origine pour annoter (voir section 5.6). Dans ce type d'annotation, il serait donc utile de pouvoir inclure un accès aux images sources dans l'outil. Par ailleurs, si l'ajout de commentaires ou d'incertitudes, si possible typées, est possible dans `Glozz`, par exemple, il faut prévoir ceux-ci dans le modèle de données, car ces métacatégories ne sont pas proposées par défaut alors qu'elles nous semblent indispensables. Enfin, certains outils comme `Glozz` souffrent encore d'un manque de robustesse qui les rend inadaptés pour l'annotation de fichiers de grande taille.

12. « Annuler » l'action précédente.
13. « Annoter tout », par analogie avec « remplacer tout ».
14. En interne, Java utilise Unicode et stocke les caractères encodés en UTF-16.

Des formalismes en voie de normalisation

Nous constatons également la généralisation de XML comme format d'export, voire de stockage, des annotations, couplé, pour la plupart des outils, à l'utilisation de l'annotation déportée. Si celle-ci est considérée par de nombreux auteurs comme devant faire partie des bonnes pratiques d'annotation (Leech, 1997; Ide *et al.*, 2003), certains gestionnaires de campagnes préfèrent laisser l'accès aux données sources aux annotateurs. Cela a été le cas pour la campagne d'annotation en entités nommées structurées à laquelle nous avons participé (voir section 5.5). Si ce choix présente l'avantage de permettre des modifications dans le fichier source, il peut aussi poser problème, par exemple lorsqu'un annotateur vient modifier les balises insérées de manière non conforme et rend ainsi le XML non valide. `GATE` propose en option la possibilité de modifier les données sources, le gestionnaire de la campagne pouvant ainsi décider ou non de permettre de telles modifications. Les annotations déportées permettent d'annoter des groupes de mots discontinus, des factorisations, des recouvrements, des relations, orientées ou non. `Cadixe`, qui manipule du XML mais ajoute directement les annotations dans le source et non en déporté, est de ce fait très limité. Dans une expérience que nous avons menée en microbiologie avec `Cadixe` (voir section 5.1), le fait de ne pas pouvoir annoter les coordinations d'entités nommées a posé problème.

L'annotation de relations, orientées ou non, ou de chaînes (dans le cas de l'anaphore), est une fonctionnalité de plus en plus répandue, même si `Callisto` et `GATE` sont limités de ce point de vue, et que `Cadixe`, `Dexter`, `UAM CorpusTool` et `Eulia` ne le permettent pas.

Certains outils permettent en outre la définition et l'utilisation de différentes « couches » d'annotation (en particulier `MMAX2`, `Glozz`, `UAM CorpusTool`), qui correspondent à des niveaux linguistiques (parties du discours, syntaxe, etc) ou à des « groupes » définis par le créateur du modèle de données, comme dans `Glozz`. La souplesse de la définition autorise le regroupement d'éléments sémantiquement proches (voir notre campagne d'annotation de matchs de football en section 5.3) sans qu'ils aient de signification linguistique particulière. Ces « groupes » sont utilisés pour l'annotation (un groupe étant entièrement annoté avant un autre), l'affichage/masquage des annotations et l'évaluation.

3.2.3 Des tendances fortes

Nous observons que les outils d'annotation arrivent peu à peu à maturité et évoluent dans trois grandes directions : la généricité, la collaboration et la gestion de l'annotation.

Évolution vers une plus grande souplesse et généricité

Nous constatons ces dernières années une évolution depuis des outils d'annotation spécifiques à une tâche particulière (voir Annexe A.2) vers des outils qui se veulent génériques et souples, souvent *via* des plug-ins (voir Annexe A.1) ou une API commune (comme pour les outils du LDC, présentés en annexe dans la sous-section A.2.1).

Cette généricité, quand elle est le résultat d'une évolution d'un outil dédié à une tâche spécifique ou un objectif différent, pose cependant un problème de complexité de l'outil résultant, qui en devient souvent difficile à installer et à paramétrer. C'est notamment le cas de `GATE` ou de `Callisto`, dont la prise en main, nous avons pu le constater lors de nos tests, est complexe pour l'administrateur. Les outils directement créés dans une optique générique sont souvent mieux conçus de ce point de vue et plus faciles à paramétrer. C'est en particulier le cas pour `Glozz` ou pour les outils les plus récents, comme `CCASH`.

Enfin, la généralisation de XML permet plus facilement de s'adapter à des normes comme par exemple le format TEI, ce qui est un atout pour qui veut pouvoir facilement partager ses données.

Évolution vers l'annotation dite « collaborative »

Petasis (2012) utilise l'adjectif collaboratif/distribué (*collaborative/distributed*) pour distinguer les outils d'annotation collaborative sous forme d'application Web, de celui dont il parle (qui n'est pas un client léger). Ce faisant, il tente de détricoter un amalgame entre deux termes qui sont souvent confondus aujourd'hui. Le terme d'annotation collaborative est en effet devenu très ambigu, signifiant pour les uns l'annotation par *crowdsourcing*, pour les autres l'annotation par une communauté d'experts, voir les deux (l'appel à soumissions pour le VIe Linguistic Annotation Workshop en est un bon exemple [15]), et pour d'autres encore, dont Petasis (2012), la participation à un projet commun d'annotation.

Selon nous, la collaboration dans l'annotation se définit selon deux axes, sa visibilité pour les annotateurs et les éventuels moyens utilisés pour sa mise en œuvre, la collaboration pouvant en effet être directe ou indirecte.

Ainsi, le mode d'annotation à la Wikipedia, dans lequel chaque annotateur voit ce que font les autres et peut modifier leurs annotations, mode également proposé dans `brat` et dans `SYNC3` en option, est éminemment collaboratif, car l'annotation est dans ce cas à la fois directe et visible pour les annotateurs.

15. *« The special theme for LAW VI is Collaborative Annotation (both community-based and crowdsourced) »* : http://faculty.washington.edu/fxia/LAWVI/cfp.html

Á l'opposé, l'adjudication par un expert d'annotations réalisées par d'autres est une forme de collaboration indirecte, puisque l'expert « profite » du travail précédemment effectué et s'en inspire. Au-delà, le fait de faire annoter en parallèle un échantillon du corpus et d'utiliser l'accord inter-annotateurs qui en résulte pour modifier le guide d'annotation, qui va être à son tour utilisé par les annotateurs, est également une forme de collaboration, car le travail fourni en amont par certains influence l'annotation à venir pour tous. Ce type de collaboration indirecte est peu visible pour les annotateurs, puisque ceux-ci n'en voient que la manifestation négative : leurs désaccords.

Une forme de collaboration plus évidente est la possibilité d'échanger avec les autres annotateurs et de garder une trace de ces échanges (Lortal *et al.*, 2006; De la Clergerie, 2008). Si `EasyRef` (De la Clergerie, 2008) le fait sous forme indirecte de rapports de bugs (peu visibles par les annotateurs), `AnT&CoW` (Lortal *et al.*, 2006) [16] le propose directement et visiblement, sous forme de forum. Cette possibilité n'est malheureusement pas encore prise en compte dans les outils de gestion d'annotation existants.

La collaboration est donc un mécanisme présent depuis longtemps dans l'annotation. Il est cependant indubitable que nous assistons, depuis l'apparition du Web 2.0, au développement de nouvelles formes de collaboration. Outre Wikipedia, déjà citée, un jeu sérieux comme `Phrase Detectives` permet une collaboration indirecte (*via* les scores obtenus et les classements des joueurs) et visible pour les joueurs. Quant aux plate-formes de myriadisation du travail parcellisé comme `Amazon Mechanical Turk`, le principe qui les sous-tend est une collaboration indirecte par le consensus entre travailleurs, invisible pour ceux-ci (ils n'ont que très peu de retours sur le travail qu'ils effectuent).

Cette évolution va de pair avec une prise en compte accrue de l'importance de la formation des annotateurs et de l'évaluation de l'annotation. `Phrase Detectives` met par exemple l'accent sur ces deux points, en rendant la phase de formation obligatoire, et en réalisant des comparaisons régulières des performances des annotateurs avec une référence. Un des objectifs de la collaboration est d'ailleurs de faciliter la formation des annotateurs, ne serait-ce que par la validation de leurs annotations.

Un autre objectif de la collaboration, en particulier dans les jeux sérieux, est la motivation des participants. Ce bénéfice de la collaboration devrait selon nous davantage être pris en compte, y compris dans les modes d'annotation traditionnels, dans lesquels les annotateurs ont encore trop peu d'espaces d'échanges. Les outils d'annotation n'offrent en effet généralement pas de fonctionnalité de type forum qui permettrait aux annotateurs et au gestionnaire de garder une trace des décisions prises au fil de l'eau.

16. `AnT&CoW` n'est pas un outil d'annotation pour le TAL, raison pour laquelle nous ne le présentons pas en détail en annexe A.

Évolution vers la prise en compte de la gestion d'annotations

A notre connaissance, le premier article à faire mention explicite de la gestion de l'annotation est (Kaplan *et al.*, 2010), qui présente SLATE, un outil offrant non seulement des fonctionnalités d'aide à l'annotation mais également, et c'est ce qui fait son originalité, une vision plus « macro » de celle-ci, dont une définition claire des acteurs (« administrateur » et « annotateur »), considérés comme complètement distincts. Grâce à SLATE, l'administrateur peut répartir et assurer le suivi des textes à annoter et donc gérer le corpus, un versionnage de celui-ci au cours de l'annotation est d'ailleurs prévu et toutes les annotations sont identifiées par le numéro de version du projet, qui correspond également à celui du jeu d'étiquettes, au moment de l'annotation. L'outil permet également d'effectuer des comparaisons et fusion d'annotations. Cette définition plus formelle et plus explicite des rôles se retrouve aussi dans GATE Teamware, qui en identifie trois (gestionnaire de la campagne, éditeur ou curateur et annotateur). Quant aux fonctionnalités de gestion de l'annotation, elles sont semblables dans Djangology (Apostolova *et al.*, 2010) et GATE Teamware (Bontcheva *et al.*, 2010) et sont en cours de développement dans CCASH et Callisto.

Cette évolution vers la gestion de l'annotation se confirme donc. Elle apparaît cependant bien avant 2010. Ainsi, des interfaces de comparaison des annotations et de calcul des accords inter-annotateurs ont été ajoutées dans beaucoup d'outils (Knowtator, MMAX2, Glozz, Serengeti, SYNC3). Par ailleurs, si les plate-formes de TAL, comme GATE, intègrent évidemment l'utilisation de traitements automatiques pour optimiser l'annotation manuelle, la plupart des outils prévoient la prise en compte de ces traitements (moyennant l'adaptation au format de l'outil, comme dans Glozz) et certains vont jusqu'à en proposer, comme la propagation automatique d'annotations (ou *tag dictionary*), disponible dans Djangology et CCASH. Étant donnés les biais potentiels induits par les traitements automatiques (Fort *et al.*, 2009), nous considérons qu'il revient au gestionnaire de la campagne de décider d'appliquer ces traitements, qui relèvent donc de la gestion de l'annotation et non de l'annotation *stricto sensu*. Il en va de même pour la possibilité de modifier le schéma d'annotation en cours de campagne (proposée par UAM CorpusTool, GATE et ANALEC). Enfin, des outils comme Slate ou EasyRef offrent la possibilité de définir des contraintes sur l'annotation (par exemple, dans EasyRef, des menus contextuels qui ne permettent que les actions autorisées dans le contexte), qui, là encore, sont du ressort du gestionnaire.

Une autre fonctionnalité proposée par de nombreux outils d'annotation « simples » est indubitablement liée à la gestion de l'annotation, même si elle peut aussi être utile aux annotateurs : le suivi de la campagne d'annotation. PALinkA offre ainsi la possibilité de voir le temps passé sur l'annotation. De même, il est possible de configurer brat pour enregistrer le temps passé par un annotateur sur un document et sur chaque action d'édition et de typage (une fonctionnalité semblable est proposée dans CCASH). EasyRef conserve quant à lui des traces des activités sur le système, grâce à des logs. Ce suivi,

effectué en local dans les outils d'annotation, s'enrichit d'un suivi plus global dans les outils de gestion d'annotation comme SLATE ou GATE Teamware, suivi permettant de visualiser l'état d'avancement de la campagne et des annotateurs. Cette dernière fonctionnalité implique cependant la prise en compte de la notion de corpus, qui n'est pas encore présente dans tous les outils d'annotation (elle est inexistante dans Glozz, par exemple).

3.2.4 L'impossible outil d'annotation à tout faire ?

La diversité des annotations (voir section 2.4) implique une diversité des outils. De l'écrit à la vidéo en passant par l'oral, du niveau micro (POS-tagging) au niveau macro (discours), il est difficilement envisageable que l'on puisse créer un outil totalement générique répondant aux besoins et aux contraintes de chacun (conservation ou non du format d'origine, par exemple). Beaucoup de gestionnaires de campagnes d'annotation préfèrent en outre développer un nouvel outil adapté aux contraintes de leur campagne, et qui peut être un simple plugin XEmacs (voir section 5.5), plutôt que d'essayer de se fondre dans la logique d'un outil existant, ce qui est coûteux en temps, peut biaiser la campagne du fait de limitations intrinsèques et risque de se révéler décevant. Ainsi, utiliser GATE pour de l'annotation manuelle était, jusqu'à sa version 5.0, très compliqué, car cette partie était peu documentée et l'outil présentait des problèmes d'affichage pour les gros corpus.

Si certains outils sont plus utilisés que d'autres, souvent parce que plus complets et mieux maintenus (c'est le cas par exemple de GATE, de CCASH, et, dans une moindre mesure, de Glozz et MMAX2), il n'existe à ce jour pas d'outil faisant l'unanimité.

Il faut souligner également que développer un outil d'annotation générique, fiable et documenté est une tâche de longue haleine. Ainsi, l'outil Glozz a nécessité le travail de deux personnes (Yann Mathet et Antoine Widlöcher), pour une durée de six mois pour la conception et autant pour le développement (Yann Mathet, communication personnelle, le 12 janvier 2011).

S'il existe pléthore d'outils d'aide à l'annotation, très peu permettent également de gérer une campagne d'annotation. Les trois seuls dont nous ayons connaissance sont Slate (Kaplan *et al.*, 2010) et GATE Teamware (Bontcheva *et al.*, 2010), toujours en cours de développement, auxquels il faut ajouter Djangology (Apostolova *et al.*, 2010), qui ne semble pas maintenu (voir annexe A.4). Ces outils, bien que très complets (en particulier GATE Teamware), ne proposent aucune fonctionnalité de préparation de campagne. Ils n'offrent aucun moyen de prévoir les difficultés que va présenter la campagne et donc de sélectionner les traitements automatiques adéquats à mettre en place.

Notre travail vise à combler ce vide et à proposer des outils permettant de préparer une campagne d'annotation, quelle qu'elle soit, de la manière la plus efficace possible.

3.3 Évaluation de l'annotation manuelle

3.3.1 Motivations

La qualité des annotations manuelles produites est fondamentale car elle a un impact direct sur les applications concernées (voir section 2.3). Ainsi, il a été démontré que les outils d'apprentissage automatique apprennent à faire les mêmes erreurs que les annotateurs si ces erreurs suivent un modèle et ne correspondent pas à un simple « bruit » dans l'annotation (Reidsma et Carletta, 2008). Il est par ailleurs évident que des erreurs dans la référence annotée peuvent avoir des conséquences graves sur une évaluation, qui ne pourrait pas être considérée comme fiable. Enfin, une annotation de mauvaise qualité entrainerait des résultats trompeurs dans le cas d'une analyse linguistique pour la création de systèmes à base de règles, par exemple.

Pour être considérée comme « bonne », une annotation (au sens de l'annotation de corpus) doit être valide, autrement dit, les catégories associées à la source doivent être correctement typées et correctement ancrées au bon empan de texte. Cela étant, l'annotation manuelle est, par définition, un processus d'interprétation, il n'existe donc pas de « vérité absolue ». Nous ne pouvons donc pas mesurer directement la validité de l'annotation manuelle, si tant est que cela ait un sens. Nous devons nous contenter de mesurer sa fiabilité, autrement dit, la cohérence avec laquelle les annotateurs ont réalisé l'annotation, ce qui démontre qu'ils ont (ou pas) bien assimilé le schéma d'annotation et que ce dernier était cohérent.

Or, cette fiabilité ne peut être évaluée que par le calcul de l'accord entre les annotateurs (inter-annotateurs). On fait alors annoter par ceux-ci un même échantillon du corpus, afin de comparer leurs annotations. Gut et Bayerl (2004) ajoutent qu'outre l'accord inter-annotateur, qui permet de mesurer la stabilité de l'annotation, il faut calculer l'accord intra-annotateur (de l'annotateur avec lui-même, plus tard dans la campagne), qui donne une indication de la reproductibilité de l'annotation. S'il est important de calculer l'accord inter-annotateur, il n'est pas nécessaire de le réaliser sur tout le corpus, ne serait-ce que pour des raisons de coût. Il est en revanche conseillé de le calculer tôt, afin d'identifier les problèmes rapidement et de modifier l'annotation en conséquence pour en améliorer la qualité. Il en va de même en ce qui concerne l'accord intra-annotateur.

Un important travail sur les coefficients pour le calcul de l'accord inter-annotateurs en TAL a été réalisé par Artstein et Poesio (2008), nous nous en inspirons ici largement.

3.3.2 Coefficients simples

Nous reprenons dans cette section les notations d'Artstein et Poesio (2008).

Accord observé

La mesure la plus évidente d'accord inter-annotateurs est le pourcentage d'accord ou accord observé (A_o), qui correspond à la proportion d'éléments sur lesquels les annotateurs sont d'accord, autrement dit, le nombre total d'éléments (i) « annotables » pour lesquels il y a accord (agr), divisé par le nombre total d'éléments annotables. Il est défini comme suit :

$$A_o = \frac{1}{i} \sum_{i \in I} agr_i$$

Ce coefficient ne prend cependant pas en compte le hasard, qui peut influencer les résultats. Ainsi, un schéma d'annotation comprenant un petit nombre de catégories donnera de meilleurs accords observés, uniquement du fait du hasard. En outre, cette mesure d'accord ne compense pas la distribution des éléments dans les catégories : une catégorie prépondérante va très largement influencer l'accord observé (Artstein et Poesio, 2008).

Des coefficients simples plus évolués existent, qui permettent de prendre en compte le hasard. Les plus courants de ces coefficients (κ et π) sont décrits en détail dans (Artstein et Poesio, 2008) et sont présentés dans la suite, accompagnés de coefficients moins connus (S et R de Finn). (Artstein et Poesio, 2008) présentent S, κ et π de manière regroupée, car ils présentent de grandes similarités. Ils sont en particulier obtenus tous les trois à partir de la formule suivante, dans laquelle seul l'accord attendu (A_e) diffère selon le coefficient :

$$S, \kappa, \pi = \frac{A_o - A_e}{1 - A_e}$$

La différence entre ces coefficients réside dans la manière de calculer l'accord attendu en fonction des hypothèses concernant le comportement des annotateurs dans le cas d'une annotation des éléments au hasard.

S

S suppose que les annotations réalisées au hasard suivent une distribution uniforme dans les différentes catégories (k) (Bennett *et al.*, 1954). L'accord attendu (A_e^S) est donc ici directement fonction du nombre de catégories. il est calculé comme suit :

$$A_e^S = \frac{1}{k}$$

Par conséquent, plus le nombre de catégories est élevé, plus l'accord attendu est faible, ce qu'il est en général, sa valeur maximale étant de $0,5$ ($\frac{1}{2}$) pour deux catégories.

S se calculant comme suit :

$$S = \frac{A_o - A_e^S}{1 - A_e^S}$$

il suffit d'ajouter des catégories « vides » pour obtenir un coefficient S élevé. Ce coefficient n'est donc pas satisfaisant. Il est heureusement peu utilisé.

R de Finn

Le coefficient R (Finn, 1970), peu présent en TAL, a cependant été utilisé par Laignelet et Rioult (2009) dans le cadre d'une campagne d'annotation présentant une importante disproportion entre catégories, suivant en cela une suggestion de Hripcsak et Heitjan (2002).

Le coefficient R est calculé selon la formule suivante :

$$R = 1 - \frac{\text{Variance observée}}{\text{Variance attendue}}$$

la variance observée étant la moyenne des variances sur les éléments annotés et la variance attendue étant la variance de la distribution uniforme discrète à n catégories (ci-dessous *nb catégories*), soit [17] :

$$\text{Variance attendue} = \frac{(\text{nb catégories})^2 - 1}{12}$$

ce coefficient modélise le hasard comme S, en considérant une distribution uniforme des catégories et n'est donc pas plus sensible que S à la répartition des éléments dans les catégories. Le coefficient R de Finn n'apporte donc pas plus que S dans des cas de dispersion des annotations et donc de dissymétrie des catégories.

π de Scott (ou κ de Carletta)

Le coefficient π (Scott, 1955), appelé également K dans (Siegel et Castellan, 1988) ou Kappa dans (Carletta, 1996), considère comme S que les distributions réalisées par les annotateurs par hasard sont équivalentes, mais il suppose que la répartition des éléments (i) entre catégories (k) n'est pas homogène et qu'elle peut être estimée par la répartition moyenne réalisée par les annotateurs. L'accord attendu (A_e^π) est donc calculé de la façon suivante, n_k étant le nombre d'affectations à k pour les deux annotateurs :

$$A_e^\pi = \sum_{k \in K} (\frac{n_k}{2i})^2$$

17. Finn (1970) ne détaille pas le calcul de cette variance attendue, mais on le trouve dans les sources de la librairie *irr* du logiciel `R`. Pour une explication plus approfondie, voir : http://mathworld.wolfram.com/DiscreteUniformDistribution.html.

π se calcule ensuite comme suit :

$$\pi = \frac{A_o - A_e^\pi}{1 - A_e^\pi}$$

κ de Cohen

Le coefficient Kappa (κ) (Cohen, 1960) suppose dans sa modélisation du hasard que la répartition des éléments entre catégories peut être différente pour chaque annotateur. Dans ce cas, la probabilité pour qu'un élément (i) soit assigné dans une catégorie (k) est le produit de la probabilité que chaque annotateur l'assigne dans cette catégorie. L'accord attendu (A_e^κ) est donc calculé de la façon suivante, n_{c1k} étant le nombre d'affectations à k pour l'annotateur 1 :

$$A_e^\kappa = \sum_{k \in K} \frac{n_{c1k}}{i} . \frac{n_{c2k}}{i}$$

κ se calcule ensuite comme suit :

$$\kappa = \frac{A_o - A_e^\kappa}{1 - A_e^\kappa}$$

Il faut noter que $\pi \leqslant \kappa$ et que κ et π sont souvent très proches (Di Eugenio et Glass, 2004), ce qui signifie qu'il y a peu de biais entre les annotateurs. Il est donc intéressant de calculer ces deux coefficients.

Généralisation à plus de deux annotateurs

Multi-π ou κ de Fleiss Une fois encore, les termes employés sont trompeurs et le κ de Fleiss est en fait une généralisation du π de Scott et non du κ de Cohen. Cette ambiguïté est accentuée par le fait qu'il correspond au coefficient appelé K par (Siegel et Castellan, 1988).

Quoi qu'il en soit, le κ de Fleiss permet de prendre en compte plus de deux annotateurs (Fleiss, 1971). Il le fait de manière simple, en considérant le nombre de paires d'annotateurs en accord par rapport à toutes les paires de jugements possibles pour l'unité considérée (Artstein et Poesio, 2008).

Il est calculé selon la formule générale des Kappas, en y intégrant les accords par paires :

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

avec :

$$\bar{P} = \frac{1}{i}\sum_{i \in I} agr_i = \frac{1}{ic(c-1)}\sum_{i \in I}\sum_{k \in K} n_{ik}(n_{ik}-1)$$

et, bien sûr, à l'image de π :

$$\bar{P}_e = \sum_{k \in K} (\frac{n_k}{ic})^2$$

c étant le nombre d'annotateurs (*coders*).

Multi-κ Ce coefficient est lui une généralisation du κ de Cohen à plus de deux annotateurs (Davies et Fleiss, 1982). Il reprend donc le mode de calcul de l'accord attendu utilisé dans le κ de Cohen.

3.3.3 Coefficients pondérés

Selon Artstein et Poesio (2008), π et κ ont pour défaut de traiter tous les désaccords de la même manière et seuls des coefficients pondérés permettent de donner plus d'importance à certains désaccords.

κ_ω et α

Artstein et Poesio (2008) détaillent deux coefficients pondérés : la version pondérée de κ, κ_ω (Cohen, 1968) et Alpha (α) (Krippendorff, 1980, 2004). Ces deux coefficients prennent pour base le désaccord entre annotateurs et utilisent une distance entre les catégories décrivant à quel point deux catégories sont distinctes l'une de l'autre.

Les coefficients pondérés κ_ω et α sont calculés à partir de la formule suivante :

$$\kappa_\omega, \alpha = 1 - \frac{D_0}{D_e}$$

où D_0 représente le désaccord observé entre les annotateurs et D_e représente le désaccord attendu (*expected*), autrement dit, si l'affectation est réalisée au hasard. Le désaccord attendu de κ_ω et d'α suit la même logique que κ et π respectivement, et inclut la notion de distance entre catégories.

On trouve dans (Artstein et Poesio, 2008) une discussion sur la définition de cette distance en fonction du type d'annotation. Elle permet entre autres de traiter des annotations de structures complexes en introduisant plusieurs valeurs de distance entre

annotations. Cette méthode présente l'inconvénient de complexifier l'interprétation des résultats.

Pour pondérer l'accord inter-annotateurs, les distances entre catégories sont définies à partir de connaissances préalables sur la tâche d'annotation. Or, cela pose le problème de cette définition, fondée, en quelque sorte, sur l'« intuition » et non sur une réalité terrain. Nous ne nous attarderons donc pas plus sur ces coefficients.

Mesure Glozz

Les créateurs de l'outil `Glozz` (voir annexe A), Yann Mathet et Antoine Widlöcher (GREYC/CNRS), ont développé une mesure originale pour le calcul de l'accord inter-annotateurs, qui est présentée dans (Mathet et Widlöcher, 2011a) et qu'ils appellent « Mesure Glozz ». Cette mesure n'est pas directement pondérée, mais elle se veut plus tolérante que les coefficients de type Kappa, en particulier en ce qui concerne la segmentation des unités, et nécessite en outre la définition d'une distance. Cette mesure s'inspire également de la famille des Kappas, au sens où elle prend en compte une forme de hasard. Enfin, elle se veut « holiste » car elle prend les annotations de tout le texte en considération plutôt que de faire des comparaisons locales entre unités, et unifiée, car elle ne dissocie pas alignement et mesure d'accord, procédant aux deux simultanément, et considérant que l'un ne peut se faire indépendamment (ni en amont) de l'autre.

Pour ce faire, les auteurs s'appuient sur la notion de « désordre du système constitué par l'ensemble des jeux d'annotations d'un texte » (Mathet et Widlöcher, 2011a). Ce désordre prend la forme de deux types de « dissimilarités », les dissimilarités positionnelles et catégorielles. Nous renvoyons à (Mathet et Widlöcher, 2011a) pour le détail des calculs de ces dissimilarités. Notons cependant que la dissimilarité catégorielle implique, comme pour α et κ_ω, la définition d'une distance entre catégories et que cette définition, telle que proposée dans l'article, est empirique [18].

La mesure Glozz est calculée selon la formule suivante, pour tout jeu d'annotations j sur un corpus c :

$$accord(j) = \frac{e_{aleatoire}(c) - e(j)}{e_{aleatoire}(c)}$$

L'« entropie » (le désordre) d'un alignement unitaire correspond à la moyenne des dissimilarités de ses unités constituantes. L'entropie aléatoire, $e_{aleatoire}(c)$, peut être calculée selon différentes méthodes, dont celle privilégiée dans (Mathet et Widlöcher, 2011a), qui est utilisée par `Glozz`. Cette méthode consiste à observer les productions faites par les annotateurs sur l'ensemble des textes, et à générer ensuite de façon indépendante du contenu du texte des multi-annotations qui respectent la distribution

18. Elle est maintenant établie par une matrice calculée automatiquement par observation du corpus (Y. Mathet, GREYC-CNRS, communication personnelle, le 17 août 2012)

statistique du corpus, à la fois en termes positionnels et en termes catégoriels. Une version simplifiée de cette mesure est intégrée à `Glozz` et permet de réaliser directement des calculs d'accords inter- et intra-annotateurs sur les textes annotés dans l'outil. Elle n'est toutefois pas encore applicable aux relations. En outre, la complexité de cette mesure la rend difficilement utilisable en dehors de `Glozz`.

Une dernière limite de cette mesure est qu'elle ne permet pas encore de se comparer à d'autres campagnes, car elle est malheureusement encore peu utilisée. Elle est *a priori* plus tolérante que le Kappa (qu'il soit de Cohen ou de Carletta), mais cela reste à démontrer mathématiquement. Une solution méthodologiquement intéressante consiste donc à la calculer en complément d'un Kappa, ce que nous avons fait lors de notre campagne d'annotation de matchs de football (voir section 5.3).

3.3.4 Utilisation de métriques d'évaluation des outils

F-mesure

Dans certaines campagnes d'annotation, les responsables de l'évaluation de l'annotation manuelle utilisent la F-mesure plutôt que les coefficients classiques d'accord inter-annotateurs. Ce choix permet en particulier de contourner le problème posé par la définition des annotables pour le calcul du κ dans le cas des entités nommées (Alex *et al.*, 2010; Grouin *et al.*, 2011).

La F-mesure est une mesure courante en TAL, issue de la recherche d'informations. Elle correspond à la moyenne harmonique du rappel et de la précision :

$$\text{F-mesure} = 2.\frac{\text{précision.rappel}}{\text{précision} + \text{rappel}}$$

La F-mesure est aussi appelée F1-mesure, car il s'agit d'un cas particulier de $F\beta$, la moyenne harmonique pondérée du rappel et de la précision :

$$F\beta = (1 + \beta^2).\frac{\text{précision.rappel}}{\beta^2.\text{précision} + \text{rappel}}$$

la valeur de β permettant de favoriser :
– le rappel ($\beta = 2$)
– la précision ($\beta = 0.5$)
Dans les deux cas, le rappel et la précision sont calculés comme suit :

$$\text{Rappel} = \frac{\text{Nb d'annotations correctes trouvées}}{\text{Nb d'annotations correctes attendues}}$$

$$\text{Précision} = \frac{\text{Nb d'annotations correctes trouvées}}{\text{Nb total d'annotations trouvées}}$$

Par définition, le rappel et la précision ne peuvent être calculés qu'à partir d'une référence. Or, nous n'en avons pas dans le cas de l'accord inter-annotateurs. Cela étant, on peut considérer que le travail de chacun des annotateurs sert de référence à l'annotation de l'autre. Il est en outre inutile de calculer la F-mesure dans les deux sens, car le rappel de l'un est la précision de l'autre (voir (Hripcsak et Rothschild, 2005)). Une seule F-mesure suffit donc pour obtenir une estimation de la fiabilité de l'annotation.

Cependant, si (Hripcsak et Rothschild, 2005) mettent l'accent sur l'utilité de cette mesure dans certains cas d'annotation, ils reconnaissent aussi ses limites. En effet, contrairement à π et κ (voir ci-dessus), la F-mesure ne prend pas en compte le hasard. Toutefois, il a été démontré dans (Hripcsak et Rothschild, 2005) que dans les cas où le nombre d'annotables est très grand, les Kappas tendent vers la F-mesure (ce que nous avons pu observer dans (Grouin *et al.*, 2011)), ce qui justifie l'utilisation de celle-ci plutôt que d'un Kappa.

Slot Error Rate

A la différence des mesures de type Kappa, le *Slot Error Rate* (Makhoul *et al.*, 1999) (SER) est, comme le *Word Error Rate* dont il est inspiré, une mesure d'erreur et non de succès. Le SER correspond au ratio du nombre total d'erreurs d'annotation (de substitution, de suppression et d'insertion) divisé par le nombre total d'annotations de la référence.

Il est calculé comme suit :

$$SER = \frac{S + D + I}{C + S + D}$$

avec :

- substitute (S) : le nombre de substitutions
- delete (D) : le nombre de suppressions
- insert (I) : le nombre d'insertions
- correct (C) : le nombre de réponses correctes

Cette mesure a été adaptée pour l'évaluation Quæro sur les entités nommées (Galibert *et al.*, 2010). Dans cette expérience, les auteurs ont considéré :

- delete (D) : le nombre de réponses attendues non ramenées
- insert (I) : le nombre de réponses ramenées non attendues
- type (T) : le nombre de réponses typées correctement (avec bonnes frontières)
- front (F) : le nombre de réponses avec de bonnes frontières (mais mal typées)
- type-front (TF) : le nombre de réponses attendues avec mauvaises frontières et mauvais type
- reference (R) : le nombre de réponses attendues

Le SER est alors calculé comme suit :

$$SER = \frac{D + I + TF + 0,5.(T + F)}{R}$$

Le SER a le mérite de permettre la prise en compte de différents types de désaccords (sur les types et sur les frontières, par exemple). Si cette mesure semble *a priori* nécessiter une référence, on peut, de même que pour la F-mesure, considérer les résultats d'un annotateur comme la référence de l'autre et vice-versa. Par contre, dans le cas du SER, il faut calculer tous les SER et en faire la moyenne, car ils ne sont pas symétriques. En outre, le SER ne prend pas en compte le hasard.

Nous avons utilisé le SER pour l'évaluation de l'annotation du mini corpus de référence des campagnes d'annotation en entités nommées structurées (voir section 5.5).

3.3.5 Signification des résultats

Artstein et Poesio (2008) détaillent dans une section de leur article les différentes échelles qui ont été fournies au fil des ans (voir figure 3.1) pour interpréter les Kappas et soulignent à quel point il est difficile de définir un seuil qui ait du sens.

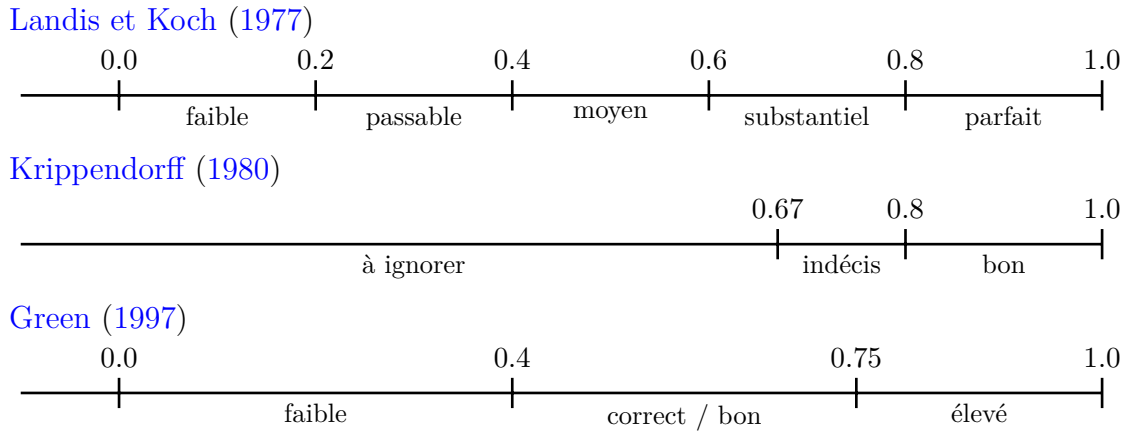


FIGURE 3.1: Échelles d'interprétation des Kappas

Ils concluent prudemment en proposant un seuil de « qualité raisonnable » (« *reasonable quality* ») de 0,8 pour les Kappas, tout en ajoutant douter « qu'un seul seuil puisse convenir à tous les types d'annotation » (« *doubt that a single cutoff point is appropriate for all purposes* »). Les travaux de Gwet (2012), qui présentent divers coefficients d'accord inter-annotateurs, insistent également sur les problèmes liés à leur interprétation.

D'autres études, concernant l'évaluation de la qualité de l'annotation manuelle, ont permis d'identifier des facteurs influençant l'accord inter-annotateurs, donnant par là même des clefs sur les comportements des mesures utilisées. Ainsi, Gut et Bayerl (2004) ont démontré que l'accord inter-annotateurs et la complexité de la tâches sont corrélés : plus le jeu d'étiquettes est important, plus l'accord est faible. Les mêmes auteurs montrent toutefois que les catégories prêtant à confusion sont en nombre limité. La méta analyse présentée dans (Bayerl et Paul, 2011) étend cette recherche et permet à ses auteurs d'identifier huit facteurs influençant l'accord inter-annotateurs : le « domaine » (nous dirions plutôt le type d'annotation), le nombre d'annotateurs, la formation des annotateurs, l'utilisation faite de l'annotation, la connaissance du domaine, la langue, le nombre de catégories et la méthode de calcul de l'accord. Ils en déduisent des recommandations afin d'améliorer la qualité de l'annotation manuelle. Cependant, aucune de ces analyses ne donne une image claire du comportement des mesures d'accord ou de leur signification.

Les expériences détaillées dans (Reidsma et Carletta, 2008) constituent une avancée intéressante de ce point de vue. Elles se focalisent sur les effets produits par les erreurs d'annotation manuelle sur les systèmes d'apprentissage automatique et montrent l'importance qu'a la forme du désaccord éventuel sur la qualité des systèmes (les annotations « bruitées » de manière aléatoire étant tolérables, mais pas les désaccords systématiques). Ce travail, s'il met en perspective les résultats obtenus avec des coefficients de type Kappa, ne présente qu'un point de vue « outils », qui plus est limité à ces coefficients en particulier.

Le domaine manque donc d'un outil proposant une image claire et complète du comportement des mesures d'accord inter-annotateurs, qui permettrait de mieux qualifier les résultats obtenus avec tel ou tel coefficient, lors de telle ou telle campagne d'annotation.

3.3.6 Conclusion

Il est aujourd'hui difficile, voire impossible, de publier un article de recherche sur un corpus annoté sans fournir de mesure précise de sa qualité (en l'occurrence sa fiabilité), si possible sous forme d'une métrique unique et bien connue, ladite métrique devant en outre permettre de comparer un corpus annoté à un autre. Cette évolution est sans aucun doute bénéfique sur le principe, mais donne lieu à toutes sortes d'excès dans le domaine.

En effet, étant données la complexité et la multiplicité des mesures existantes, on voit apparaître quatre tendances peu satisfaisantes. La première consiste à utiliser un coefficient connu, souvent un Kappa, de Cohen (Cohen, 1960) ou de Carletta (Carletta, 1996) (donc π (Scott, 1955)) [19], sans pour autant détailler la définition des annotables

19. Pour plus de détails sur ce problème de terminologie, voir (Artstein et Poesio, 2008)

choisie pour cela. La seconde consiste à essayer de trouver un autre coefficient, qui, soit donnerait de meilleurs résultats sur la campagne, soit y semblerait plus adapté (Laignelet et Rioult, 2009), mais qui n'est généralement reconnu ou utilisé par personne d'autre. La troisième tendance, la plus « conservatrice », consiste à utiliser une F-mesure, qui permet certes de satisfaire au critère de comparabilité, mais ne tient pas compte du hasard (Hripcsak et Rothschild, 2005). Enfin, la dernière tendance consiste à donner un résultat d'« accord inter-annotateurs » sans préciser ce qu'il couvre (de fait, souvent un simple accord observé) (Paroubek *et al.*, 2010).

Malgré les efforts pédagogiques louables de Artstein et Poesio (2008), le domaine manque aujourd'hui cruellement d'un inventaire précis et argumenté des différentes mesures disponibles, en fonction des types de campagnes.

Il nous semble également nécessaire d'y associer les différents moyens de synthèse des données disponibles (matrice de confusion, tableau de Krippendorff, etc.) et de mettre l'accent sur l'importance des choix effectués lors de leur création.

Surtout, le problème de l'interprétation des résultats obtenus est connu et souligné dans (Artstein et Poesio, 2008) : un Kappa de $0,75$ est-il satisfaisant ? et pour quelle tâche d'annotation ? De même, comment comparer un Kappa et une mesure Glozz ?

3.4 Quelques solutions proposées

3.4.1 Annotation assistée par ordinateur

Nous regroupons sous le terme « annotation assistée par ordinateur » toutes les techniques, notamment issues du TAL, permettant de faciliter le travail de l'annotateur humain, à l'exception de l'outil d'aide à l'annotation (qui peut en intégrer).

Cette assistance consiste en général à proposer automatiquement des annotations de même nature que ce qu'aurait fait un annotateur humain. Il existe plusieurs techniques de ce type, la propagation d'étiquettes, la pré-annotation automatique et l'apprentissage actif. Nous les présentons ici, brièvement pour la première et la dernière, de manière plus approfondie pour la seconde.

Propagation d'étiquettes

La propagation d'étiquettes consiste à ré-utiliser automatiquement les annotations déjà réalisées. Autrement dit, un annotateur ayant associé une ou plusieurs étiquettes à un empan de texte voit ces mêmes étiquettes automatiquement proposées pour l'annotation du même empan de texte, un peu plus loin dans la source. Cette technique très simple permet de faire gagner du temps à l'annotateur tout en lui permettant de

garder un niveau de qualité équivalent (Carmen *et al.*, 2010). Elle est bien entendu d'autant plus efficace que le nombre d'annotations déjà réalisées (ou disponibles) est élevé.

Cela étant, si le fait de proposer certaines étiquettes à l'annotateur plutôt que toutes permet de gagner en temps, il n'est pas certain que cela n'induise pas un biais de l'annotation (l'annotateur privilégiant les étiquettes proposées). Cette tendance est d'ailleurs observée dans (Carmen *et al.*, 2010) : le bénéfice de la propagation d'étiquettes, dans les cas d'une faible couverture de cette propagation, est alors remis en cause.

Pré-annotation totalement automatique

Un autre type de pré-annotation, davantage utilisé, est l'annotation réalisée de manière totalement automatique par un outil (de TAL). L'annotation par l'humain se réduit alors à de la correction d'annotations automatiques.

Cette technique a été utilisée pour l'annotation morpho-syntaxique du *Penn Treebank* (Marcus *et al.*, 1993) et a permis un gain très significatif en temps (deux fois plus rapide que l'annotation directe) et en qualité (taux d'erreurs inférieur de 50 %). Cet article n'étudie cependant ni l'influence de la formation des annotateurs sur les biais potentiels induits par la correction, ni l'impact de la qualité de l'outil de pré-annotation sur le temps et la qualité de la correction.

Dans un article très complet, Dandapat *et al.* (2009) montrent que, dans le cas d'annotations morpho-syntaxiques complexes (celles de l'hindi et du bengali), la pré-annotation du corpus permet un gain de temps, mais pas forcément de cohérence, qui elle, dépend largement de la qualité de la pré-annotation. Ils montrent également que les annotateurs non formés sont davantage influencés par la pré-annotation que les annotateurs formés. Cette étude, bien que très intéressante et très complète, manque cependant d'une référence qui permettrait d'évaluer plus précisément la qualité des annotations et des corrections. En outre, elle ne considère que deux niveaux de qualité de l'outil de pré-annotation (élevé et faible).

Alex *et al.* (2008) ont mené des expériences dans le domaine biomédical, dans le cadre d'une tâche de « curation » d'interactions entre protéines. La curation consiste à lire des articles et à en entrer les informations extraites dans un formulaire. Ils montrent qu'une pré-annotation parfaite du corpus permet une réduction de plus d'un tiers du temps de curation, ainsi qu'un meilleur rappel. Une pré-annotation de qualité moindre apporte également un gain de temps, bien qu'inférieur (moins d'un quart). Ils ont également testé les effets d'un meilleur rappel et d'une meilleure précision de la pré-annotation sur un annotateur, qui a jugé le rappel plus utile que la précision.

Rehbein *et al.* (2009) ont mené des expériences relativement complètes sur le sujet, dans le domaine de l'annotation de cadres sémantiques. Ils ont demandé à 6 annotateurs d'annoter ou de corriger une annotation de cadres sémantiques. Là encore, la pré-annotation a été réalisée à l'aide d'outils de deux niveaux de qualité, état de l'art et amélioré. Les résultats de ces expériences sont un peu décevants, les auteurs n'ayant pas pu démontrer un gain de temps dû à la pré-annotation. Selon eux, cela tient, au moins en partie, à une « interaction entre le gain de temps dû à la pré-annotation et celui dû au niveau de formation des annotateurs ». Ils ont dû exclure, pour la même raison, certains des résultats concernant l'évaluation de la qualité. Ils ont en revanche pu démontrer qu'une pré-annotation de mauvaise qualité et bruitée ne corrompait globalement pas le jugement humain.

Nous avons par ailleurs remarqué que la pré-annotation introduit un biais dans l'annotation d'entités nommées (Fort *et al.*, 2009) (voir section 5.6.3), les annotateurs privilégiant ce qui est déjà annoté, au détriment des entités non pré-annotées présentes dans le texte. Ce type de biais ne devrait pas apparaître dans une tâche telle que l'annotation morpho-syntaxique, puisque tout le texte doit être annoté. Cependant, les annotateurs de ce type de tâches ne sont pas à l'abri d'un excès de confiance en la pré-annotation.

Dans un domaine complètement différent, Barque *et al.* (2010) ont utilisé une suite d'outils de TAL, `MACAON`, pour identifier automatiquement les composantes centrales et périphériques de définitions lexicographiques. Ce pré-traitement s'est révélé décevant, n'ayant pas permis de gain de temps significatif. Les auteurs considèrent que ces mauvais résultats sont dus à la qualité de la pré-annotation, ce type d'annotation automatique n'en étant qu'à ses débuts, à la différence de l'annotation morpho-syntaxique.

Malgré un état de l'art conséquent, nous constatons donc que, d'une part, peu d'études se penchent sur les biais de la pré-annotation et que, d'autre part, il reste à établir le seuil de qualité de l'outil d'annotation à partir duquel on peut prévoir que celle-ci sera bénéfique pour l'annotation.

Apprentissage actif

L'apprentissage actif, ou *active learning* (Cohn *et al.*, 1995; Engelson et Dagan, 1996), consiste à annoter manuellement une petite partie d'un corpus, puis à entraîner à partir de ce sous-corpus un outil d'annotation automatique. Cet outil est ensuite utilisé pour ré-annoter le corpus. Les scores obtenus par l'outil sur le corpus[20] sont utilisés pour proposer aux annotateurs humains de corriger les données qui permettront d'améliorer l'outil le plus efficacement possible.

L'apprentissage actif se décompose donc en plusieurs temps :

20. Ce processus implique donc que l'outil soit en mesure d'associer une mesure de confiance à chaque annotation.

1. annotation manuelle d'une portion du corpus,
2. entrainement d'un modèle (statistique) sur ce corpus,
3. application du modèle et obtention de scores,
4. proposition de données à corriger manuellement.

Le processus est itératif et continue jusqu'à obtention d'un modèle satisfaisant, ce qui, de notre point de vue, revient à créer un outil de pré-annotation de bonne qualité.

L'apprentissage actif est encore peu utilisé dans la construction de corpus annotés manuellement, ce qui peut sans doute s'expliquer par le fait que cette technique n'a pas été créée dans ce but.

Cela étant, nous manquons aujourd'hui de données sur l'efficacité de l'apprentissage actif par rapport à la pré-annotation automatique telle que définie plus haut.

3.4.2 Collaboration en ligne

Une multiplicité de systèmes et de dimensions à prendre en compte

La volonté de partager et de participer est inhérente au Web lui-même. Ainsi, le projet Gutenberg [21] de numérisation de livres libres de droit, né avant même Internet (1971), a pu commencer à recruter des volontaires dès le début des années 90 par le biais du Web naissant. Au début des années 2000, le Web 2.0, en permettant aux internautes d'interagir de façon simple avec le contenu des pages Web, a vu l'essor de la participation en ligne. Le terme *crowdsourcing*, que Gilles Adda a traduit par « myriadisation » dans (Sagot *et al.*, 2011) recouvre l'ensemble des systèmes participatifs développés dans le cadre du Web. En effet, le principe du *crowdsourcing* est que le travail est délocalisé (*outsourced*) et est effectué par un grand nombre de personnes (*crowd*, la foule), payées ou non. L'exemple type de système participatif est Wikipedia [22], l'encyclopédie en ligne créée en 2001, qui est devenue aujourd'hui l'un des sites Web les plus visités au monde. Ces systèmes participatifs ont depuis fait florès.

Parmi les systèmes de type wiki, on trouve non seulement des plate-formes génériques, comme l'encyclopédie Wikipedia mais également des systèmes plus spécifiques, comme, dans le domaine de l'annotation, l'outil `Serengeti` (Stührenberg *et al.*, 2007), pour l'annotation sémantique des textes, qui a été utilisé dans le cadre du projet ANAWIKI (Chamberlain *et al.*, 2009a) [23]. On trouve également aujourd'hui des sites proposant de participer à la création ou à l'évaluation d'un projet (potentiellement, une ressource) par le biais d'un jeu ou d'un travail rémunéré.

21. http://www.gutenberg.org
22. http://fr.wikipedia.org
23. http://www.anawiki.org

Si, dans le cas de Wikipedia, le système est transparent (le volontaire sait à quoi il participe) et collaboratif (les volontaires collaborent ensemble à l'écriture d'articles de l'encyclopédie), dans le cas d'ANAWIKI adapté pour l'annotation de relations de co-référence par le jeu (`Phrase Detectives`[24]), le système est moins transparent (les volontaires peuvent savoir à quoi ils participent, mais n'y sont pas obligés pour effectuer la tâche) et il n'est plus directement collaboratif (chacun annote ses textes, sans se préoccuper de ce qu'ont fait les autres), mais parcellisé (chacun réalise une toute petite partie de l'annotation). En ce qui concerne les plate-formes de travail rémunéré de type `Amazon Mechanical Turk` (MTurk), le système n'est pas transparent du tout (les participants ne savent pas quelle est la finalité du travail réalisé) et il est parcellisé (en *Human Intelligence Tasks* pour MTurk).

Quant à la rétribution, si elle est au centre du système dans MTurk, elle est totalement absente de Wikipedia et de nombreux jeux, dont `JeuxDeMots` (Lafourcade et Joubert, 2008). Elle n'est en revanche pas totalement exclue de `Phrase Detectives`, puisque des récompenses, sous forme de bons d'achat, sont attribués aux plus productifs ou par tirage au sort.

Enfin, en ce qui concerne la disponibilité des ressources créées, elle correspond en général au degré de transparence du système. Ainsi, Wikipedia est disponible sous licence *Creative Commons, paternité partage à l'identique* et les données lexicales créées par `JeuxDeMots` (dont les caractéristiques sont semblables à celles de `Phrase Detectives`) sont accessibles[25] et sous licence *Creative Commons Paternité-Partage des Conditions Initiales à l'Identique*. Quant au corpus annoté en relation de co-référence par le biais de `Phrase Detectives`, il est supposé être disponible sur un site Web (Kruschwitz *et al.*, 2009)[26], qui, en date du 26 janvier 2011, ne répond plus. Quant aux ressources créées sur MTurk, elles ne sont généralement pas accessibles, si ce n'est de manière très indirecte, *via* les articles de certains chercheurs ayant utilisé MTurk dans leurs expériences, dont par exemple (Kaisser et Lowe, 2008).

Jeux « ayant un but »

On a vu apparaître en ligne depuis quelques années des jeux sérieux (ou *serious games*). Parmi ceux-ci, les jeux dit « ayant un but » ou *Games With A Purpose* (GWAP) dont le premier est le jeu ESP (ESP game) d'annotation d'images (Von Ahn et Dabbish, 2004; Von Ahn, 2006), ont eu un succès immédiat en TAL. Ainsi, `Phrase Detectives`[27] (Chamberlain *et al.*, 2008b) permet de collecter les jugements de joueurs sur des relations anaphoriques. En moins de trois mois, les auteurs de (Chamberlain *et al.*, 2008a) disent avoir fait annoter 100 000 relations anaphoriques (en 2008)

24. http://anawiki.essex.ac.uk/phrasedetectives/
25. http://www.lirmm.fr/~lafourcade/JDM-LEXICALNET-FR/
26. http://www.anaphoricbank.org/
27. http://anawiki.essex.ac.uk/phrasedetectives/

par plus de 500 joueurs différents. En France, le premier jeu de type GWAP fut `JeuxDeMots` [28] (Lafourcade et Joubert, 2008), permettant la construction d'un réseau lexical d'aujourd'hui plus d'un million de relations.

Ces jeux permettent donc de « déguiser » une tâche, de la rendre plus attractive et ludique pour les participants (à la différence, par exemple, de Wikipedia). Leur succès montre l'intérêt de cette solution. Ils nécessitent cependant une conception non seulement du logiciel, mais également du jeu (Lafourcade et Joubert, 2008; Chamberlain *et al.*, 2009a), mettant l'accent sur la motivation des joueurs, des développements lourds et une attention toute particulière portée à l'interface. Cela représente un coût élevé. Pour `Phrase Detectives`, ce coût a été de six mois de développement, puis d'environ 34 000 dollars de maintenance (Chamberlain *et al.*, 2012). La publicité du jeu, permettant d'attirer des joueurs (Chamberlain *et al.*, 2009b) représente également un coût, au moins en temps. Ainsi, pour `Phrase Detectives`, la publicité et les prix ont représenté un coût d'environ 9 000 dollars (Chamberlain *et al.*, 2012).

Il faut y ajouter la mise en place de moyens de prévention contre les tricheurs, particulièrement nécessaires pour les jeux permettant aux plus actifs de gagner des prix, comme c'est le cas pour `Phrase Detectives` (Chamberlain *et al.*, 2008a). Il est important de noter ici que s'il est possible de la réduire, il est illusoire de penser pouvoir anéantir totalement la tricherie. Il est ainsi facile de gagner beaucoup de points très rapidement dans `Phrase Detectives`, en étant toujours en accord avec les autres annotateurs. Le résultat de cette manœuvre sera encore beaucoup plus efficace en participant à plusieurs, avec cette stratégie en tête [29].

En ce qui concerne la qualité des ressources produites, nous manquons aujourd'hui de recul pour l'évaluer de manière précise. Nous avons cependant travaillé sur le sujet en collaboration avec les créateurs de `Phrase Detectives` et de `JeuxDeMots` (Chamberlain *et al.*, 2012). Si, dans `Phrase Detectives`, l'accord inter-annotateurs entre les experts et le vote majoritaire des joueurs est de $0,7$ (Kappa de Cohen (Cohen, 1960) et de Carletta (Carletta, 1996)) sur les relations anaphoriques, il est de 0 sur l'interprétation des propriétés. Ce mauvais résultat peut être dû à la complexité de la tâche elle-même ou à des directives peu claires. Quant à `JeuxDeMots`, il n'existe pas de référence sur laquelle se fonder pour évaluer les résultats. Cependant, les créateurs du jeu ont créé un jeu complémentaire, `AKI` [30], qui permet d'évaluer la ressource créée de manière indirecte (Lafourcade *et al.*, 2011). Si les résultats obtenus ne correspondent pas à des mesures d'évaluation traditionnelles, les auteurs en tirent la conclusion suivante : « On peut déduire des performance de AKI que 75 % des termes pour lesquels il a été sollicité sont bien indexés, en tout cas suffisamment bien pour permettre le bon choix en cas de désambiguïsation lexicale (avocat : profession *vs* fruit). ». Quoi

28. http://www.lirmm.fr/jeuxdemots/
29. Test réalisé avec des étudiants de Master 2 Sciences Cognitives, option TAL, à Nancy.
30. http://www.jeuxdemots.org/AKI.php

qu'il en soit, utiliser un jeu pour évaluer une ressource produite avec un autre jeu est une idée qui nous semble riche en possibilités.

Microworking ou myriadisation du travail parcellisé

`Amazon Mechanical Turk` (MTurk) (voir Annexe A.1.4) est la première et la plus connue des plates-formes de travail parcellisé (les tâches sont découpées en sous-tâches) et myriadisé (les tâches sont réalisées par une « foule », des non-experts). A la différence des jeux ayant un but, MTurk ne « déguise » pas la tâche derrière une interface ludique. Par ailleurs, le système met directement l'accent sur la rémunération, il suffit, pour le constater, d'aller voir la page d'accueil du site[31] et de la comparer à celle de `Phrase Detectives`[32], par exemple.

Pour nous, ces systèmes, malgré leur succès actuel, ne représentent pas une solution satisfaisante à la question du coût de l'annotation manuelle.

Nous avons travaillé avec Gilles Adda (LIMSI-CNRS) et Kevin B. Cohen (University of Colorado School of Medicine/University of Colorado at Boulder) sur `Amazon Mechanical Turk`, dans le but de faire le point sur le système et de mettre au jour les dangers qu'il présente (Fort *et al.*, 2011a). Nous avons poursuivi et approfondi ce travail par la suite avec Benoît Sagot (Alpage INRIA/Paris 7), Gilles Adda, Joseph Mariani (IMMI) et Bernard Lang (INRIA) (Sagot *et al.*, 2011). Ces travaux présentent l'évolution de l'utilisation de MTurk dans les domaines du traitement automatique des langues et de la parole ces dernières années, et détricotent la légende selon laquelle le système permettrait de faire développer toutes sortes de ressources linguistiques de qualité, pour un prix imbattable et en un temps très réduit, par des gens pour qui il s'agit d'un passe-temps.

Nous démontrons en particulier que :

- la qualité des ressources produites n'est pas toujours satisfaisante, en particulier pour des tâches complexes, du fait de limitations liées à la non-expertise des *Turkers* (travailleurs) et au fonctionnement même de MTurk,
- l'omniprésence des spammeurs rend l'utilisation de MTurk difficile, voire fragilise son existence même,
- si le coût de production des ressources est effectivement bas, il ne l'est pas autant qu'annoncé, car les adaptations nécessaires pour MTurk ne sont jamais prises en compte dans le calcul du coût,
- la grande majorité des *Turkers* les plus actifs considère MTurk comme sa source de revenus principale ou secondaire,

31. https://www.mturk.com
32. http://anawiki.essex.ac.uk/phrasedetectives/

– les droits élémentaires des *Turkers* ne sont pas respectés (salaire non assuré, impossibilité de se syndiquer, d'ester en justice, pas de salaire minimum, même adapté au pays).

Nous reprenons ici certains développements concernant le coût et la qualité des ressources produites.

Des coûts mal évalués Dans la plupart des articles ayant utilisé MTurk, le faible coût de développement de la ressource est mis en avant. Il est vrai que MTurk permet de proposer des rétributions si faibles aux *Turkers* que le coût en est forcément réduit, par exemple 0,005 dollars pour transcrire un segment d'environ cinq secondes de parole téléphonique (Novotney et Callison-Burch, 2010). Il faut cependant nuancer ces chiffres. Tout d'abord, le coût effectif n'est pas toujours calculé avec rigueur. En effet, le temps de développement de l'interface et de mise en place des garde-fous contre le spam est non nul (Callison-Burch et Dredze, 2010). De même, le coût de validation (Kaisser et Lowe, 2008) ou de développement (Xu et Klakow, 2010) post-MTurk permettant de compenser la mauvaise qualité des résultats n'est généralement pas précisément évalué. Ces coûts supplémentaires ne sont jamais pris en compte dans le calcul final. De plus, certaines tâches peuvent se révéler plus coûteuses que prévues. Ainsi, si l'on ne trouve pas de *Turkers* pour faire la tâche, on peut être obligé d'augmenter la rémunération, comme Novotney et Callison-Burch (2010), qui, partant d'un coût très bas (cinq dollars de l'heure transcrite), ont été obligés de le multiplier par sept (37 dollars de l'heure) pour transcrire du coréen par manque de *Turkers* qualifiés.

Une qualité discutable Les *Turkers* étant des non-experts, le *Requester* (fournisseur de tâches) doit découper les tâches complexes en tâches plus simples, afin de les rendre réalisables. Ce faisant, il est amené à faire des choix qui peuvent biaiser les résultats. Un exemple de ce type de biais est analysé dans (Cook et Stevenson, 2010), où les auteurs reconnaissent que le fait de ne proposer qu'une phrase par type d'évolution lexicale (amélioration ou péjoration) influence le résultat.

Plus grave encore que ces biais potentiels, certains chercheurs ont observé que, lorsque la complexité de la tâche augmente, la qualité produite sous MTurk est insuffisante. C'est notamment le cas dans (Bhardwaj *et al.*, 2010), qui démontre que, pour leur tâche de désambiguïsation lexicale, un petit nombre d'annotateurs bien formés produit de bien meilleurs résultats qu'un grand nombre de *Turkers* (le nombre étant supposé contrebalancer la non expertise). De ce point de vue, leurs résultats contredisent ceux de Snow *et al.* (2008) dont la tâche était semblable mais beaucoup plus simple. Cette même difficulté d'obtenir une qualité suffisante sur des tâches complexes apparaît dans (Gillick et Liu, 2010), qui démontre que l'évaluation par des non-experts de systèmes de résumé automatique est « risquée », les *Turkers* n'étant pas capables d'obtenir des résultats comparables à ceux des experts. On retrouve ce problème de qualité dans de nombreux articles, dans lesquels les auteurs ont dû faire valider les

résultats des *Turkers* par des spécialistes (des étudiants en thèse pour (Kaisser et Lowe, 2008)) ou leur faire subir un post-traitement assez lourd (Xu et Klakow, 2010). Enfin, la qualité du travail des annotateurs non-experts varie considérablement (Tratz et Hovy, 2010).

Nous considérons, pour notre part, que la qualité des ressources produites est fondamentale et que, pour obtenir cette qualité, il faut mettre l'évaluation et l'annotateur au centre du processus d'annotation. Notre travail ne se situe donc clairement pas dans la même optique que MTurk.

3.4.3 Méthodologies partielles

Corpus

De nombreux travaux ont été menés en linguistique de corpus sur la sélection de corpus, en particulier par John Sinclair, dont l'article (Sinclair, 2005) résume les bases de la création de corpus (représentativité, échantillonnage, équilibre, etc).

Cela étant, la réalité d'une campagne pour le TAL est souvent plus triviale. D'une part, la sélection du corpus est souvent réalisée par le client (pour entraîner son outil) et celui-ci prend rarement en considération les connaissances acquises en linguistique de corpus. D'autre part, les corpus sélectionnés le sont souvent avant tout pour des raisons de disponibilité, parfois au détriment de leur intérêt. Ainsi, en microbiologie, on utilise principalement les résumés (*abstracts*) de MEDLINE, qui sont disponibles, plutôt que les articles entiers, qui pourtant apportent souvent des informations plus pertinentes (voir section 5.1), mais ne sont pas disponibles.

Acteurs

Les acteurs ne sont que peu ou pas décrits dans la littérature concernant les campagnes d'annotation. On y trouve au mieux un nombre d'annotateurs associé à un statut : expert, senior, junior, etc (Böhmová *et al.*, 2001; Abeillé *et al.*, 2003; Kim *et al.*, 2008; Alex *et al.*, 2010). Généralement, le gestionnaire de la campagne n'est pas cité, si ce n'est pour dire qu'il a participé à l'annotation en tant qu'expert (Laignelet et Rioult, 2009), sans plus de détails. Une rare exception à cette règle est (Alex *et al.*, 2010), dans lequel le gestionnaire est identifié nommément. Enfin, le client n'apparaît que lorsque l'application est finale, et encore, uniquement en filigrane (Laignelet et Rioult, 2009). Cet état de fait n'est pas surprenant, les campagnes étant en général très peu détaillées. Les articles cités ici sont d'ailleurs de ce point de vue des exceptions.

Certains acteurs apparaissent toutefois dans la description de deux outils de gestion d'annotation. Ainsi, `Slate` prévoit deux types d'utilisateurs (Kaplan *et al.*, 2010),

l'administrateur et les annotateurs, que nous détaillons en annexe A.4.1. Dans `Slate`, les deux rôles sont totalement étanches et l'administrateur ne peut donc pas annoter. `GATE Teamware` (Bontcheva *et al.*, 2010) prévoit lui trois acteurs : les gestionnaires de la campagne, les éditeurs ou curateurs (annotateurs experts) et les annotateurs. Là encore, le détail de leur rôle est présenté en annexe A.4.3.

Cependant, sorties du contexte précis de la campagne, très riche en informations, qui permettrait, entre autres, de mettre au jour d'éventuels biais, liés à la fusion de certains rôles, ces descriptions d'acteurs ne sont pas très informatives et il faut utiliser les outils pour en découvrir toutes les subtilités.

Formation

Il a été démontré dans (Dandapat *et al.*, 2009; Bayerl et Paul, 2011) qu'une bonne formation des annotateurs reste le moyen le plus efficace pour réduire le temps d'annotation et en améliorer la qualité. Cette phase de formation est inévitable, mais elle est souvent sous-estimée, en temps et en importance pour la campagne. Ainsi, peu d'articles donnent le temps d'apprentissage des annotateurs. De ce point de vue, comme de beaucoup d'autres, (Marcus *et al.*, 1993) fait exception, en annonçant les temps d'apprentissage des annotateurs (un mois pour l'annotation morpho-syntaxique, deux pour l'annotation syntaxique).

Nous imaginons que cette phase d'apprentissage, si elle n'est pas formellement établie et documentée, est au moins reconnue en interne et exprimée par les annotateurs. Mais sans données fiables sur le temps effectif d'annotation et la qualité obtenue (sous forme d'accord inter-annotateurs ou de précision par rapport à une référence), comment être sûr que la phase d'apprentissage a été dépassée et que la phase de formation est terminée ?

Dans le jeu `Phrase Detectives` (Chamberlain *et al.*, 2008b), la phase de formation est obligatoire et le joueur n'a le droit de véritablement commencer à annoter que lorsque ses résultats sont jugés suffisants. Par la suite, il doit régulièrement annoter un texte extrait du *gold standard* et ses résultats sont enregistrés, permettant ainsi un tri des « mauvais annotateurs » *a posteriori*. Cette validation de la formation et cette évaluation tout au long de l'annotation nous semblent être de bonnes pratiques à mettre en œuvre dans toutes les campagnes d'annotation.

Documentation

Nédellec *et al.* (2006) montrent que la disponibilité d'un guide d'annotation permet d'améliorer les résultats de l'annotation automatique d'entités nommées en microbiologie, entraînée à partir du corpus annoté avec ce guide. Les corpus annotés sans guide

d'annotation présentent effectivement de graves incohérences, qui impactent les outils entraînés sur ces corpus (Alex *et al.*, 2006).

En effet, une formation suffisante, associée à une documentation présentant les types ambigus en parallèle avec les tests nécessaires, comme pour le *Penn Treebank* (Marcus *et al.*, 1993), permet aux annotateurs de mieux identifier les éléments à annoter et de désambiguïser les types, donc de produire une annotation plus cohérente.

Pour Sampson (2000), non seulement cet effort de documentation est indispensable, mais il devrait être considéré avec au moins autant d'intérêt que le développement d'outils. L'auteur utilise pour sa démonstration un parallèle avec l'évolution du développement informatique, depuis la programmation (écriture de code) vers l'ingénierie logicielle, qui implique des spécifications détaillées, à partir desquelles la programmation devient quasiment triviale.

Annotation agile

Ce parallèle avec les méthodes de développement a été étendu à la méthodologie même d'annotation, avec l'apparition récente de l'« annotation agile », par analogie avec le développement agile (Beck, 2011).

De notre point de vue, Bonneau-Maynard *et al.* (2005), même s'ils ne sont en général pas cités comme référence de l'annotation agile, sont des précurseurs en la matière. Ils ont en effet montré que des calculs d'accord inter-annotateurs dès le début de la campagne permettent d'identifier les problèmes rapidement et de mettre à jour le guide d'annotation, afin de limiter leur impact .

L'annotation agile (Voormann et Gut, 2008; Alex *et al.*, 2010) va un pas plus loin, en réorganisant également les phases traditionnelles de l'annotation manuelle (voir figure 3.2) et en préconisant plusieurs aller-retours entre l'annotation et la mise à jour du guide.

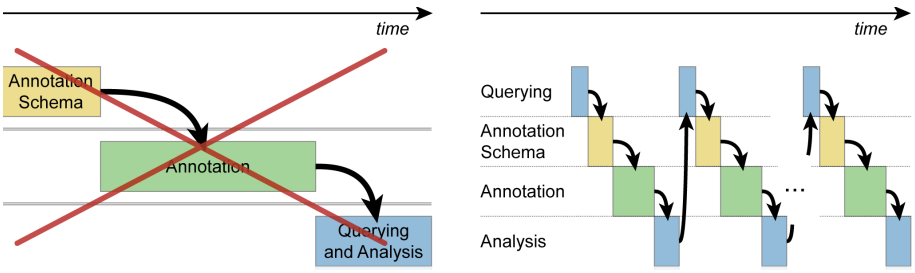


FIGURE 3.2: Phases de l'annotation traditionnelle (à gauche) et cycles de l'annotation agile (à droite). Reproduction de la figure 2 de (Voormann et Gut, 2008)

Cette méthode d'annotation est tout à fait récente et n'a fait l'objet pour l'instant que d'une seule application réelle, dans (Alex *et al.*, 2010). Il nous est donc difficile de dire dans quelle mesure elle est vraiment différente de la méthode préconisée par Bonneau-Maynard *et al.* (2005) et si elle permet de produire de meilleurs résultats.

3.5 Problématique

Comme nous venons de le voir dans les sections précédentes, si des pistes sont explorées sur des sujets en particulier, la gestion de l'annotation n'est pas réellement considérée dans son ensemble. Cette méta-annotation n'est pas pensée comme telle et il en découle de nombreux problèmes de coût et de qualité de l'annotation produite.

Selon nous, la solution à ces questions doit être une méthodologie d'ensemble, prenant en compte tous les aspects d'une campagne d'annotation et en priorité analysant de manière détaillée la campagne elle-même, dès ses débuts. Le gestionnaire de la campagne est la clé de voute de cette méthodologie. Il doit faire en sorte que les annotateurs puissent produire des annotations de qualité, en fonction de l'application visée et de manière la plus efficace possible.

Les annotateurs eux-mêmes, bien que « simples » exécutants, sont la source de l'interprétation recherchée et producteurs de la grande majorité du travail d'annotation. Nous pensons qu'il est important de mieux les prendre en compte, en particulier dans le contexte actuel où l'on tend à considérer (à tort) qu'un grand nombre de non-experts produit la même qualité d'interprétation qu'un petit nombre d'experts.

Deuxième partie

Méthodologie proposée

Chapitre 4

Organiser une campagne d'annotation

Comme nous allons le voir plus loin (section 4.1.1), le gestionnaire est la clef de voute d'une campagne d'annotation. Il est donc indispensable qu'il soit sensibilisé aux difficultés propres à une campagne d'annotation. Or, il n'existe aujourd'hui aucun guide du gestionnaire de la campagne. On trouve tout au plus quelques « bonnes pratiques » (Wynne, 2005) ou des ouvrages plus descriptifs sur l'annotation de corpus, peu adaptés à la tâche du gestionnaire de la campagne (Garside *et al.*, 1997).

Nous détaillons dans ce chapitre l'organisation d'une campagne d'annotation, qu'elle soit de type traditionnel ou par collaboration en ligne (*crowdsourcing*), que nous décomposons en trois phases principales, précédées d'un travail préparatoire.

4.1 Travail préparatoire

Une campagne d'annotation ne commence pas avec l'annotation elle-même, mais comprend un important travail d'identification des acteurs, de connaissance du corpus et d'écriture d'une première version du guide d'annotation. Ce travail préparatoire peut être long mais ne doit pas être négligé, parce qu'il assure la productivité et la qualité de la campagne d'annotation.

4.1.1 Identifier les acteurs

Cette sous-section correspond à l'approfondissement d'un travail de réflexion débuté en collaboration avec Sophie Rosset (LIMSI-CNRS). L'objectif est ici d'identifier clairement les différents intervenants d'une campagne d'annotation et de montrer les tensions produites par leurs visions – souvent divergentes – de la campagne, tensions qui peuvent aller jusqu'à provoquer l'échec de celle-ci.

Des rôles variés

Notre expérience et l'état de l'art nous amènent à distinguer cinq rôles principaux au sein d'une campagne d'annotation :

1. le ou les financier(s) : représentant(s), plus ou moins distants, des instances finançant la campagne d'annotation (ANR, OSEO, etc) et éventuellement son évaluation ;
2. le(s) client(s) ou donneur(s) d'ordre : personne(s) ou équipe(s) ayant besoin du corpus annoté pour entraîner, créer ou évaluer leurs outils ;
3. le gestionnaire de la campagne : personne chargée de mettre en place et de s'assurer de la bonne marche de la campagne (méthodologie, avancement). Généralement, le gestionnaire assure le lien entre le client, les évaluateurs et les experts (plus rarement avec les financiers) ;
4. le ou les annotateur(s) expert(s) : annotateurs spécialisés dans le domaine (parfois, dans la tâche) qui sélectionnent les annotateurs, les forment, les évaluent, répondent à leur questions et font l'adjudication si nécessaire ;
5. les annotateur(s) : personnes réalisant l'annotation ; dans le *crowdsourcing*, elles sont appelées « non experts », mais elles peuvent être des expertes du domaine (notamment dans les domaines très pointus) ;
6. le ou les évaluateur(s) : personnes chargées d'évaluer la qualité du corpus annoté ou/et les outils qui vont être créés, entraînés ou évalués grâce à ce corpus.

Ces rôles ne sont pas toujours tous présents dans une campagne d'annotation. Ainsi, les évaluateurs peuvent être absents d'une campagne purement interne, sans évaluation externe.

Quant au nombre d'acteurs par rôle, il est dans la plupart des cas variable, mais le gestionnaire doit être une personne unique, pour éviter les incohérences dans la campagne. Au cas où celui-ci est également un expert, il peut n'y avoir qu'un expert supplémentaire, mais ils doivent être au moins deux en tout si on veut construire une mini-référence (voir sous-section 4.2.1).

Des annotateurs en nombre... suffisant

Enfin, si l'on veut obtenir une annotation de qualité, il faut la faire réaliser par au moins deux annotateurs, afin de vérifier que leurs annotations convergent (voir section 3.3). Bayerl et Paul (2011) concluent leur méta étude en suggérant d'utiliser au moins cinq annotateurs pour les tâches les plus « critiques » et au moins trois ou quatre pour les autres. Avant eux, Klebanov et Beigman (2009) avaient montré qu'utiliser davantage d'annotateurs permettait de faire diminuer l'influence du hasard sur les accords inter-annotateurs. Pour autant, Bhardwaj *et al.* (2010) ont démontré que des annotateurs bien formés produisent une meilleure qualité d'annotation qu'une myriade de non-experts. Le nombre ne fait donc pas tout.

Dans le cas du *crowdsourcing*, les annotateurs sont supposément une myriade. Les participants sont en effet très nombreux. `Phrase Detectives`, par exemple, a rassemblé plus de 2 000 joueurs en 32 mois. Cependant, nous avons eu l'occasion de

collaborer avec les auteurs de ce jeu et nous avons constaté que seul un nombre limité des annotateurs apportent vraiment leur pierre à l'édifice, ce que nous détaillons dans (Chamberlain *et al.*, 2012). La figure 4.1 illustre le profil clairement zipfien de la courbe du nombre de joueurs en fonction de leur score (qui, dans le cas de `Phrase Detectives`, est dépendant non seulement du nombre d'annotations effectuées, mais également de leur qualité, voir annexe A.2.4). Les données utilisées pour créer cette courbe ont été récoltées entre février 2011 et février 2012 (inclus). Les 93 joueurs représentés sont les seuls à avoir passé la phase de formation, puis effectué une annotation complète (soit un joueur sur six sur la période). Le même phénomène a été observé par Mathieu Lafourcade dans `JeuxDeMots` (communication personnelle, le 16 septembre 2012), les résultats sont d'ailleurs accessibles à tous sur le site du jeu [1].

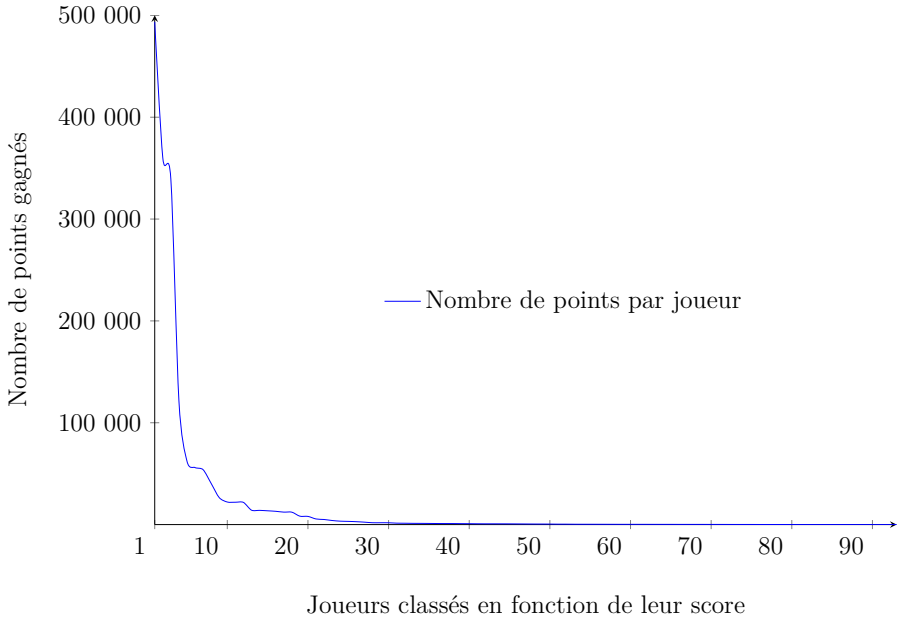


FIGURE 4.1: Nombre de joueurs sur `Phrase Detectives` en fonction de leur classement en points

Quant au nombre de *Turkers* actifs sur MTurk, si le nombre de personnes enregistrées sur le site comme *Turkers* est de plus 500 000, nous avons estimé les actifs entre 15 059 et 42 912 dans (Fort *et al.*, 2011a). Nous relevons en outre que 80 % des tâches (HIT) sont réalisées par les 20 % de *Turkers* les plus actifs (Deneme, 2009). La myriade participe, mais sa participation est à nuancer.

1. Il faut pour cela ouvrir une session sur `JeuxDeMots`, puis aller à http://www.jeuxdemots.org/generateRanking-4.php?crit=nbplayed

Des visions de l'annotation parfois divergentes

Chaque type d'acteur de la campagne la voit sous un angle particulier, et ces visions peuvent être dissonantes. Nous avons donc essayé d'identifier, pour chaque rôle, le point de vue dominant, c'est-à-dire ses priorités dans le cadre de la campagne d'annotation.

Ainsi, le financier vise un rapport qualité prix optimal, parfois au détriment de la qualité. Il peut cependant être sensible au fait que le corpus soit ré-utilisable (donc de qualité). L'évaluateur est lui peu préoccupé par le coût ou la vitesse d'annotation, mais il recherche la cohérence et la fidélité à la documentation. Le client est également peu préoccupé par le coût, mais est impatient de pouvoir tester ses outils ou de commencer son développement. La qualité est liée pour lui à son besoin, qui peut être une application finale ou un outil à entraîner. Le gestionnaire de la campagne a pour mission de faire réaliser l'annotation dans les délais prévus et d'assurer une qualité suffisante de celle-ci, définie avec le client, en fonction de l'application. Il assure le lien avec les experts et les annotateurs. Il est donc le seul à avoir une vision d'ensemble de la campagne, dont l'équilibre repose largement sur lui. L'annotateur expert (ou expert) est en général peu préoccupé par le coût, mais le domaine traité est le sien et lui tient à cœur, la qualité de l'annotation est donc importante pour lui. Il connaît souvent mal l'application finale. Enfin, l'annotateur recherche en général la qualité, surtout s'il connaît l'objectif final de l'annotation, ce qui est plus rarement le cas dans le *crowdsourcing*. Il est par contre a priori peu préoccupé par la vitesse, à moins qu'un effort en ce sens lui soit demandé.

En outre, nous sommes en présence de différents rapports de force, il faut donc ajouter à ces objectifs divergents les tensions qui s'exercent entre les acteurs, en particulier sur les annotateurs (experts ou non), qui sont souvent de simples exécutants (voir figure 4.2). Il est à noter que les annotateurs sur `Amazon Mechanical Turk` sont particulièrement en position de faiblesse, puisque le *Requester* n'est pas obligé de les payer [2]. En revanche, dans les jeux ayant un but comme `Phrase Detectives`, le rapport de force entre les annotateurs et le gestionnaire est en quelque sorte inversé, puisque celui-ci cherche à attirer ceux-là pour qu'ils jouent, et, ce faisant, produisent des annotations.

Fusion des rôles et conséquences

Les fusions concernent tous les rôles et peuvent provoquer des conséquences portant atteinte à la campagne d'annotation de différentes manières.

Nous avons ainsi pu constater que le fait que l'INRA MIG soit à la fois client, expert et, au début, gestionnaire de fait de la campagne 1 (voir section 5.1) a provoqué plusieurs problèmes dans la campagne, qui sont probablement à l'origine de son échec.

2. Voir la page d'accueil du système https://www.mturk.com/.

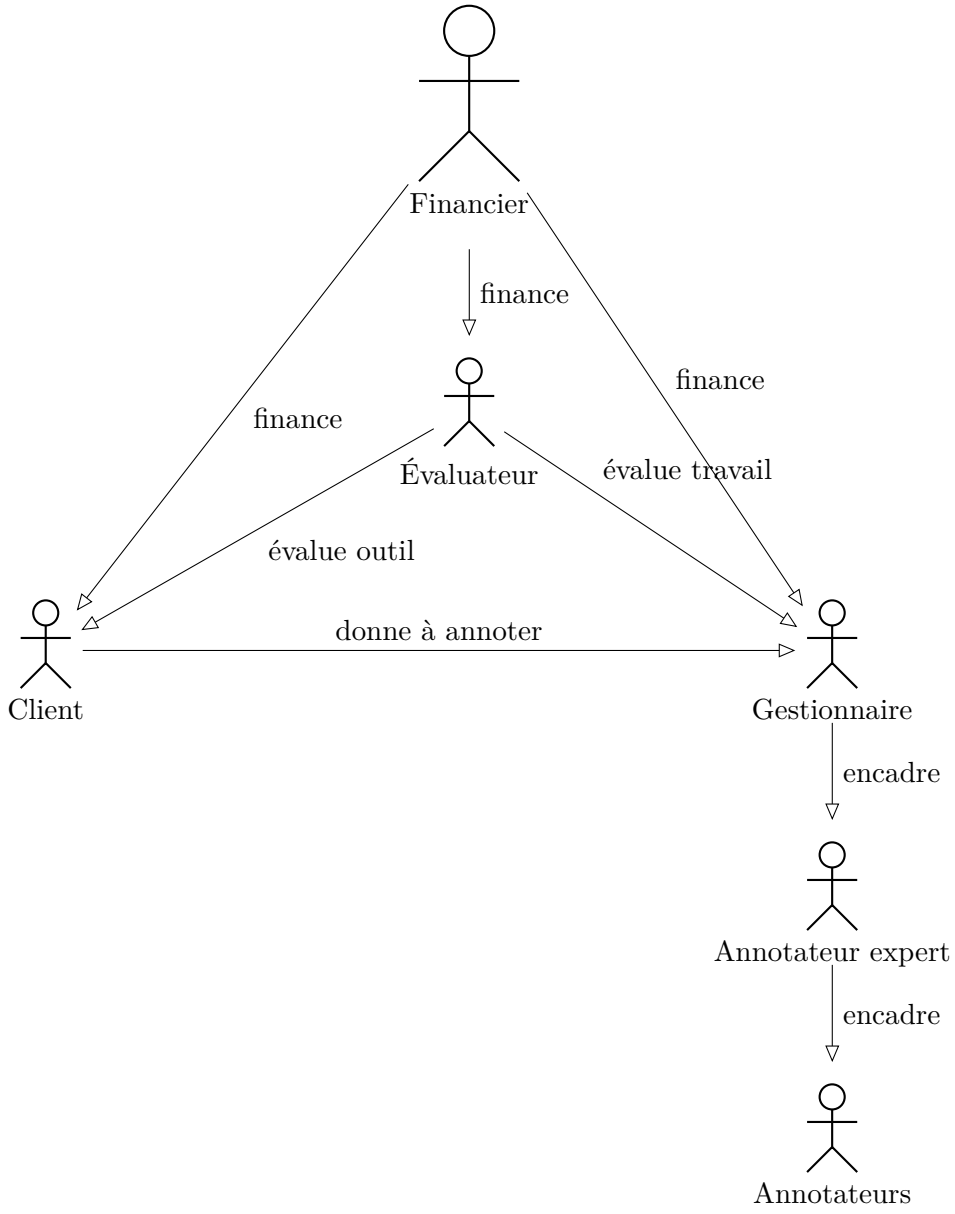


FIGURE 4.2: Hiérarchie des acteurs de l'annotation

L'absence d'un expert ou d'un gestionnaire distinct du client, qui aurait pu faire le lien avec les annotateurs les a isolés et ils n'ont pas réussi à faire remonter leurs difficultés d'annotation auprès du gestionnaire/expert. La distance physique entre les annotateurs et le gestionnaire n'a bien sûr rien arrangé. En outre, MIG, en tant que client, avait en vue les performances de son outil plutôt que l'application finale ou le domaine. La logique de l'annotation s'en est trouvée brouillée avec, par exemple, le mélange dans le guide d'annotation entre préoccupations linguistiques (*nom propre* ou

nom commun) et du domaine (*taxon* ou *eucaryote*), peu cohérentes entre elles, ce que nous avons montré dans (Fort *et al.*, 2009).

Bien entendu, on peut avoir une confusion des rôles encore plus grande. Dans ce cas, les risques sont évidemment multipliés. Ainsi, dans ESTER2, la DGA était à la fois annotateur, financier, évaluateur et gestionnaire, et l'équilibre de la campagne s'en est trouvé perturbé. Le guide d'annotation a été modifié après la distribution des données d'apprentissage et si les données pour le développement et le test des systèmes ont été ré-annotées partiellement, celles pour leur entrainement ne l'ont pas été. Il y a donc une différence nette entre l'annotation des deux jeux de données, remarquée par exemple dans (Raymond et Fayolle, 2010).

Enfin, notre absence de lien hiérarchique direct avec les annotateurs au sein de l'INIST-CNRS a posé problème dans la campagne 3 d'annotation de matchs de football, car, dans la mesure où cette tâche n'était pas leur tâche prioritaire, nous n'avons pas pu, en tant que gestionnaire, nous assurer de leur disponibilité pour permettre la continuité de l'annotation. La campagne s'en est trouvée rallongée et la qualité s'en est ressentie (les coupures ont nuit à la formation des annotateurs). Le même problème s'est d'ailleurs posé lors de la campagne 4 d'annotation de brevets en pharmacologie, que nous n'avons encadrée que partiellement.

Il n'est pas question ici de dire que tous les rôles doivent être parfaitement distincts et hiérarchisés dans chaque campagne, mais de mettre l'accent sur le fait qu'il faut identifier ces fusions et manques, afin de prendre conscience des biais éventuels qu'elles pourraient provoquer.

Ainsi, le financier, de par sa position de force, pourrait perturber toute la campagne, il ne doit donc pas intervenir directement sur les annotateurs ou sur le guide d'annotation. Enfin, un financier différent du client aide aussi à maintenir l'équilibre de la campagne. Sans financier, le client a beaucoup d'influence.

Le gestionnaire est lui garant de l'équilibre global, il ne devrait donc pas tenir d'autre rôle, sous peine de perturber la campagne de manière importante. S'il est également expert (ce qui est souvent le cas), il est préférable qu'il travaille avec d'autres experts pour compenser cet état de fait.

Enfin, l'expert doit non seulement encadrer les annotateurs, mais également se faire leur porte-parole. On voit que l'annotateur est tout en bas de l'échelle hiérarchique. Or, c'est son interprétation qui va donner sa valeur ajoutée au corpus. Il est donc indispensable d'assurer une bonne prise en compte de ses remarques et suggestions. Les annotateurs sur les plate-formes à la `Amazon Mechanical Turk` sont payés à la tâche et n'ont pas intérêt à perdre du temps en faisant remonter des problèmes au *Requester* (et la plate-forme ne les y encourage pas), leur point de vue est donc très peu pris en compte.

Les annotateurs, éléments clefs

Lors de toutes les campagnes d'annotation auxquelles nous avons participé, il est apparu que les annotateurs se comportaient de manières très différentes vis-à-vis de la tâche d'annotation. Bien entendu, cette « typologie » est loin d'être exhaustive ou constante, mais nous avons identifié deux types d'annotateurs.

Ainsi, un annotateur se distingue souvent des autres par ses questions, très pointues, et ses analyses et propositions, très utiles en début de campagne. Nous appelons cet annotateur « annotateur-analyste ».

Si les interventions de celui-ci sont indispensables, elles le paralysent souvent lors de l'annotation elle-même et il peut être moins efficace que d'autres, ou aller trop loin. Ainsi, un de nos annotateurs sur la campagne de football (voir section 5.3) s'est révélé capable de très bonnes analyses concernant le corpus et la campagne. Mais il est allé trop loin dans cette analyse, au point d'inférer des informations à partir de sources lointaines, plutôt que d'annoter le texte fourni. Il a donc fallu le recadrer rapidement. En revanche, l'autre annotateur qui participait à la campagne, s'il se montrait assez peu intéressé par l'analyse, a été très efficace pour l'annotation elle-même. Ce type d'annotateur est indispensable pour mener la campagne à son terme, nous le nommerons « annotateur-exécutant » (sans que le terme soit péjoratif).

Ces deux exemples montrent à quel point il est important d'identifier assez tôt les caractéristiques des annotateurs, sans pour autant les caricaturer à outrance, afin de valoriser au mieux ces différentes compétences.

Nous avons également parfois observé l'existence de biais « idéologiques ». Il est important de les détecter pour limiter leur impact sur l'annotation. Ainsi, dans une campagne d'annotation de matchs de football comme celle que nous avons menée, un supporter de telle ou telle équipe en particulier pourrait être tenté malgré lui d'annoter plus d'actions positives concernant son équipe fétiche. Il est très facile, dans un tel cas, de ne pas lui donner à annoter des matchs impliquant son équipe, ou de le prévenir contre d'éventuels biais. Certains cas sont plus difficiles à résoudre. Ainsi, un exemple issu de la campagne d'annotation d'entités nommées structurées (voir section 5.5) est l'annotation de l'expression « territoires palestiniens », qui peut être considérée comme une région, un pays ou une organisation administrative, selon le point de vue de l'annotateur.

Cette identification des annotateurs et de leurs biais éventuels ne peut être réalisée que dans le cas d'une annotation traditionnelle, car dans le cas de l'annotation par *crowdsourcing*, ceux-ci sont majoritairement des inconnus[3]. Dans ce cas, on compte sur le nombre important d'annotateurs pour neutraliser les effets de ce type de biais.

3. Certaines plate-formes de myriadisation du travail parcellisé éthiques, comme `Samasource`, permettent cependant la consultation du CV des travailleurs inscrits.

Enfin, il nous paraît important de nuancer une affirmation courante qui voudrait que les chercheurs eux-mêmes, la plupart du temps à la fois gestionnaires de la campagne et experts, font les meilleurs annotateurs. Notre expérience montre en effet que (i) même s'ils sont experts de la tâche, ils ne sont pas nécessairement experts du domaine et peuvent éprouver des difficultés à comprendre le contexte, comme dans le cas de l'annotation d'entités nommées en presse ancienne (voir sous-section 5.6.4), et (ii) qu'ils remettent trop facilement en cause le guide d'annotation qu'ils ont écrit ou ne s'y réfèrent pas assez. Ainsi, lors de la campagne 5a d'annotation des entités nommées structurées de nouvelles radio-diffusées, les quatre experts qui ont annoté la mini-référence (dont nous), ont obtenu entre eux des scores d'accords inter-annotateurs (Slot Error Rate et F-mesure) qui n'étaient pas meilleurs que ceux des annotateurs d'ELDA [4].

4.1.2 Prendre en compte le corpus

Nous reprenons ici les grandes lignes d'un travail réalisé en collaboration avec Adeline Nazarenko et Claire Ris (INIST-CNRS) (Fort *et al.*, 2011b). Dans cet article, nous nous appuyons sur la campagne 3 d'annotation de football (voir section 5.3) pour montrer que l'hétérogénéité du corpus affecte tous les aspects de la campagne : la sélection d'un sous-corpus pour la formation des annotateurs, la durée de cette formation, la complexité du schéma d'annotation et la qualité de l'annotation résultante. Le gestionnaire de la campagne, qui ne choisit pas forcément le corpus sur lequel l'annotation va porter, doit donc adapter la campagne à ces impératifs.

Cela suppose de bien connaître le corpus. Un bon moyen d'acquérir cette connaissance consiste, lorsque c'est possible (c'est-à-dire lorsque le domaine et la langue s'y prête), à annoter une partie, petite mais représentative, de ce corpus, avant même de commencer la campagne. C'est ce que nous avons fait pour les campagnes 3 et 5, et cela nous a permis non seulement de mettre au jour des problèmes dans le guide d'annotation avant même que les annotateurs commencent à travailler, mais également de créer une petite référence pour l'évaluation. Dans certains cas (comme dans notre campagne 3 d'annotation de football), faire annoter cette mini-référence par le client lui permet de valider les choix réalisés et de vérifier que l'annotation ne s'éloigne pas trop de l'application prévue. Dans les cas où le gestionnaire n'est pas un spécialiste du domaine, il doit donc s'adjoindre un ou plusieurs experts pour le travail préparatoire.

Qu'elle soit faite *a priori*, lors de la sélection du corpus, ou *a posteriori*, une fois le corpus sélectionné, une analyse précise de la composition de celui-ci et des conséquences de cette composition sur la campagne doit donc être réalisée le plus tôt possible.

4. Il faut noter cependant que les annotateurs d'ELDA avaient accès à tout le contexte des nouvelles, alors que nous les avions par phrases, décorrélées du contexte (pour limiter l'impact de la courbe d'apprentissage sur nos résultats). Si ce biais a certainement une influence sur nos accords avec les annotateurs d'ELDA, il est difficile d'évaluer son influence sur nos accords entre experts.

4.1.3 Créer et modifier le guide d'annotation

Nous l'avons vu dans l'état de l'art (voir section 3.4.3), le guide d'annotation est un élément indispensable dans une campagne d'annotation. Ainsi, pour la campagne 5a d'annotation en entités nommées structurées (voir section 5.5), la mise au point du guide d'annotation a nécessité six mois de travail. Le coût de ce travail préparatoire est élevé (d'autant qu'il concerne plusieurs chercheurs), mais le guide d'annotation produit a pu être ré-utilisé pour la campagne suivante (5b), ainsi que par le projet ANR ETAPE.

La rédaction de ce guide ne doit cependant pas être considérée comme une tâche à réaliser et finaliser au début de la campagne, tout au plus avec quelques modifications par la suite. Bien au contraire, le guide d'annotation évolue tout au long de la campagne d'annotation, jusqu'à l'utilisation effective du corpus par le client. C'est la condition *sine qua non* de son utilisabilité en tant que documentation du corpus annoté.

Une première version doit être écrite rapidement, avant la campagne, en collaboration avec le client, puis testée en annotant une mini-référence, si c'est possible. Une première vague de modifications a alors normalement lieu. Ensuite, la phase de rodage permet, grâce au retour des annotateurs, de continuer les améliorations. Ces améliorations doivent à leur tour permettre une meilleure qualité de l'annotation, ainsi qu'un gain de temps, les catégories mal définies ou mal comprises faisant perdre beaucoup de temps aux annotateurs. Plusieurs cycles annotation/révision du guide peuvent être nécessaires pour atteindre une certaine stabilité de celui-ci, qui se manifeste par une qualité constante et suffisante de l'annotation.

Il nous semble que l'annotation agile (voir section 3.4.3) se distingue de la méthodologie d'annotation préconisée par Bonneau-Maynard *et al.* (2005) par la continuation de ces cycles jusqu'au bout de la campagne. Or, dès lors que l'annotation est stabilisée en termes de qualité et de temps d'annotation, il nous semble peu utile de continuer formellement le processus, même si, bien entendu, d'autres évaluations doivent être faites pour s'assurer de la non-régression de l'annotation. Quoi qu'il en soit, les deux méthodes sont très proches dans leur logique sous-jacente.

Enfin, les catégories mal définies ou mal comprises sont génératrices de stress et d'erreurs. Afin de les alléger de ce stress perturbateur et de garder une trace précise des problèmes qu'ils rencontrent lors de cette phase, il est important d'offrir aux annotateurs la possibilité d'ajouter un trait d'incertitude lorsqu'ils ont des doutes sur leurs décisions. Ce trait d'incertitude peut être typé, afin de pouvoir être plus facilement utilisé par la suite et doit bien entendu être décrit dans le guide d'annotation.

Nous donnons un certain nombre de recommandations concernant l'écriture d'un guide d'annotation dans (Fort *et al.*, 2009). Nous les résumons ici :

– indiquer ce qu'il faut annoter plutôt que comment annoter,

– ne pas exclure *a priori* ce qui serait douteux ou trop difficile à reproduire par un outil automatique,
– donner aux annotateurs une vision claire de l'application visée,
– ajouter des définitions précises, justifier les choix méthodologiques effectués et donner la logique de l'annotation visée (ne pas se contenter de donner des exemples).

Le fait de suivre ces recommandations devrait permettre de responsabiliser et de motiver les annotateurs en leur donnant accès à la logique sous-jacente. On passe ainsi d'un « rapport de père à un rapport de pair » (Akrich et Boullier, 1991), ce qui est important pour la qualité de l'annotation et d'autant plus nécessaire qu'elle porte sur des corpus spécialisés, les annotateurs étant des experts à qui on a intérêt à donner la plus grande autonomie possible.

Il est donc fondamental de ne pas tout décrire, de laisser une marge d'interprétation suffisante aux annotateurs pour qu'ils puissent apporter une vraie valeur ajoutée au corpus. Un guide trop détaillé et trop long à consulter sera moins utile aux annotateurs qu'un guide présentant l'essentiel, agrémenté de quelques exemples bien choisis et de tests concrets pour distinguer les catégories connues pour être ambiguës. Celui du Penn Treebank pour l'annotation morpho-syntaxique nous semble être un bon exemple de ce point de vue.

Le *crowdsourcing* pousse la logique minimaliste à son extrême, puisque le guide d'annotation y est réduit à des directives de quelques lignes (sur `Amazon Mechanical Turk`), à quelques pages (pour le jeu `Phrase Detectives`). La tâche à effectuer doit donc rester simple.

Le travail préparatoire a normalement permis de définir précisément l'application visée, d'écrire un premier jet du guide d'annotation, de découvrir le corpus et d'identifier les acteurs impliqués. La campagne d'annotation peut alors commencer. Elle comprend trois phases principales : la pré-campagne, qui voit la mise au point d'une mini-référence et la formation des annotateurs, l'annotation elle-même, qui commence par un rodage et est émaillée d'évaluations et de mises à jour, et la finalisation, qui comprend une correction, manuelle ou non, du corpus annoté, avant sa publication. L'organisation générale d'une campagne d'annotation est présentée dans la figure 4.3.

4.2 Pré-campagne

La pré-campagne implique principalement le gestionnaire de la campagne. Il est généralement associé à des experts dans la construction d'un échantillon de corpus annoté qui servira de mini-référence et il encadre les annotateurs lors de leur formation. Il reste cependant l'acteur clef de cette phase fondamentale de la campagne d'annotation, qui voit la finalisation du guide d'annotation et de la mini-référence.

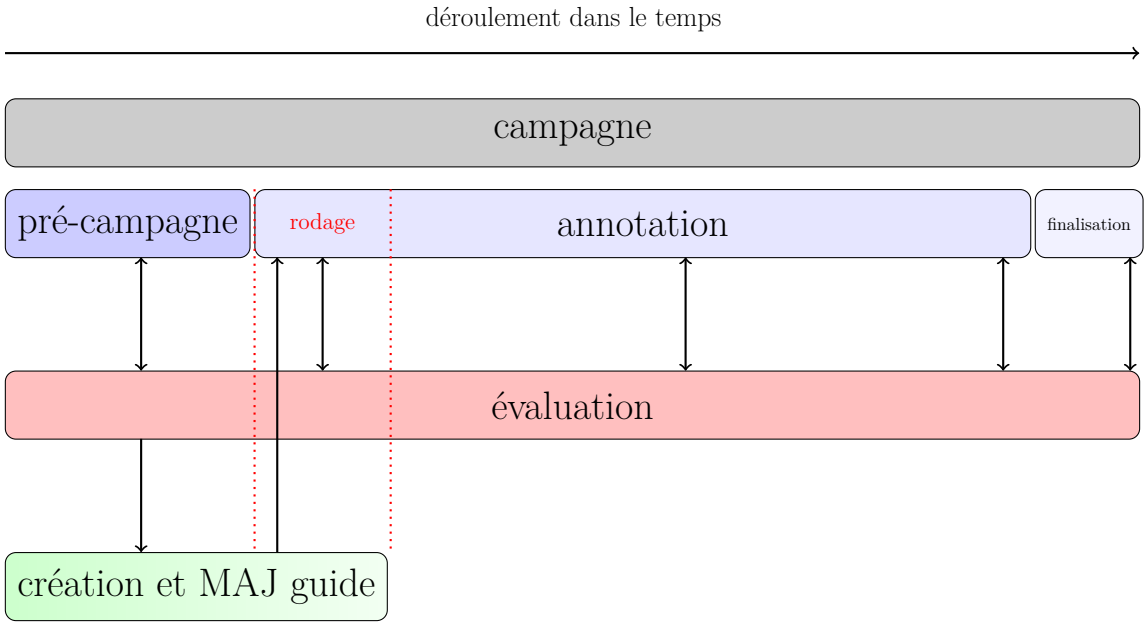


FIGURE 4.3: Organisation générale d'une campagne d'annotation.

4.2.1 Mise au point d'une mini-référence

Construire une mini-référence dès le début la campagne présente de nombreux avantages. Cela permet d'abord de tester en conditions réelles le premier jet du guide d'annotation, provenant par exemple de discussions entre le client et le gestionnaire. La construction de la mini-référence permet également d'obtenir une mesure d'évaluation de la fiabilité de l'annotation dès le début de la campagne. Cette mesure pourra par la suite être comparée avec d'autres, effectuées plus tard. En outre, la mini-référence, une fois finalisée, contient les informations nécessaires au calcul des dimensions de complexité de la campagne (voir chapitre 6), qui donneront des indications fines pour sélectionner les outils appropriés, qu'il s'agisse de pré-annotation automatique (voir chapitre 7), d'outils d'annotation (voir section 3.2) ou d'autres solutions (voir section 3.4). Enfin, cette étape permet la sélection d'une mesure d'évaluation de l'accord inter-annotateurs appropriée à la campagne (voir chapitre 8).

Les différentes étapes de la création de la mini-référence sont décrites dans la figure 4.4. Le sous-corpus de référence (ou mini-référence) est un échantillon du corpus d'origine non annoté, sélectionné si possible de manière à être représentatif. Le travail préparatoire (voir section 4.1.2) a normalement permis d'établir une typologie détaillée du corpus, et la création d'un sous-corpus représentatif pour la mini-référence peut donc en principe se faire en sélectionnant des fichiers (ou parties de fichiers) correspondant à chaque type identifié, de manière proportionnée. Le but n'est pas ici d'être parfaitement représentatif, mais de couvrir suffisamment de phénomènes pour mettre au

jour un maximum de problèmes durant l'annotation de la mini-référence. La taille de ce sous-corpus va surtout dépendre du temps disponible pour annoter, mais une taille trop réduite ou une représentativité insuffisante peuvent être source d'erreurs importantes dans le calcul des dimensions de complexité de la campagne. Ainsi, nous avons constaté, en effectuant le calcul des dimensions de complexité de la campagne d'annotation d'entités nommées structurées (voir section 5.5), que l'échantillon choisi pour le calcul était de trop petite taille. L'ambiguïté théorique est en effet relativement faible sur la mini-référence (entre $0,12$ et $0,15$, selon la tâche élémentaire considérée) et bien plus élevé sur le corpus total ($0,49$ et $0,36$, respectivement). Ces résultats sont présentés en détail dans (Fort *et al.*, 2012b).

Cette mini-référence est annotée par le gestionnaire de la campagne (ou un expert si le domaine traité est inconnu du gestionnaire), associé à au moins un ou des experts. L'annotation, entrecoupée de réunions informelles, doit permettre l'amélioration progressive du guide d'annotation et du jeu d'étiquettes. Nous avons créé des mini-références dans le cadre des campagnes d'annotation de football (voir section 5.3) et d'entités nommées structurées (voir section 5.5). Dans les deux cas, ces mini-références ont été finalisées tardivement, mais elles ont été utilisées pour l'évaluation.

Il n'est pas rare que de telles mini-références soient créées dans le cadre du *crowdsourcing*, afin de valider le travail des annotateurs. Ainsi, dans `Phrase Detectives` (Chamberlain *et al.*, 2008b), un corpus de référence annoté par des experts permet de former les annotateurs puis de les évaluer tout au long du jeu.

Il est important de noter qu'il s'agit d'une « mini-campagne » dans la campagne, et que par conséquent les étapes décrites dans les sections 4.3 et 4.4 s'appliquent en théorie elles aussi. En pratique, seuls le rodage et la publication n'ont pas lieu lors de cette mini-campagne.

4.2.2 Formation des annotateurs

La formation des annotateurs est reconnue comme une étape fondamentale d'une campagne d'annotation (voir notamment (Dandapat *et al.*, 2009; Bayerl et Paul, 2011)). Il est par conséquent important d'y passer le temps nécessaire.

Nous présentons cette étape dans la figure 4.5. Nous confondons dans ce schéma formation à la tâche d'annotation et à l'outil d'annotation, car elles sont de fait souvent associées. Ces deux formations posent cependant des problèmes distincts. Pour les annotateurs qui sont peu à l'aise avec les outils informatiques mais très compétents dans leur domaine d'expertise et dans la tâche d'annotation, il importe de trouver l'outil qui leur sera le plus adapté, quitte à perdre un peu en efficacité (par exemple, en utilisant

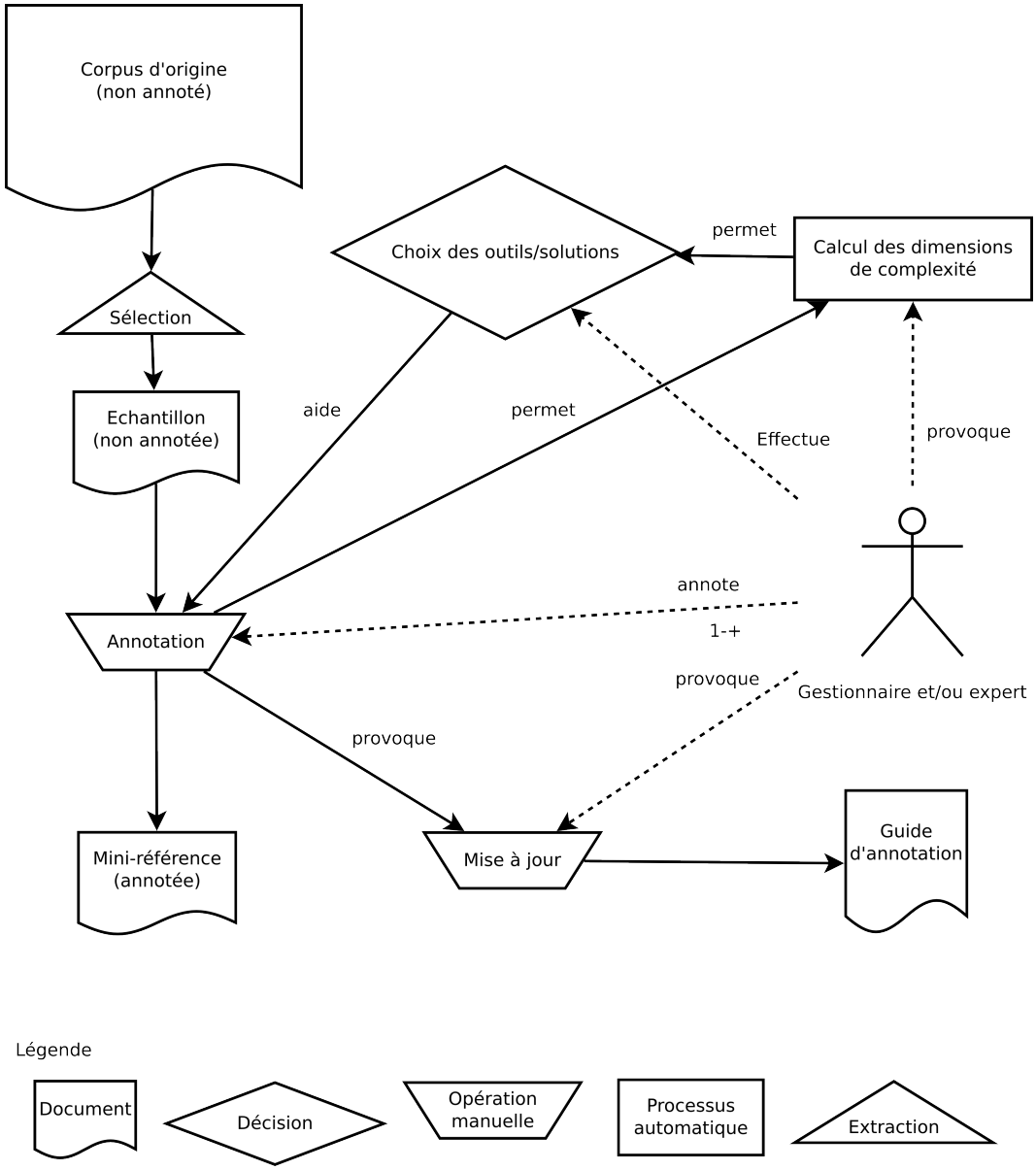


FIGURE 4.4: Création de la mini-référence

un outil « tout souris [5] » comme Glozz). La prise en main de l'outil peut en effet se révéler plus longue que prévue pour ces annotateurs.

La phase de formation permet également de repérer les annotateurs ayant des difficultés importantes sur la tâche et de les sortir de la campagne. Nous avons fait l'expérience d'un tel cas dans le cadre de la campagne 3 d'annotation de matchs de football : les validations réalisées en fin de formation (comparaison des annotations

5. En anglais, "point and click"

des annotateurs entre eux et avec le gestionnaire de la campagne) nous ont permis de constater les difficultés d'adaptation d'un annotateur à la tâche et de mettre fin à sa participation à la campagne.

La formation s'articule autour d'un extrait de la mini-référence, qui doit être annoté par les annotateurs, en fonction du guide d'annotation qui leur est fourni. Une première phase de formation en présence des différents acteurs, avec possibilité de poser les questions au moment où elles apparaissent, est toujours plus profitable qu'une formation à distance, mais n'est pas toujours possible.

Cette phase initiale doit être suivie d'une autre pendant laquelle les annotateurs travaillent en parallèle, sans se consulter, sur un même sous-corpus, en notant précisément leur temps d'annotation. L'évolution du temps d'annotation doit permettre de visualiser la courbe d'apprentissage des annotateurs, comme dans la figure 7.1 du chapitre 7. Cette courbe est un premier indicateur du niveau de formation des annotateurs. Le second indicateur est la stabilisation de la qualité produite.

L'évaluation de la formation peut se faire sur la mini-référence seule, comme indiqué sur la figure 4.5, mais également entre annotateurs (un accord inter-annotateur doit alors être calculé). Une mise au point avec les annotateurs doit ensuite avoir lieu, afin de revenir sur les points difficiles.

Cette phase de la campagne peut mettre au jour des erreurs ou des imprécisions dans le guide d'annotation et donc entraîner une mise à jour de celui-ci, ainsi que de la mini-référence. Au final, une deuxième passe d'annotation suivie d'une validation peut être utile, afin de valider ou invalider les annotations réalisées et donc le niveau des annotateurs.

Dans le jeu `Phrase Detectives` (Chamberlain *et al.*, 2008b), la phase de formation est automatisée (des indications sont fournies à l'annotateur-joueur pour l'aider à se former) et ne se termine que lorsque le joueur fait suffisamment peu d'erreurs (moins de 50 %). À l'opposé, dans les plate-formes de myriadisation du travail parcellisé comme `Amazon Mechanical Turk` (voir annexe A.1.4), les annotateurs peuvent au mieux subir un test de compétences avant de commencer à travailler, mais aucune phase de formation n'est prévu par le système. Formation et *crowdsourcing* ne sont donc pas fondamentalement antinomiques, mais les associer nous semble remettre en cause ce que beaucoup considèrent comme l'un des principes fondamentaux du système, l'utilisation de « non experts ». En effet, former un « non-expert » ne revient-il pas à en faire un expert, au moins de la tâche ?

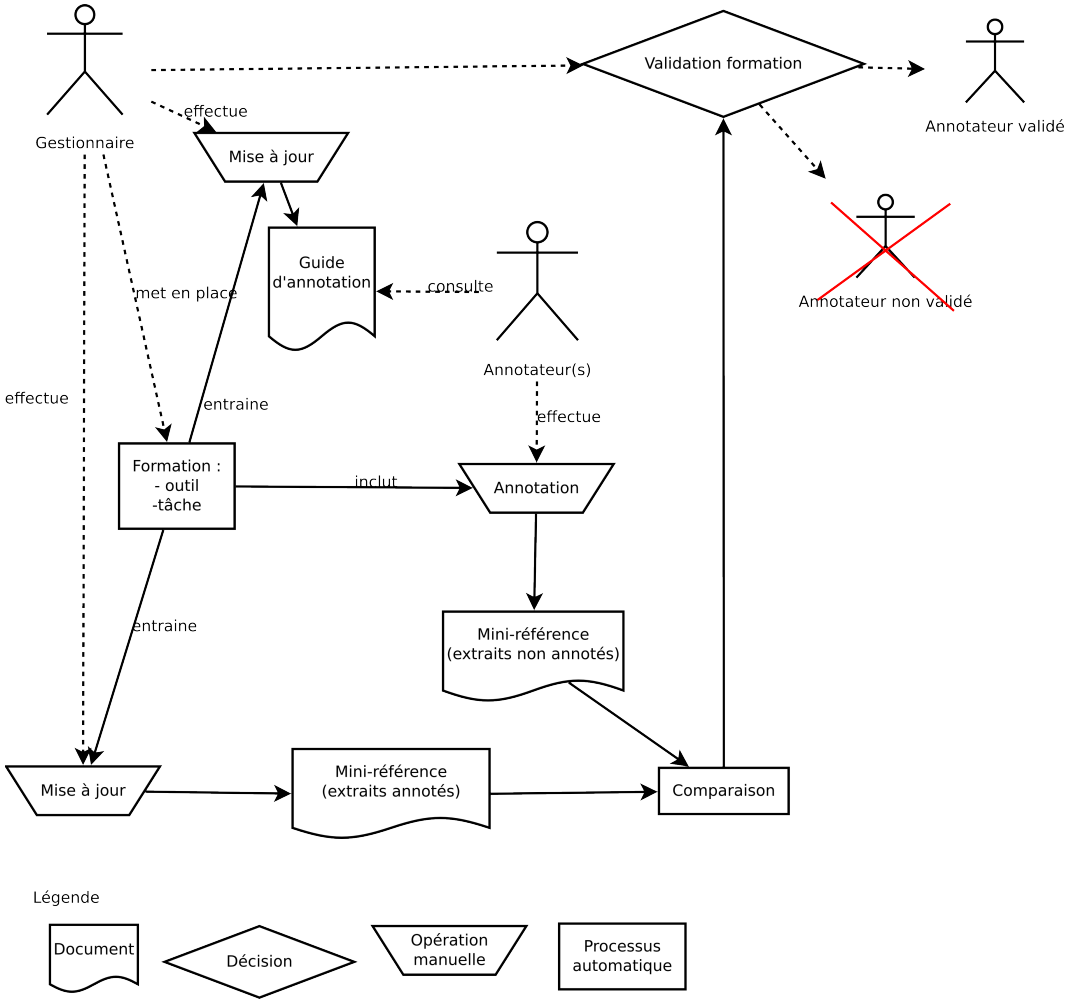


FIGURE 4.5: Formation des annotateurs

4.3 Annotation

4.3.1 Rodage

La fin de la pré-campagne ne signifie évidemment pas que tout est définitivement stabilisé. D'une part l'apprentissage des annotateurs va se poursuivre, car ils sont rarement dès la fin de la formation au maximum de leurs possibilités en termes de qualité produite et de rapidité d'annotation (la phase d'apprentissage a nécessité un mois pour la partie morpho-syntaxique du *Penn Treebank*). D'autre part, le guide d'annotation va continuer à être modifié, suite aux remarques des annotateurs. Il peut donc être nécessaire de corriger ultérieurement leurs annotations.

Une phase de « rodage » plus ou moins longue prolonge donc la pré-campagne. En

fonction des moyens disponibles, le gestionnaire va permettre des modifications du guide plus ou moins tardivement dans la campagne. L'idéal serait de pouvoir remettre en cause celui-ci jusqu'au bout, afin de prendre en compte tous les éléments découverts dans le corpus. En pratique, il faut stabiliser le guide pour pouvoir avancer dans l'annotation et ne pas perdre trop de temps à corriger ce qui a déjà été annoté. Un bon moment pour cela est sans doute celui où les annotateurs atteignent leur vitesse de croisière (on en trouve une illustration dans la figure 7.1).

Il faut noter que cette phase de rodage existe également dans le *crowdsourcing*, qu'il s'agisse de jeux ou de travail parcellisé. En effet, la mise en place du jeu ou de l'interface sur la plate-forme de travail parcellisé nécessite en général plusieurs essais avant de stabiliser les directives (« guide d'annotation » minimaliste), les interfaces (« outil d'aide à l'annotation ») et les conditions d'annotation (avec ou sans limite de temps, par exemple). Un bon exemple de ces itérations est présenté dans (Hong et Baker, 2011).

4.3.2 Travail d'annotation

La grande majorité du travail va être réalisée pendant cette phase d'annotation, par les annotateurs. Les étapes précédentes ont normalement permis de la préparer. Il est cependant important que le suivi des annotateurs, par l'expert et/ou le gestionnaire de la campagne, soit régulier.

Il est ainsi généralement nécessaire de calculer des accords inter-annotateurs, afin de s'assurer de la fiabilité de l'annotation. Une parallélisation au moins partielle de l'annotation doit donc être prévue. En fonction des moyens disponibles, l'annotation peut être réalisée en totalité en double, par au moins deux annotateurs, ou seulement partiellement en parallèle. Le premier cas de figure est relativement rare en annotation traditionnelle du fait de son coût élevé. Il est en revanche très courant quand il y a collaboration en ligne ou *crowdsourcing* (voir sous-section 3.4.2) : de nombreux participants (joueurs ou travailleurs) annotent les mêmes extraits.

La phase d'annotation elle-même peut comprendre une étape de pré-annotation automatique, le travail des annotateurs revenant alors à corriger cette annotation et éventuellement à la compléter. Dans ce cas de figure, il est important d'avertir les annotateurs/correcteurs des risques de biais dus à la pré-annotation (relâchement de l'attention et oublis) et de les notifier dans le guide. Cependant, qu'il y ait ou non pré-annotation automatique ne change pas fondamentalement la tâche, nous parlerons donc d'annotation y compris pour la correction de pré-annotations automatiques.

Chaque annotateur de la campagne s'est normalement vu attribuer un certain nombre de fichiers à annoter. Il a à sa disposition le guide d'annotation à jour, cohérent avec le modèle de données utilisé, et un outil d'aide à l'annotation. Il est formé à l'outil et à la tâche et doit avoir assimilé les grands principes énoncés dans le guide d'annotation.

La période de rodage doit lui permettre d'affiner sa connaissance du guide, qu'il doit faire sien rapidement.

L'idéal serait que le guide d'annotation soit directement intégré dans l'outil et que ce dernier soit capable de vérifier la conformité de l'annotation par rapport au guide[6]. Des fonctionnalités intermédiaires existent cependant, qui permettent d'ors et déjà une utilisation du guide plus efficace. La première est de rendre le guide facilement accessible, par un lien prévu dans l'outil, qui pourrait être mis à jour par le gestionnaire, de manière transparente pour l'annotateur. Une autre consiste à faire appliquer par l'outil des contraintes présentes dans le guide (comme cela a été fait avec `EasyRef` pour la campagne EASy, voir Annexe A.2.3). Le minimum est d'assurer la cohérence entre version du guide et modèle de données utilisé par l'outil (`Slate` prévoit ainsi un versionnage général de l'annotation, voir Annexe A.4.1).

L'outil d'aide à l'annotation utilisé par les annotateurs doit leur permettre non seulement d'annoter, mais également de suivre leur progression par rapport à l'ensemble des fichiers qui leur sont assignés, d'enregistrer le temps qu'ils passent sur chaque fichier (voire, plus finement, sur chaque couche d'annotation) et de signaler d'éventuels problèmes à l'expert et au gestionnaire. Il doit également offrir des fonctionnalités de recherche avancée (sur les catégories ou dans le texte) dans les fichiers attribués à chaque annotateur, afin qu'ils puissent corriger leurs annotations efficacement.

Durant la phase d'annotation, une évaluation régulière de la conformité de l'annotation par rapport à la mini-référence est réalisée, ainsi que des calculs réguliers d'accords intra- et inter-annotateurs.

4.3.3 Mises à jour

Même s'il est décidé de stabiliser le guide à la fin de la phase de rodage, des mises à jour sont inévitables pendant la pré-campagne et la phase de rodage. Ces mises à jour doivent donner lieu à des corrections dans le corpus déjà annoté, afin de conserver la cohérence entre les deux.

Pendant la pré-campagne, les mises à jour sont décidées de manière informelle, entre les experts. La mini-référence étant par définition de taille réduire, les corrections sont effectuées immédiatement.

Pendant le rodage, les mises à jour sont déclenchées soit formellement par le gestionnaire, suite à des mesures d'évaluation décevantes ou moins formellement par les annotateurs, qui les demandent à l'expert et au gestionnaire. Le gestionnaire peut décider de renoncer à certaines mises à jour pour des raisons de coût.

6. Mais on pourrait alors se demander où serait l'intérêt d'utiliser des annotateurs humains.

4.4 Finalisation du corpus annoté

4.4.1 Une finalisation par itérations

Une fois le corpus annoté, le gestionnaire doit finaliser la campagne. Il a à sa disposition les annotations réalisées et toute une série d'indicateurs, dont au minimum des mesures d'évaluation (conformité et accords inter- et intra-annotateur), auxquelles s'ajoutent parfois les traits d'incertitudes ajoutés par les annotateurs. Le gestionnaire peut également demander leur avis aux annotateurs en ce qui concerne leur ressenti sur la campagne. Il doit ensuite décider de la suite à donner et quatre options s'offrent alors à lui (voir figure 4.6) :

1. publier le corpus, considéré comme étant dans un état suffisamment satisfaisant pour être final,
2. réviser le corpus en adaptant le guide d'annotation,
3. faire réaliser une adjudication des annotations par un expert,
4. renoncer à la révision et à la publication (échec).

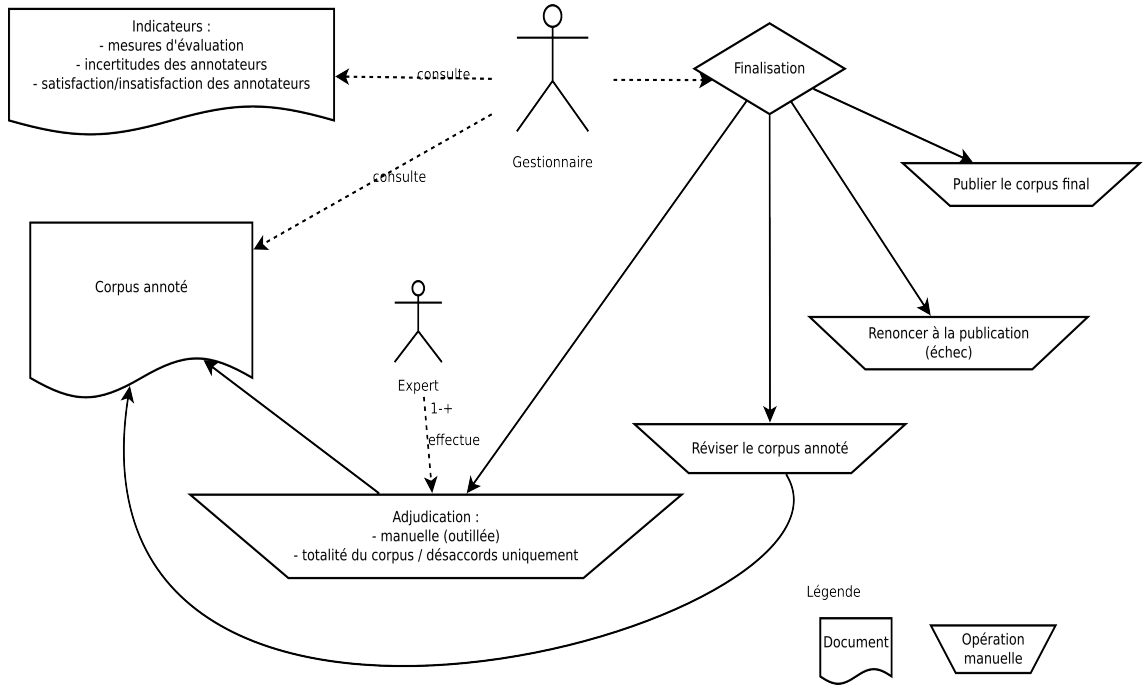


FIGURE 4.6: Finalisation du corpus annoté

Dans la grande majorité des cas, une phase de correction est nécessaire. Dans le cas d'une annotation réalisée totalement en parallèle par au moins deux annotateurs, cette

correction peut prendre la forme d'une adjudication par un expert, mais elle est plus souvent effectuée de manière plus ou moins automatique, à partir des indicateurs fournis lors de la campagne [7].

Dans les cas de correction (adjudication et révision), le corpus doit ensuite faire l'objet d'une évaluation et être de nouveau soumis, avec les indicateurs qui lui sont associés, à la décision du gestionnaire, qui peut soit publier le corpus, soit de nouveau le faire corriger (voir figure 4.6).

Échec Un échec qui ne serait constaté qu'au moment de la finalisation de la campagne est signe d'une absence de gestion de celle-ci et reste rare. Nous en avons cependant été témoin lors de la campagne 1 d'annotation en microbiologie (voir section 5.1).

Adjudication L'adjudication est la correction par un ou plusieurs experts des annotations réalisées par les annotateurs.

Cette correction porte en général uniquement sur les désaccords entre annotateurs (d'où son nom), mais nous en étendons la définition à la correction de toutes les annotations réalisée par un expert (cas rare dans l'annotation traditionnelle). Dans le premier cas, l'expert valide ou non l'une des annotations concurrentes. Un tri préalable des annotations doit donc être réalisé pour ne laisser à l'expert que les désaccords à trancher. Nous avons utilisé cette méthode lors de la campagne d'annotation du projet TCOF-POS (Benzitoun *et al.*, 2012).

L'intervention de l'expert peut aussi avoir lieu ponctuellement, sur une catégorie particulièrement difficile à annoter, par exemple.

Dans tous les cas, les corrections de l'expert peuvent être outillées, par exemple par le biais d'interfaces adaptées montrant en parallèle les annotations discordantes, à l'image de celle que nous avons réalisée pour la validation de lexique syntaxique (Fort et Guillaume, 2008).

Il nous semble important de noter ici que le travail parcellisé à la *Mechanical Turk* ne dispense pas nécessairement de corriger manuellement le travail réalisé. Ainsi, Kaisser et Lowe (2008) ont dû utiliser des étudiants en thèse pour valider leur corpus de questions/réponses. Quant à `Phrase Detectives`, les corrections y sont réalisées par les joueurs eux-mêmes, qui ont à juger les annotations réalisées par les autres.

7. Il est à noter que des corrections dans le corpus annoté peuvent également avoir lieu tout au long de la campagne d'annotation. L'outil d'annotation doit donc permettre une recherche complexe sur le corpus et les annotations, par exemple par le biais d'expressions régulières.

Révision Dans la grande majorité des campagnes d'annotation les ressources disponibles ne sont pas suffisantes pour corriger manuellement tout le corpus annoté. Cette correction se fait donc de manière plus ou moins automatisée, à partir des informations recueillies sur les erreurs commises par les annotateurs, par exemple lors du calcul des accords inter-annotateurs, ou à partir d'autres indicateurs. Des régularités dans les erreurs peuvent en effet être identifiées et corrigées de manière globale sur tout le corpus, sans l'intervention d'un expert.

Le gestionnaire, associé à un expert s'il ne l'est pas lui-même, peut ainsi prendre la décision de fusionner deux catégories trop ambiguës entre elles (il faut dans ce cas adapter le guide d'annotation). Il peut également supprimer une ou des catégories posant problème (il faut là-aussi adapter le guide). Il peut enfin décider de ne pas prendre en compte les annotations d'un annotateur si elle se révèlent trop différentes des autres (cas où plus de deux annotateurs participent à la campagne, notamment dans le *crowdsourcing*).

Des procédures de correction semi-automatiques ont été utilisées dans la campagne 5b d'annotation d'entités nommées structurées dans la presse ancienne (voir section 5.5). Ces corrections ont été réalisées à partir d'une analyse manuelle des erreurs sur un échantillon du corpus annoté.

Publication Il est important que la qualité du corpus annoté révisé (ou corpus « final ») soit évaluée. Dans le cas d'une correction par adjudication des désaccords, une évaluation réalisée par un expert d'un échantillon pris au hasard parmi les éléments non corrigés peut suffire à évaluer la qualité finale du corpus. Dans le cas exceptionnel d'une correction totale du corpus annoté, l'évaluation finale ne nous paraît pas indispensable, mais peut être effectuée sur un échantillon par un expert différent.

Cette évaluation servira en quelque sorte de label au corpus annoté et pourra être prise en compte, par exemple, lors de l'évaluation d'outils entraînés avec ce corpus. Le corpus est publié accompagné de son guide d'annotation à jour, si possible versionné.

Dans tous les cas, les indicateurs fournis avec le corpus annoté sont au cœur de la décision du gestionnaire de la campagne. Or, ils sont trop souvent sous-utilisés. Nous montrons dans la sous-section suivante deux exemples de ces indicateurs.

4.4.2 Exemples d'indicateurs

Utilisation des incertitudes

Suite à la campagne 5.1 et face à l'impossibilité d'obtenir une vision précise de la cohérence du travail des annotateurs, nous avons décidé de procéder à une seconde validation, *a posteriori*, centrée cette fois-ci sur la confiance des annotateurs en leur

annotation. Étant donné le temps imparti et la disponibilité des annotateurs, nous avons choisi de travailler sur un sous-ensemble du corpus. Pour s'assurer de la représentativité de l'échantillon, les annotateurs ont cherché à identifier une éventuelle typologie des textes du corpus selon l'objet d'étude. Ils sont revenus rapidement sur les fichiers et ont listé les « types » suivants, éventuellement associés à certains mots clefs (*entre parenthèses*) :
- test de mutagénicité (*Salmonella*)
- articles mixtes procaryote/eucaryote : relation hôte agent
- « essais cliniques », épidémiologie (*prevalence, patient, clinical trial*) : pas de protéine, que des taxons
- écologie microbienne (bacterial community, microcosme) : contenant surtout des taxons
- « système d'expression », par exemple de protéines eucaryotes dans une bactérie, une levure
- virus eucaryote

Nous avons ensuite extrait 25 fichiers parmi les 499 notices du corpus (soit 5%) en faisant en sorte d'y inclure un ou des fichiers pour chaque type ainsi identifié. L'objectif de cette seconde validation étant d'évaluer la confiance des annotateurs en leur annotation, nous avons choisi de profiter de l'étiquette *uncertainty* (incertitude), déjà proposé dans Cadixe et de l'utiliser de manière systématique en cas de doute.

Nous avons donc redéfini les types d'incertitudes, en fonction de l'expérience des annotateurs. Nous obtenons les types suivants :
- `common-proper` : doute entre *common* et *proper*
- `too-generic` : trop générique (déjà présent dans la liste)
- `common-proper-too-generic` : à la fois trop générique et doute entre *common* et *proper*
- `other` : en cas de doute non prévu, un commentaire doit alors être ajouté pour l'expliquer.

Cette évaluation n'a nécessité que quelques heures de travail (entre 3 et 5, échantillonnage compris) aux annotateurs et nous a permis de mieux qualifier et surtout quantifier leurs doutes. Au final, sur 555 annotations, les annotateurs ont déclaré 113 incertitudes (au lieu de 5 précédemment), ce qui représente environ 20 % des étiquettes (voir tableau 4.1).

On voit ici que les *gene-common-bacterial* concentrent la majorité des incertitudes (plus de 75 %), et que ces incertitudes sont très largement (à 77 %) liées à une difficulté à distinguer les *common* des *proper*.

Ces chiffres sont d'autant plus intéressants que les *gene-common-bacterial* ne sont pas majoritaires dans l'échantillon, puisqu'ils ne représentent que 17,3 % des étiquettes (voir tableau 4.2). De fait, 86 *gene-common-bacterial* sur 96 portent des *uncertainty-types*. En se penchant sur le détail des annotations, on se rend rapidement compte que

Etiquette + type	Nb	%
gene-common-EukVirus uncertainty-type=common-proper-too-generic	4	3,54 %
gene-common-EukVirus uncertainty-type=too-generic	2	1,77 %
gene-common-Bacterial uncertainty-type=too-generic	19	16,81 %
gene-common-Bacterial uncertainty-type=common-proper	49	**43,36 %**
gene-common-Bacterial uncertainty-type=common-proper-too-generic	16	14,16 %
gene-common-Bacterial uncertainty-type=other	2	1,77 %
gene-common-Eukaryotic uncertainty-type=common-proper	16	14,16 %
gene-common-Eukaryotic uncertainty-type=common-proper-too-generic	2	1,77 %
gene-proper-Bacterial uncertainty-type=annotator-uncertainty	1	0,88 %
taxon-common uncertainty-type=too-generic	2	1,77 %
Total	113	100 %

TABLEAU 4.1: Répartition des incertitudes par étiquette et type sur l'échantillon (campagne 1)

ce chiffre élevé est dû au fait que les mêmes « entités » sont répétées de nombreuses fois, multipliant ainsi leur impact sur les résultats. On trouve notamment 11 occurrences de « cyclophilins » (ou dérivés) dans l'échantillon, or, l'étiquetage de cet élément a posé problème aux annotateurs, ce qui augmente considérablement les chiffres des incertitudes.

Cette constatation est une bonne nouvelle, car elle signifie qu'il sera facile de faire baisser le nombre des incertitudes en levant les ambiguïtés sur un petit nombre de cas.

Quant au type d'incertitude *other*, il a été utilisé pour l'expression « DNA-binding protein 1 », avec le commentaire « double : common-proper et boundaries (acronyme = MDP1) ». Un autre a été utilisé pour « beta » dans la phrase « To investigate the secretion function of the beta-domain », sans commentaire (en fait, problème de frontière).

Enfin, un type *annotator-uncertainty* a été conservé pour l'expression « tfdCDEF ».

Utilisation d'évaluations fines

Dans le cadre de la campagne d'annotation de matchs de football (voir section 5.3), il nous a paru, à Vincent Claveau et nous-même, que la disparité de type des éléments annotés (unités, actions, relations) et la disparité des média sources impliquaient une analyse plus localisée.

Etiquette	**Occurrences**	**%**
taxon-proper	151	27,21 %
taxon-common	16	2,88 %
gene-proper-EukVirus	25	4,51 %
gene-proper-included-Bacterial	2	0,36 %
gene-proper-Bacterial	224	**40,36 %**
gene-proper-Phage	2	0,36 %
gene-proper-Eukaryotic	12	2,16 %
gene-proper-included-Eukaryotic	2	0,36 %
gene-common-EukVirus	6	1,08 %
gene-common-Bacterial	96	**17,3 %**
gene-common-Eukaryotic	18	3,24 %
Note	1	0,18 %
Total	555	100 %

TABLEAU 4.2: Répartition des étiquettes dans l'échantillon (campagne 1)

Nous avons donc regroupé les catégories en différents ensembles qui nous semblaient pertinents pour l'analyse : *acteurs* (regroupant les entités nommées *Joueur*, *Entraineur*, *Président*, etc), *circonstants* (regroupant *EspaceSurTerrain*, *LieuDuMatch* et *TempsDansMatch*), *actions impliquant l'arbitre* (regroupant *FaireFauteDeJeu*, *HorsJeu*, etc.), *autre actions*, *relations impliquant l'arbitre* (regroupant *FaireFauteSurJoueur*, *TaclerFaute* et *RemplacerJoueur*), *autres relations*. Ces regroupements sont motivés par le fait qu'il nous est apparu que, d'une part, les entités nommées et les circonstants sont de natures très différentes et d'autre part que l'intervention de l'arbitre permet en général de plus facilement repérer une action ou une relation. Les résultats obtenus pour chaque ensemble sont présentés dans le tableau 4.3.

	CR		Transcriptions	
	κ de Cohen	κ de Carletta	κ de Cohen	κ de Carletta
Acteurs	0,9228	0,9228	0,8974	0,8973
Circonstants	0,4827	0,4826	0,4441	0,4440
Actions impliquant l'arbitre	0,5999	0,5999	0,5082	0,5082
Autres actions	0,3240	0,3240	0,1407	0,1403
Relations impliquant l'arbitre	0,6355	0,6354	0,4520	0,4503
Autres relations	0,5540	0,5540	0,3793	0,3789

TABLEAU 4.3: Accords par modalité et catégorie pour la campagne 3 d'annotation de matchs de football

Ce tableau montre qu'il existe des disparités significatives entre les ensembles de catégories que nous avons identifiés, ce qui valide notre hypothèse de départ et nos regroupements.

Ces résultats confirment d'une part qu'il n'y a pas de biais entre les annotateurs et, d'autre part, que les transcriptions ont généré davantage de désaccords entre les annotateurs. Comme nous l'avions escompté, les actions ou relations impliquant une décision de l'arbitre ont été annotées de manière plus cohérente que les autres, sans doute parce qu'elles étaient plus facilement identifiables. En ce qui concerne les unités, les résultats sont très différents selon qu'ils concernent les acteurs (qui génèrent des Kappas supérieurs à $0,8$) ou les circonstants (Kappa entre $0,44$ et $0,48$), davantage sujets à interprétations. Enfin, certaines catégories ont été annotées de manière peu cohérente (notamment *PosséderBallon* et *FaireTentative2Passe*)

Une analyse plus détaillée encore montre que les annotateurs sont rarement en désaccord sur les catégories associées aux éléments annotés, mais qu'ils annotent des éléments différents. Ce dernier point confirme l'intérêt d'estimer le plus précisément possible les annotables (voir sous-section 8.3.3).

Au final, ces analyses poussées des accords inter-annotateurs nous ont permis d'identifier les catégories « à problème » et devraient ainsi limiter le coût de correction du corpus annoté.

Chapitre 5

Conduire une campagne d'annotation

Nous présentons dans ce chapitre les campagnes d'annotation dans lesquelles nous avons été directement impliquée, soit en tant que gestionnaire de la campagne (campagnes 1 à 3), soit en tant que participant à la gestion de la campagne (campagnes 4 et 5) de notre arrivée à l'INIST-CNRS en octobre 2008 à juin 2012 (la campagne de pharmacologie étant toujours en cours). Toutes ces campagnes ont pour cadre le programme Quæro. Leur répartition temporelle est présentée dans la figure 5.1.

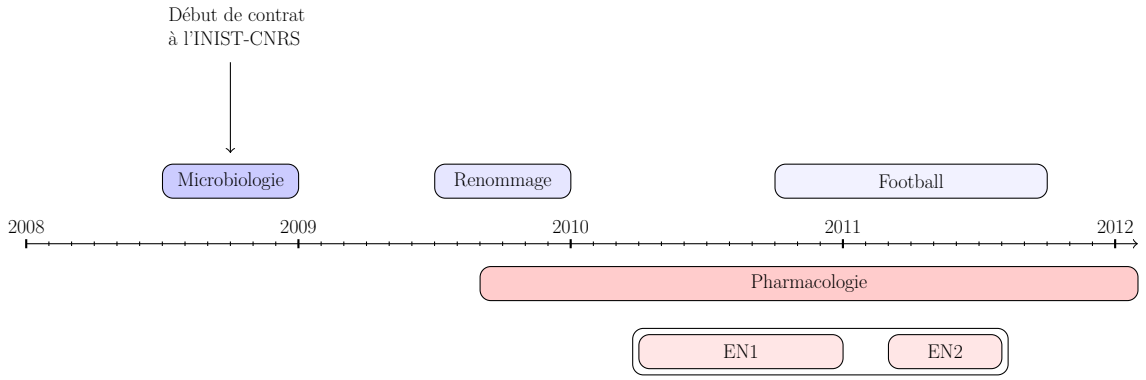


Figure 5.1: Répartition temporelle des campagnes d'annotation (en haut et en bleu celles que nous avons gérées, en bas et en rouge, celles auxquelles nous avons participé).

La description de ces campagnes est de granularité variée, les campagnes terminées étant relativement détaillées, alors que les autres, en particulier celles que nous n'avons pas gérées directement ou qui sont encore en cours, sont présentées plus rapidement.

5.1 Campagne 1 : noms d'espèces, de gènes et de protéines

5.1.1 Motivations et généralités

Dans le cadre de la tâche de reconnaissance des entités nommées du programme Quæro, l'équipe de l'unité Mathématique, Informatique et Génome (MIG) de l'INRA de Jouy

en Josas et l'INIST-CNRS ont collaboré afin de réaliser la validation manuelle d'un corpus en microbiologie devant servir à l'entraînement des outils d'annotation automatique.

La tâche a consisté à corriger une annotation automatique de noms de gènes, de protéines et d'espèces en microbiologie. Un premier corpus de 499 notices bibliographiques extraites de la base MEDLINE (titres et résumés en anglais, soit environ 110 000 tokens) pré-annotées par l'équipe MIG par la simple application d'un dictionnaire et d'un anti-dictionnaire a été fourni aux annotateurs. Leur tâche a consisté en la correction de ces annotations.

La correction de ce premier corpus s'est déroulée entre juin et octobre 2008. La première version du corpus fourni en juin, trop générale, a dû être revue, pour être recentrée sur les bactéries. La validation « finale » n'a donc réellement commencé que début septembre. A raison de deux à trois matinées par semaine, cela représente environ 80 heures de travail par personne, soit plus de six heures par notice [1]. Nous n'avons commencé à travailler sur cette campagne, en tant que gestionnaire de la campagne à l'INIST-CNRS, qu'en octobre. Nous avons donc mené des expériences complémentaires avec la participation des annotateurs, en fin d'année 2008. Avant notre arrivée, le client (INRA MIG) faisait office de gestionnaire de la campagne, à distance.

Le travail de correction a été réalisé à l'INIST-CNRS par deux annotateurs experts : Bernard Taliercio, ingénieur documentaliste en microbiologie et Françoise Tisserand, ingénieure documentaliste en biotechnologies. Ces deux annotateurs ont été choisis par l'INIST-CNRS pour leurs compétences dans le domaine de la microbiologie. Il est cependant important de noter qu'ils ne sont pas spécialisés en annotation, mais en indexation. Par ailleurs, s'ils lisent l'anglais, ils ne sont pas bilingues.

5.1.2 Conditions d'annotation

Les annotateurs ont utilisé le logiciel `Cadixe` (voir en annexe, sous-section A.1.2), qui a été fourni par l'équipe MIG à l'INIST-CNRS, accompagné d'un manuel d'utilisation en français.

Le logiciel a été fourni sans procédure de gestion de bugs particulière, mais les interactions par courrier électronique ont été plutôt efficaces et les bugs ont été rapidement corrigés. Par ailleurs, lorsque la configuration de `Cadixe` a dû être mise à jour suite à l'ajout d'étiquettes, l'équipe MIG a fourni les fichiers modifiés (DTD et CSS) aux annotateurs. Cette partie de la gestion du projet, menée de manière pragmatique, s'est révélée assez efficace, puisque les problèmes pratiques rencontrés par les annotateurs ont été résolus rapidement.

1. Ce temps comprend également les interactions avec MIG, les différentes mises au point réalisées et la rédaction des documents.

Ceux-ci se sont réunis dans un même bureau plusieurs matinées par semaine pendant presque deux mois. Ils ont travaillé à deux, l'un sur l'ordinateur, l'autre en retrait, ce qui leur a permis d'avoir un certain recul par rapport à la manipulation de l'outil lui-même, suivant en cela une division des rôles connue, décrite dans (Demirsahin *et al.*, 2012). Ils se mettaient d'accord en direct sur les annotations. En cas de doute sur la méthodologie, ils consultaient le guide d'annotation ou leur propre documentation (courriers ou décisions précédentes). En cas de doute sur le fond, ils lançaient une recherche sur Internet, essayant de retrouver l'article en son entier, afin d'obtenir des éléments de contexte supplémentaires. Cette recherche les guidait souvent vers les services Web de l'INIST-CNRS et les annotateurs allaient parfois jusqu'à y accéder en interne pour lire l'article. Dans tous les cas, la négociation se faisait en direct et semblait assez efficace, à l'image de ce qui est décrit dans (Demirsahin *et al.*, 2012). Cependant, en l'absence de comparaison avec une annotation non commune (chaque annotateur travaillant de son côté), il est impossible d'évaluer l'impact de cette collaboration sur l'annotation et en particulier l'influence d'un annotateur sur l'autre.

Les annotateurs utilisaient peu les raccourcis clavier et avaient toujours un certain nombre de documents à portée de main pendant leur travail :
- le guide d'annotation,
- le document d'historique qu'ils s'étaient créé,
- la compilation des mails échangés avec l'équipe MIG,
- un cahier sur lequel étaient notés les notices traitées et leur statut (*OK*, *problème* sur telle entité, etc).

Les annotateurs se sont donc adaptés efficacement à l'outil d'annotation, en créant leur propre méthode de prise de notes (cahier) et en collaborant directement.

5.1.3 Documentation

Les annotateurs ont reçu au début du projet un guide d'annotation, écrit en anglais, qui a été mis à jour au fur et à mesure. Ce guide d'annotation détaille les choix à effectuer en matière d'annotation et fournit de nombreux exemples très utiles. Il fournit également un anti-dictionnaire de termes trop courants pour être annotés. Il invite à une mise à jour, qui a effectivement été faite par les annotateurs.

Outre le guide d'annotation et le manuel d'utilisation de `Cadixe`, l'équipe MIG a fourni à l'INIST-CNRS un certain nombre de documents, dont, en particulier, un manuel détaillant l'utilisation de `Cadixe` pour cette campagne d'annotation. Par la suite, d'autres documents sont venus s'ajouter, notamment des commentaires concernant la première passe d'annotation, ainsi qu'une FAQ compilée par l'équipe MIG à partir d'un échange de courriers sur un sujet spécifique.

Les annotateurs ont par ailleurs eu le bon réflexe de créer un document leur permettant de conserver un historique des choix complexes qu'ils ont dû faire. Ce document

contient une liste d'entités dont l'annotation a posé problème et nécessité des recherches plus approfondies. Ils ont également créé un document regroupant tous les échanges menés par mail avec l'équipe MIG au sujet de l'annotation. Ce document, plus complet que la FAQ de l'équipe MIG, mais non organisé, s'est révélé très utile en tant que document de référence.

5.1.4 Étiquettes

Le jeu d'étiquettes est relativement simple. En effet, les catégories sont regroupées dans deux groupes :

- des noms d'organismes (nommés *Taxon*) : procaryotes, eucaryotes, virus,
- des noms de gènes et de protéines (nommés *Gene*) dans un sens élargi (objets biologiques composés de protéines, de gènes ou de parties de ceux-ci).

Le guide d'annotation fourni distinguait d'un côté les noms propres (*proper*, des entités nommées) et de l'autre les noms communs (*common*, des termes), ce qui correspondait au fait que l'annotation automatique met en œuvre des techniques différentes pour la reconnaissance de noms propres et de noms communs. Ce choix d'influencer l'annotation par l'outil s'est révélé par la suite inopportun, les annotateurs étant qualifiés en microbiologie, non en TAL.

Au départ, MIG a proposé six étiquettes pour l'annotation :

- *taxon-proper*,
- *taxon-common*,
- *gene-proper-Bacterial*,
- *gene-proper-Eukaryotic*,
- *gene-common-Bacterial*,
- *gene-common-Eukaryotic*.

Les annotateurs ont très rapidement proposé deux nouvelles étiquettes :

- *gene-proper-Both-BactEuk*,
- *gene-common-Both-BactEuk*

Les quatre dernières étiquettes ont été ajoutées plus tardivement :

- Par expérience, sur le corpus :
  - *gene-proper-Phage*,
  - *gene-common-Phage*.
- Puis, par déduction :
  - *gene-proper-EukVirus*,
  - *gene-common-EukVirus*.

Ce type d'évolution est inévitable dans une tâche d'annotation ou de validation. Il est donc important que l'outil d'annotation soit assez souple pour les gérer facilement, ce qui est le cas de `Cadixe`.

Il est à noter que si la stratégie d'annotation privilégiait la correspondance la plus longue, elle permettait aussi les inclusions. Ainsi, « thymidylate synthase (thyA) gene » a donné lieu aux annotations suivantes :
– thymidylate synthase (thyA) gene : gene-common-Bacterial,
– thyA : gene-proper-included-Bacterial.

5.1.5 Évaluation de la correction réalisée

La pré-annotation automatique a généré 5 387 annotations (voir le tableau 5.1), dont près de 70 % de *gene*, surtout *proper* (76,5 % d'entre eux). Qu'ils soient *proper* ou *common*, on ne trouve que des *Bacterial*.

L'annotation validée manuellement comprend 6 705 étiquettes (soit 24,4 % de plus). Les gènes (*gene*) y représentent plus de 60 % des étiquettes. Plus de 90 % des *genes* sont encore des *Bacterial* (contre 100 % auparavant), dont près de 66 % de *proper*, mais on voit apparaître d'autres *genes*, en particulier des *Eukaryotic* (moins de 8 % des *genes*), mais également des *EukVirus* (0,7 %), des *Phages* (0,1 %), des *include-Bacterial* (0,3 %) et des *include-Eukaryotic* (0,07 %). Par contre, on ne voit toujours pas d'étiquette *include-Both-BactEuk*, et *Both-BactEuk* n'apparaît que deux fois, dans les *genes-common*.

Si l'augmentation la plus importante numériquement est celle touchant les *taxon-common* (+1196,4 %), l'évolution la plus significative concerne à notre avis le ré-équilibrage *gene-proper*/*gene-common*, au profit de ces derniers, qui, bien qu'ils restent minoritaires (35 % des *genes*), ont vu leur nombre augmenter de plus de 68 %.

Les annotateurs avaient la possibilité d'ajouter un attribut d'incertitude (*uncertainty*) à leurs annotations, éventuellement en le typant, les types prévus étant les suivants :
– *uncertain-boundaries* (frontières incertaines),
– *annotator-uncertainty* (incertitude de l'annotateur),
– *maybe-too-generic* (peut-être trop générique),
– *lack-of-info-in-text* (manque d'information dans le texte).
Pourtant, on ne trouve en tout que 55 occurrences de cet attribut (0,8 % des annotations), soit 26 *uncertain-boundaries*, 21 *annotator-uncertainty* et 8 *maybe-too-generic*. Une analyse un peu plus approfondie montre qu'une majorité des incertitudes a porté sur les *gene-common-Bacterial* (39 cas, soit plus de 70 % des cas), et en particulier sur les frontières de ceux-ci (17 cas). S'il n'est pas étonnant que les difficultés concernent en particulier les *Bacterial*, qui représentent plus de 56 % des étiquettes, il est plus surprenant que les étiquettes *common* soient plus touchées, alors qu'elles sont deux fois moins nombreuses que les *proper*. Apparemment, le guide d'annotation n'était pas très clair sur cette étiquette, voire parfois contradictoire, et cela avait entrainé beaucoup de discussions avec l'équipe MIG, sans pour autant résoudre les problèmes.

Étiquette			Avant valid.		Après valid.		Écart
taxon				1682		2540	+51 %
	proper		1654		2177		+31,6 %
	common		28		363		+1196,4 %
gene				3705		4165	+12,4 %
	proper			(2836)		(2708)	-4,5 %
		Bacterial	2836		2503		-13,3 %
		Eukaryotic	0		164		
		Both-BactEuk	0		0		
		Phage	0		4		
		EukVirus	0		25		
		incl.-Bacterial	0		9		
		incl.-Eukaryotic	0		3		
		incl.-Both-BactEuk	0		0		
	common			(869)		(1464)	+68,4 %
		Bacterial	869		1290		+48,4 %
		Eukaryotic	0		160		
		Both-BactEuk	0		2		
		Phage	0		2		
		EukVirus	0		6		
		incl.-Bacterial	0		4		
		incl.-Eukaryotic	0		0		
		incl.-Both-BactEuk	0		0		
Total			5387		6705		+24,4 %

Tableau 5.1: Répartition des étiquettes de l'annotation (avant et après validation, campagne 1)

Lors d'un entretien de suivi de projet, l'un des annotateurs a confirmé que l'attribut d'incertitude a été très peu utilisé, les annotateurs le réservant pour les cas où ils ne savaient vraiment pas quoi faire. Il manque donc une évaluation plus précise de la fiabilité des annotations.

Notre arrivée tardive dans la campagne ne nous a pas permis de mettre en place de méthodologie adaptée. Nous avons cependant pu réaliser une évaluation de la campagne *a posteriori* (voir sous-section 4.4.2). Les annotations réalisées dans cette campagne n'ont finalement pas pu être utilisées par MIG, les incertitudes des annotateurs et les désaccords avec les experts de l'INRA étant trop nombreux. Cette campagne est de ce point de vue un échec, mais l'expérience acquise nous a permis de mettre en place de manière plus structurée la campagne suivante d'annotation de renommage de noms de gènes (voir section 5.2).

5.2 Campagne 2 : relations de renommage de gènes

5.2.1 Motivations et généralités

Cette campagne, que nous présentons dans (Jourde *et al.*, 2011) et sur sa partie évaluation dans (Fort *et al.*, 2012a), a été menée de juin à début décembre 2009 à l'INIST-CNRS, une nouvelle fois en collaboration avec l'équipe MIG [2] de l'INRA de Jouy en Josas.

Elle a consisté à identifier, dans des résumés (*abstracts*) de la base MEDLINE, des relations de renommage de noms de gènes. Ce phénomène est en effet courant en microbiologie, où les noms donnés aux gènes lors de leur découverte sont souvent remplacés par la suite par des noms plus significatifs, dès lors qu'on en identifie la fonction. Si les nomenclatures ne sont pas toujours à jour à ce sujet, les renommages peuvent être mentionnés dans des publications scientifiques et donc identifiés par ce biais. Le but de l'équipe MIG était d'automatiser l'extraction des mentions de renommage à partir de ces textes. Le travail d'annotation manuelle devait leur permettre de développer des méthodes de construction de listes de gènes synonymes. Une première version de leur outil d'extraction des mentions de renommage était déjà opérationnelle et le corpus annoté créé par la campagne a servi à son entraînement et à son évaluation. Il est à noter que cette campagne n'a concerné que les relations de renommage des gènes de la bactérie *Bacillus Subtilis*, une bactérie très utilisée dans la recherche en microbiologie comme espèce modèle.

Ce corpus a été utilisé par la suite comme corpus d'entraînement et de test dans le cadre de la tâche commune d'annotation de relations de renommage de gènes de bactéries de BioNLP 2011 [3] (Jourde *et al.*, 2011) et est disponible librement pour la recherche, sous licence INRA [4].

5.2.2 Conditions d'annotation

L'annotation du corpus a été réalisée à l'INIST-CNRS, par les deux mêmes annotateurs que pour la campagne d'annotation de noms de gènes et de protéines (voir section 5.1), à l'aide du même outil, `Cadixe`.

A la différence de la campagne précédente, nous avons pu intervenir dès le début de la campagne en tant que gestionnaire de la campagne à l'INIST-CNRS. Nous avons donc mis en place la méthodologie définie dans (Bonneau-Maynard *et al.*, 2005), complétée par les connaissances acquises lors de la campagne 1 (Fort *et al.*, 2009). Nous avons

2. http://genome.jouy.inra.fr/bibliome/renommage/
3. https://sites.google.com/site/bionlpst/home/bacteria-gene-renaming-rename
4. http://weaver.nlplab.org/~bionlp-st/BioNLP-ST/downloads/downloads.shtml

ainsi calculé l'accord inter-annotateurs dès le début de la campagne, afin de mettre au jour les désaccords et de modifier le guide d'annotation en conséquence.

Par ailleurs, une expérience préalable d'annotation de renommage de gènes a eu lieu à l'INRA début juin 2009. Cette expérience s'est déroulée sur une journée, avec neuf annotateurs. Les retours de l'INRA sur cette première expérience nous ont permis de faire quelques modifications dans le guide d'annotation et d'extrapoler les temps d'annotation pour deux annotateurs.

La méthodologie appliquée lors de cette campagne correspond en partie à celle présentée au chapitre 4, puisqu'elle reprend les bonnes pratiques préconisées par (Bonneau-Maynard *et al.*, 2005). Cela s'est cependant avéré insuffisant. Ainsi, si l'expérience préalable d'annotation menée à l'INRA peut être considérée comme la création d'une mini-référence, celle-ci n'a pas été utilisée lors de la campagne. En outre, les annotateurs ayant déjà utilisé l'outil `Cadixe` et l'annotation étant jugée simple par le client, aucune formation particulière n'a été dispensée aux annotateurs, ce qui a eu pour conséquences de nombreux questionnements sur la tâche, donc une complexification de sa gestion. Enfin, il y a eu correction manuelle du corpus annoté par un expert de l'INRA. Notre grille d'analyse des dimensions de complexité d'une campagne d'annotation n'existait pas encore et nous n'avons pu l'appliquer qu'*a posteriori* (voir chapitre 6).

5.2.3 Sélection du corpus

L'annotation du corpus ne concerne que les relations de renommage entre noms de gènes chez *Bacillus Subtilis*. Un corpus a donc été élaboré spécifiquement pour la tâche. Il a été constitué par Julien Jourde, de l'INRA MIG, comme suit :

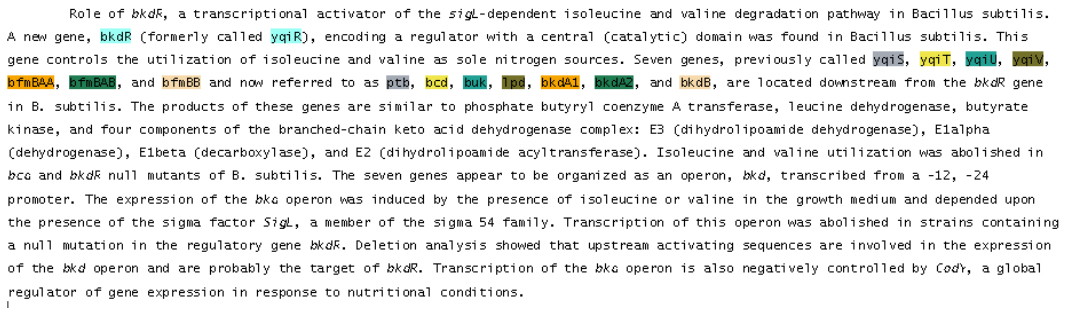
- Recherche dans PubMed des textes (en anglais) traitant de *Bacillus Subtilis*.
- Dans ces textes, recherche de co-occurrences de noms de gènes potentiellement synonymes (la liste de ces couples est issue de différentes sources).
- Si les textes ne présentent aucune co-occurrence, des marqueurs linguistiques de renommage (par exemple *renamed*, *termed*, *designated*) sont recherchés et permettent un éventuel repêchage des textes.

1 843 textes ont ainsi été sélectionnés. 703 d'entre eux mentionnent des couples de synonymes (1 014 couples détectés au total, redondances incluses). Les 1 843 fichiers (*abstracts*) ont été regroupés en 74 lots par l'INRA, et nous avons répartis ceux-ci en paquets pour l'annotation. Certains lots ont été annotés par les deux annotateurs, afin de calculer l'accord inter-annotateur. D'autres ont été annotés deux fois par chaque annotateur afin de calculer l'accord intra-annotateur.

5.2.4 Éléments à annoter

Les relations de renommage ont été annotées, grâce à l'outil `Cadixe`, en sélectionnant le nom d'origine du gène (annoté *Former*), puis son nouveau nom (annoté *New*).

`Cadixe` ne permettant pas d'annoter directement des relations, chaque couple a été traité comme une entité distincte et leur relation est matérialisée par un identifiant commun. L'annotation étant généralement séquentielle, les éléments du premier couple rencontré dans le texte sont annotés *Former id="1"* et *New id="1"*, ceux du second couple *Former id="2"* et *New id="2"* et ainsi de suite.


Role of *bkdR*, a transcriptional activator of the *sigL*-dependent isoleucine and valine degradation pathway in Bacillus subtilis.
A new gene, bkdR (formerly called yqiR), encoding a regulator with a central (catalytic) domain was found in Bacillus subtilis. This gene controls the utilization of isoleucine and valine as sole nitrogen sources. Seven genes, previously called yqiS, yqiT, yqiU, yqiV, bfmBAA, bfmBAB, and bfmBB and now referred to as ptb, bcd, buk, lpd, bkdA1, bkdA2, and bkdB, are located downstream from the *bkdR* gene in B. subtilis. The products of these genes are similar to phosphate butyryl coenzyme A transferase, leucine dehydrogenase, butyrate kinase, and four components of the branched-chain keto acid dehydrogenase complex: E3 (dihydrolipoamide dehydrogenase), E1alpha (dehydrogenase), E1beta (decarboxylase), and E2 (dihydrolipoamide acyltransferase). Isoleucine and valine utilization was abolished in *bcd* and *bkdR* null mutants of B. subtilis. The seven genes appear to be organized as an operon, *bkd*, transcribed from a -12, -24 promoter. The expression of the *bkd* operon was induced by the presence of isoleucine or valine in the growth medium and depended upon the presence of the sigma factor *SigL*, a member of the sigma 54 family. Transcription of this operon was abolished in strains containing a null mutation in the regulatory gene *bkdR*. Deletion analysis showed that upstream activating sequences are involved in the expression of the *bkd* operon and are probably the target of *bkdR*. Transcription of the *bkd* operon is also negatively controlled by *CodY*, a global regulator of gene expression in response to nutritional conditions.

Dans cet exemple, *bkdR*/*yqiR* (en bleu clair) constitue le premier couple et prend l'identifiant 1, puis le second couple, *ptb*/*yqiS* (en gris), est annoté en tant que couple d'identifiant 2 et ainsi de suite, jusqu'à l'annotation du dernier couple *bkdB*/*bfmBB* (en beige) en tant que couple d'identifiant 8.

Certains fichiers (plus d'un tiers d'entre eux) ne comportaient pas de renommage du tout. Nous avons ainsi obtenu, en moyenne sur l'échantillon, un renommage par fichier. Les accords et désaccords ont été analysés qualitativement, ce qui nous a permis d'ajouter les cas non traités dans le guide d'annotation.

5.2.5 Évaluation

Nous avons fait annoter par les deux annotateurs dès le début de la campagne (entre mi-juillet et fin août 2009), un même échantillon de 93 fichiers, correspondant à plus de 19 000 tokens, à partir duquel nous avons ensuite calculé les accords inter- et intra-annotateurs, tel que recommandé par (Gut et Bayerl, 2004). Cette annotation a été réalisée par les deux annotateurs en parallèle, sans aucune concertation.

Nous avons obtenu un Kappa de Cohen (Cohen, 1960) de $0,8$ (pour plus de détails sur ce coefficient, voir section 3.3), ce qui, si l'on se réfère à (Artstein et Poesio, 2008) [5] peut sans doute être considéré comme satisfaisant. Cependant, étant donné la

5. *« [I]f a threshold needs to be set, 0.8 is a good value »* (p.591)

très grande dispersion des renommages dans le corpus, donc la très grande prévalence d'une catégorie *Rien* ou *Nulle*, il nous a paru important de vérifier la validité de ces résultats. Les expériences réalisées dans ce but sont rapportées dans (Fort *et al.*, 2012a) et concluent que s'il existe bien une influence de la prévalence, celle-ci reste limitée, l'accord inter-annotateurs atteignant $0,73$ sans la catégorie *Rien*.

Au final, cette campagne aura permis de mettre au jour manuellement environ 200 couples de renommage.

5.3 Campagne 3 : entités nommées, actions et relations en football

5.3.1 Motivations et généralités

Cette campagne a commencé en septembre 2010, après un faux départ dû à des problèmes de financement. Elle a de nouveau été stoppée d'octobre à décembre, du fait d'importants changements organisationnels à l'INIST-CNRS qui n'ont pas permis aux annotateurs d'être aussi disponibles que nécessaire. Quelques mois plus tard, nous somme partie en congé maternité. La campagne ne s'est donc pas déroulée selon le calendrier prévu initialement et ne s'est terminée qu'en septembre 2011.

Elle a consisté à annoter un corpus très hétérogène, mêlant brefs comptes-rendus de matchs de football diffusés sur le Web et transcriptions de commentaires vidéos des mêmes matchs, télévisés, le tout en français. L'application envisagée à terme est celle de résumé automatique de matchs. Les éléments à annoter sont donc des éléments du domaine.

Le corpus que nous avons utilisé couvre 16 matchs de football. Il est composé de 24 transcriptions de commentaires tirés de vidéos (1 par mi-temps, 12 matchs) et de 16 fichiers contenant une description minute-par-minute du match (dont les 12 de la transcription et 4 matchs additionnels) tirés de sites Web spécialisés. L'ensemble du corpus a une taille d'environ 250 000 tokens. Comme cela a été souligné dans (Fort *et al.*, 2011b), sa principale caractéristique est d'être très hétérogène, que ce soit d'un point de vue des types de match (matchs de ligue 1 français, matchs internationaux, matchs de championnats étrangers), de la taille des fichiers (de 1 116 tokens par match pour les minutes à 21 000 tokens par match pour les transcriptions), ou de la source (chaînes de diffusion des vidéos, commentateurs, sites Web).

Cette campagne a fait l'objet de deux articles écrits en collaboration avec Vincent Claveau (IRISA/CNRS), l'un mettant l'accent sur le calcul et l'utilisation des accords inter-annotateurs (Fort et Claveau, 2012b) et l'autre focalisant davantage sur

l'influence du média source sur les annotations (Fort et Claveau, 2012a). Nous en reprenons ici certains éléments, en les complétant.

5.3.2 Conditions d'annotation

Cette campagne d'annotation a impliqué l'INIST-CNRS, en tant que fournisseur de l'annotation, et l'équipe TexMex de l'IRISA, plus particulièrement Vincent Claveau, en tant que client (fournisseur du corpus et demandeur de l'annotation). Nous en avons été le gestionnaire et avons à ce titre encadré deux annotateurs à l'INIST-CNRS, Claire Ris et Alain Zerouki. Tous les deux sont des ingénieurs documentalistes et des footballeurs aguerris, volontaires pour la campagne, que nous avons formés sur l'outil et la tâche.

Outils

Un inventaire poussé de l'existant (voir annexe A), nous a convaincue d'utiliser pour cette campagne l'outil `Glozz` (Widlöcher et Mathet, 2009), qui a été mis à notre disposition par ses créateurs, Antoine Widlöcher et Yann Mathet. Ce choix a été dicté par la nécessité d'annoter des relations et de le faire de façon simple pour les annotateurs. En effet, l'interface de `Glozz` est très agréable à utiliser, même si elle ne permet pas d'être très efficace, puisque tout se fait *via* la souris. Ce point particulier était fondamental car nos annotateurs ne sont pas toujours très à l'aise avec les nouvelles interfaces. La parole contenue dans les vidéos de matchs a été transcrite manuellement [6] en utilisant `transcriber` (Barras *et al.*, 1998) et le guide de transcription fourni par défaut.

Hormis pour la phase de formation des annotateurs, le corpus a été automatiquement pré-annoté avec les noms de joueurs et d'entraîneurs à partir d'informations collectées automatiquement sur des sites Web spécialisés. Nous avions en effet montré que ce type de pré-annotation permettait de gagner du temps et d'améliorer la qualité des annotations finales, notamment pour l'annotation morpho-syntaxique (voir chapitre 7). Ces pré-annotations ne sont cependant pas exemptes d'erreurs, et une partie du travail d'annotation a donc aussi consisté à corriger ces pré-annotations. Ainsi, l'un des annotateurs a pu noter que 321 entités nommées n'ont pas été pré-annotées alors qu'elles auraient dû l'être (silence), du fait, en particulier, d'erreurs typographiques. Au final, nous obtenons malgré tout un accord inter-annotateurs (en utilisant la mesure Glozz), entre la pré-annotation automatique et les annotateurs, de plus $0,9$ sur les transcriptions et de $0,8$ sur les compte-rendus.

6. Le travail de transcription a été réalisé par l'entreprise e-TT (Easy Top Transcription), qui a été choisie par le client.

Méthodologie

Outre la définition des catégories, accompagnée d'exemples, le guide d'annotation contenait un certain nombre de consignes de travail. Ainsi, il a été demandé aux annotateurs, pour un match donné, de commencer par les minutes et d'annoter ensuite les transcriptions lorsque ces deux fichiers étaient disponibles, les transcriptions étant supposées plus ambiguës. Il est important de noter qu'il a aussi été demandé que les transcriptions soient annotées sans avoir recours à la vidéo, pour s'assurer que les annotations soient propres au média considéré, et non polluées ou interprétées en fonction d'autres sources. Ce choix a également permis de gagner un temps non négligeable (plus de 2 heures par fichier de transcription selon nos tests).

Les différents fichiers à annoter ont été répartis entre les deux annotateurs de manière à ce qu'ils aient une charge de travail similaire, en prenant également en compte la taille des fichiers, le type de match (matchs de Ligue 1, matchs internationaux, etc.) et bien sûr leur source (compte-rendus ou transcriptions).

L'annotation de chaque fichier s'est faite couche par couche (voir sous-section 5.3.3), et les temps d'annotation par couche ont été mesurés par les annotateurs à l'aide d'un outil libre en ligne[7]. Ceux-ci avaient pour consigne de travailler au moins 10 heures par semaine sur ces tâches pour garder un rythme régulier et optimiser la courbe d'apprentissage et la qualité du travail. Nous les avons également invités à ajouter des commentaires sur leurs annotations, et un attribut *Incertitude* a été mis à leur disposition dans `Glozz`. Ces possibilités ont été utilisées de manière très disparate, l'un des annotateurs les ayant beaucoup utilisées et l'autre très peu.

La méthodologie suivie pendant la campagne est celle décrite par Bonneau-Maynard *et al.* (2005) : nous avons calculé des accords inter-annotateurs tôt dans la campagne pour détecter d'éventuelles incohérences dans le modèle d'annotation ou des ambiguïtés dans le jeu d'étiquettes, et ainsi améliorer le guide d'annotation. Nous avons aussi mesuré l'accord intra-annotateur, comme recommandé par Gut et Bayerl (2004), pour mettre à jour des incohérences propres à chaque annotateur.

La campagne a ainsi été divisée en plusieurs phases :

1. formation : sur le plus petit fichier de minutes (sans pré-annotation), avec l'outil d'annotation ;
2. pré-campagne 1 : annotation par les deux annotateurs (séparément) d'un même fichier de minutes, calcul d'accord inter-annotateurs, discussion des désaccords, mise-à-jour du guide d'annotation ;
3. pré-campagne 2 : annotation par les deux annotateurs (ensemble) d'un fichier de minutes, mise-à-jour du guide d'annotation ;

7. TimeTracker, disponible à http://www.formassembly.com/time-tracker/.

4. pré-campagne 3 : annotation par les deux annotateurs (séparément) d'un même fichier de transcriptions, calcul d'accord inter-annotateurs, discussion des désaccords, mise-à-jour du guide d'annotation ;
5. campagne : annotation des fichiers assignés aux annotateurs, match par match ;
6. fin de campagne : calcul des accords inter- et intra-annotateur.

Après une phase de formation « hachée » du fait du report de la campagne, nous avons pu commencer la première pré-campagne. Les annotateurs et nous-même avons annoté le même compte-rendu de match et nous avons comparé les résultats dans le détail, afin de mettre au clair les problèmes et de corriger le guide d'annotation. Malheureusement, cette pré-campagne a elle-aussi été perturbée par le manque de disponibilité des annotateurs.

Nous avons donc décidé de faire travailler les annotateurs en binôme (tous les deux sur un même ordinateur) sur un même compte-rendu, afin d'homogénéiser leurs annotations et de relever les problèmes restants. Lors de cette annotation, l'un des annotateurs a pris en note les problèmes rencontrés, que nous avons examinés, puis nous avons transmis au client les questions en suspens. Une fois un accord trouvé entre les annotateurs, le client et le gestionnaire, le fichier a été corrigé par l'un des annotateurs. Il a été considéré par la suite comme la mini-référence de la campagne.

Dès réception des premières transcriptions de commentaires de matchs [8], nous avons mené une deuxième pré-campagne, afin d'ajouter les éléments spécifiques à ce type de texte très différent au guide d'annotation et de régler les problèmes liés aux différents média mis en jeu (faut-il annoter le texte, l'audio ou la vidéo ? Comment arbitrer entre les trois ?) et à l'annotation des ellipses (par exemple, acteur manquant dans une action de corner).

La campagne s'est ensuite déroulée, en fonction des disponibilités des annotateurs et sans gestionnaire, puisque nous n'avions pas pu être remplacée. Une évaluation finale a cependant eu lieu en fin de campagne.

5.3.3 Étiquettes

Le jeu d'étiquettes utilisé a été défini en trois étapes. Nous avons exploité une ontologie du football existante (Ranwez et Crampes, 1999) développée à l'École des Mines d'Alès [9] dans laquelle nous avons sélectionné et adapté les éléments intéressants pour l'application finale. Nous avons ensuite modifié ces définitions durant la phase d'entraînement (cf. infra), et de nouveau durant la phase de pré-campagne.

8. Ces transcriptions ne sont arrivées correctement pré-annotés que bien après le début de la campagne, mi-décembre 2010.
9. http://www.lgi2p.ema.fr/~ranwezs/ontologies/soccerV2.0.daml

L'ensemble des étiquettes retenues a été divisé en trois couches, chacune correspondant à un niveau d'analyse de complexité croissante, et telles que les éléments de chaque couche soient facilement annotables simultanément. Toutes les étiquettes sont présentées dans le tableau 5.2. Les annotateurs devaient donc se concentrer sur les unités (entités nommées, temps et lieux), ensuite sur les actions et enfin sur les relations.

Unités
Joueur, *Equipe*, *Arbitre*, *Entraineur*, *ArbitreAssistant*, *Président* *EspaceSurTerrain*, *LieuDuMatch*, *TempsDansMatch*
Actions
TirerCoupFrancDirect, *TirerCoupFrancIndirect*, *TirerCorner*, *TirerPenalty*, *FaireFauteDeJeu*, *HorsJeu*, *MarquerBut*, *PrendreCartonJaune*, *PrendreCartonRouge*, *PrendreRappelALOrdre* *Centrer*, *FaireTentative2Centre*, *Dribbler*, *RaterBut*, *ArreterBut*, *IntercepterBallon*, *PossederBallon*, *ActionDuPublic*
Relations
FaireFauteSurJoueur, *TaclerFaute*, *RemplacerJoueur* *FaireCombinaison*, *FairePasse*, *FaireTentative2Passe*

TABLEAU 5.2: Couches d'annotations retenues et étiquettes correspondantes (campagne 3)

Comme nous l'avons dit, le corpus est composé de données relevant de différents médias, dont l'oral transcrit qui est particulièrement riche en ellipses (« `Makoun. Et c'est récupéré. Clerc, avec Cris. Boumsong, Makoun.` »). Nous avons donc décidé de ne pas faire porter les annotations sur les prédicats dénotant les actions ou les relations, souvent absents, mais sur les acteurs impliqués. Ce choix est cohérent avec les besoins applicatifs et permet d'obtenir un processus d'annotation homogène et de simplifier le guide d'annotation.

Cependant, dans certains cas, notamment dans les transcriptions, l'acteur impliqué dans une action peut lui-même être élidé. C'est par exemple le cas de l'acteur du dribble dans la phrase « `Grand dribble en pivot bien pris` ». Ce phénomène se produit aussi pour les relations dans lesquelles l'acteur source et/ou cible peut être manquant : « `Ribéry, avec une faute sur Gourcuff` » (pas d'acteur source pour la faute). Dans de tels cas, les annotateurs ont reçu pour consigne d'ancrer exceptionnellement l'annotation sur le prédicat et d'ajouter un attribut prédéfini (*Acteur manquant* pour les actions et *Cible/source manquante* pour les relations). Pour les relations, il a également fallu ajouter une nouvelle unité *ActionPourActeurVide* utilisée pour annoter le prédicat et ainsi permettre d'ancrer la relation sur cette annotation.

Ce problème d'annotation des ellipses est particulièrement intéressant et riche. On trouve en effet non seulement des actions elliptiques, telles que la suivante, où on ne sait pas qui tire le corner :

Le corner qui suit ne donne rien.

mais également des relations, dont l'acteur, source ou cible, peut manquer, comme dans l'exemple suivant, où on ne sait pas qui fait la passe à Bedimo :

On écarte sur la gauche vers Bedimo

Nous avons mis en place une méthodologie d'annotation de ces ellipses grâce à des traits typés. Cela a demandé un peu de temps d'adaptation aux annotateurs, mais ils ont semblé s'y habituer.

5.3.4 Évaluation

Étant donnée la complexité de l'annotation, il a été difficile d'établir une mesure précise de l'accord inter-annotateurs dès le début de la campagne et nous avons dû pour commencer nous contenter d'un accord observé sur les entités nommées et les actions. Cet accord était d'environ $0,6$ en fin de première pré-campagne, avant annotation en binôme, sur un compte-rendu de match. Une analyse fine des problèmes nous a montré une régression de l'un des deux annotateurs sur certaines catégories (en particulier *Espace_sur_terrain*), mal interprétées. Nous avons donc ré-expliqué ces catégories et modifié le guide. Étant donnée la complexité de la tâche, ce résultat nous a paru acceptable.

Par la suite, `Glozz` a intégré une mesure d'accord applicable à nos annotations (hors relations), la mesure Glozz (Mathet et Widlöcher, 2011a). Nous avons par ailleurs travaillé sur notre propre méthode d'évaluation (détaillée en sous-section 8.3.3). Les résultats des évaluations menées avec ces mesures sont présentés dans le tableau 5.3.

Accords inter-annotateurs			
	κ de Cohen	κ de Carletta	mesure Glozz
CR unités/actions	0,5992	0,5991	0,7627
CR relations	0,5707	0,5707	-
Transcriptions unités/actions	0,5979	0,5879	0,7645
Transcriptions relations	0,4050	0,4025	-
Transcriptions unités/actions	0,6490	0,6490	0,7351
Transcriptions relations	0,4640	0,4635	-
Accords intra-annotateur			
	κ de Cohen	κ de Carletta	mesure Glozz
CR unités/actions A1	0,7531	0,7531	0,8753
CR relations A1	0,6377	0,6377	-
CR unités/actions A2	0,7109	0,7109	0,8519
CR relations A2	0,5985	0,5983	-
Transcriptions unités/actions A1	0,7558	0,7558	0,8327
Transcriptions relations	0,4010	0,3904	-
Transcriptions unités/actions A2	0,6812	0,6812	0,8179
Transcriptions relations	0,4701	0,4700	-

Tableau 5.3: Accords dans la campagne d'annotation de matchs de football

Les Kappas de Cohen (Cohen, 1960) et de Carletta (Carletta, 1996) obtenus sont très proches, ce qui signifie qu'il n'y a pas de biais entre les annotateurs (Artstein et Poesio, 2008). Par ailleurs, tous les résultats montrent que l'annotation de relations génère davantage de désaccords que l'annotation d'unités ou d'actions. On constate en outre sans surprise que l'accord (aussi bien inter- qu'intra-annotateur) a tendance à être plus faible dans les transcriptions que dans les minutes, à l'exception d'une transcription pour laquelle les unités/actions ont produit un accord bien supérieur (près de $0,65$). Cette tendance générale se manifeste spécialement dans les cas d'annotations complexes comme les relations. Les spécificités de l'oral mentionnées précédemment, et en particulier le style elliptique propre aux commentaires, expliquent facilement cette différence.

La disparité des catégories et des média sources nous a cependant incitée à réaliser des évaluations complémentaires, plus fines, qui sont présentées dans la sous-section 4.4.2. Les résultats obtenus devraient faciliter la correction des annotations.

Les annotations produites seront librement disponibles sous licence LGPL-LR et seront mises à disposition sur le site du projet [10] dès qu'elles auront été corrigées, c'est-à-dire, nous l'espérons, au printemps 2013.

5.4 Campagne 4 : entités nommées, termes et relations en pharmacologie

5.4.1 Motivations et généralités

Cette campagne consiste à annoter un corpus de brevets (écrits en anglais) en pharmacologie pour une application d'aide au processus d'examen des brevets (aide à la lecture et recherche de similarités). Le début de campagne fut lent du fait, en particulier, d'incertitudes du client et l'annotation elle-même n'a commencé qu'au premier trimestre 2010. Elle est encore en cours à l'heure où nous écrivons ceci.

Il a été décidé d'annoter tout le brevet (dans sa partie en anglais), à l'exception de la partie état de l'art (*prior art* ou *background*). Le corpus comprend 110 brevets que nous avons répartis entre les annotatrices. Ces 110 brevets sont de tailles hétérogènes et contiennent la plupart du temps des balises XML et du texte dans d'autres langues que l'anglais. La taille totale du corpus en anglais est donc un peu difficile à évaluer, mais nous l'estimons à environ un million de tokens.

Nous nous sommes rapidement rendue compte que la difficulté principale de la tâche réside dans la nature même des brevets, qui doivent couvrir le plus grand nombre de cas possible sans pour autant donner trop de détails sur l'invention elle-même, et qui

10. http://www.irisa.fr/texmex/people/claveau/corpora/FootQuaero/

sont par conséquent rédigés de manière à rester dans le flou. Les annotatrices ont donc dû en quelque sorte « deviner » de quoi il est question, afin de pouvoir annoter.

5.4.2 Conditions d'annotation

Cette campagne d'annotation implique Jouve (JSI), en tant que client de l'annotation (demandeur du corpus annoté) et fournisseur du corpus brut, et l'INIST-CNRS, en tant qu'annotateur du corpus. Nous étions le gestionnaire de la campagne par intérim, avant que Sabine Barreaux rejoigne notre équipe et reprenne la campagne, immédiatement après la première pré-campagne. Deux ingénieures documentalistes de l'INIST-CNRS, spécialistes en pharmacologie, Anne Busin et Marie-Pierre Verdier, travaillent comme annotatrices sur le projet.

À la demande du client, qui désirait conserver le balisage d'origine des brevets (extrêmement complexe, donc difficilement adaptable dans `Glozz`), nous utilisons pour cette campagne un éditeur XML d'usage courant chez Jouve, `Epic` d'Arbortext. L'outil a été adapté par une personne de chez Jouve, ce qui a retardé le démarrage de la campagne. En particulier, cet outil a dû être adapté pour permettre l'annotation de relations à l'aide d'identifiants reliant entre eux les éléments de la relation. L'interface a demandé une période assez longue de prise en main par nos annotatrices. Il est à noter que cette interface offre la possibilité aux annotateurs d'annoter *via* la souris, mais également *via* des raccourcis clavier, ce qui permet de gagner du temps.

Soulignons aussi que nous avons demandé aux annotatrices de noter très précisément le temps passé sur la tâche. Pour cela, nous leur avons proposé d'utiliser `TimeTracker` [11]. Nous disposons ainsi de données détaillées sur leur temps d'annotation.

Après une phase de formation assez longue, du fait de difficultés d'utilisation de l'outil, nous avons pu commencer la première pré-campagne. Les annotatrices ont annoté le même brevet et nous avons comparé les résultats dans le détail, afin de mettre au clair les problèmes et de corriger le guide d'annotation. Outre les habituelles difficultés de compréhension dues à l'incomplétude du guide, de graves problèmes de frontières sont apparus, en particulier en ce qui concerne les exemples (*embodiements*).

Suite à une réunion avec le client, ces annotations d'exemples ont été abandonnées. Par contre, des catégories supplémentaires ont été ajoutées, ainsi que des « liens » (relations sémantiques). Le guide a été mis à jour par la nouvelle gestionnaire du projet. Les annotatrices ont ensuite annoté chacune cinq nouveaux brevets avec les nouvelles directives (deuxième pré-campagne). La campagne elle-même, avec le guide d'annotation stabilisé, n'a commencé qu'en mai 2011.

11. http://www.formassembly.com/time-tracker/

5.4.3 Étiquettes

Les éléments à annoter sont des éléments du domaine et se divisent principalement en termes (*pathologie*, *voie d'administration*, etc) et entités nommées (*substance active*). Le client a demandé au début d'y ajouter les exemples d'utilisation (*embodiements*), qui s'étendent parfois sur près d'un paragraphe :

> *In a preferred embodiment, the porcelains exhibit a CTE equal to or up to about 1.0 x 10-6C higher than the dental alloys to which they are applied as the opaque.*

Comme nous l'avons dit plus haut, ces annotations d'exemples ont été abandonnées après la première pré-campagne.

Le jeu d'étiquettes de cette campagne comprend donc 22 catégories, dont quatre principales, *Pathologie*, *Traitement*, *Procédé* et *Invention*, qui sont structurées comme présenté dans la figure 5.2, les catégories en italique étant celles qui sont utilisées pour l'annotation.

Quant aux relations sémantiques ajoutées suite à la première pré-campagne, elles sont au nombre de 32. Inspirées du réseau sémantique d'UMLS (*Unified Medical Language System*[12]), elles permettent de caractériser les molécules et leurs interactions (*interagit avec*, *contient*, *traite*).

5.4.4 Temps d'annotation et évaluation

L'ajout des relations a considérablement allongé le temps d'annotation, qui est passé de 9 heures en moyenne à 40 heures en moyenne pour cinq brevets (soit, respectivement, de près de deux heures à huit heures par brevet). A ce jour (septembre 2012), 40 brevets seulement ont été annotés.

Ce retard s'explique d'abord par le manque de précision de la demande du client et par la technicité des textes. Le guide d'annotation a donc été rédigé par les gestionnaires et les annotatrices et a nécessité beaucoup d'aller-retours avec le client. Ensuite, dans un souci de cohérence du projet et afin de valoriser la campagne d'annotation, la seconde gestionnaire, en accord avec le client, a proposé d'utiliser ce corpus pour l'évaluation des outils susceptibles d'intégrer la plateforme visée par le projet. Ce nouvel objectif a rendu nécessaire une révision de l'annotation de ces 40 brevets. Enfin, cette tâche d'annotation n'est pas la tâche principale des annotatrices de l'INIST-CNRS. Les contraintes qui en découlent pour leur disponibilité ont pesé sur les délais et sur l'efficacité de l'annotation, un travail trop fractionné ne permettant pas de se rapprocher de l'asymptote de la courbe d'apprentissage.

12. http://www.nlm.nih.gov/research/umls/

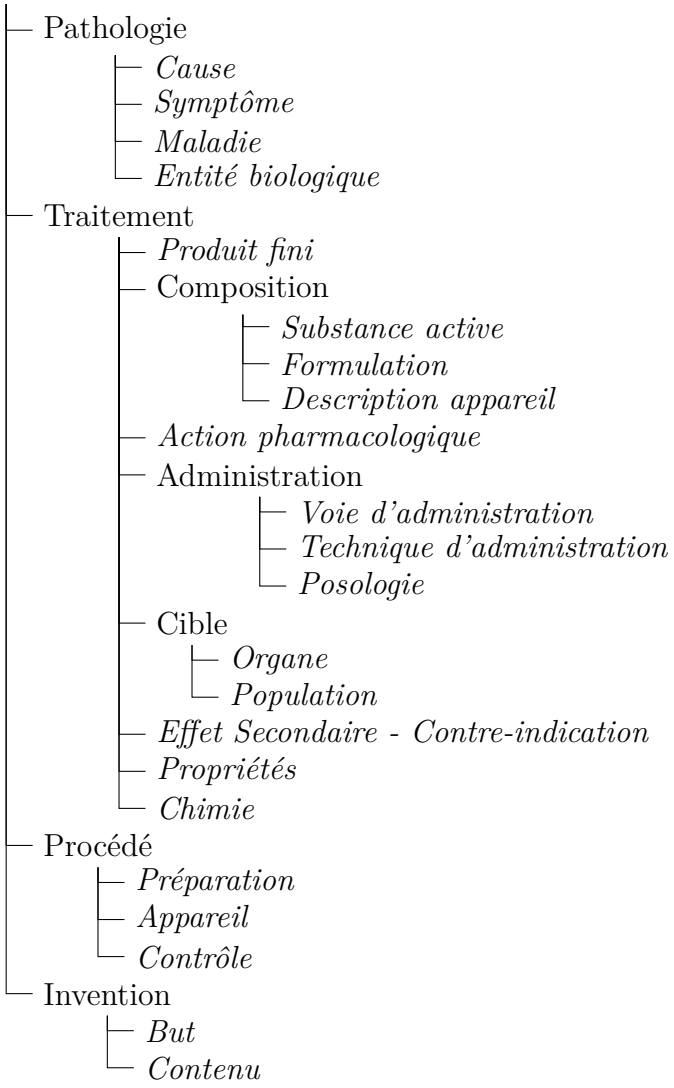


FIGURE 5.2: Structure du jeu d'étiquettes pour la campagne 4 d'annotation de brevets en pharmacologie

L'accord inter-annotateurs a ainsi été calculé à la fin de la seconde pré-campagne par la nouvelle gestionnaire, avec la mesure Glozz (voir sous-section 3.3.3) et sur un seul brevet. Il est de $0,5965$, uniquement pour les unités, les relations n'ayant pas encore pu être évaluées. Ce résultat peut être comparé, par exemple, avec les résultats obtenus lors de la campagne 3 d'annotation de matchs de football, qui ont été calculés selon la même méthode. Lors de cette campagne 3, nous avons obtenu, sur les unités et les actions (voir section 5.3), un accord inter-annotateurs situé entre $0,73$ et $0,76$, en fonction du média source (Fort et Claveau, 2012a).

Une première analyse catégorie par catégorie[13] a été menée par Sabine Barreaux et elle montre une grande hétérogénéité. Les catégories *Chimie*, *Préparation* et *Voie d'administration* semblent avoir posé particulièrement problème aux annotatrices, pour les deux premières du fait de la difficulté à décider si le segment est intéressant à annoter par rapport à la nouveauté de l'invention et pour la dernière du fait de problèmes de frontières. Suite à cette analyse, les consignes pour l'annotation ont été affinées et une relecture plus poussée et plus rigoureuse des annotations a été réalisée, avec homogénéisation systématique des divergences constatées (Sabine Barreaux, communication personnelle, le 18 septembre 2012).

5.5 Campagne 5 : entités nommées structurées

Cette campagne en recouvre en réalité deux. La première (5a) porte sur des nouvelles radio-diffusées (retranscription de journaux radiophoniques) et s'est déroulée de septembre à décembre 2010 et la seconde (5b) traite de la presse ancienne (journaux de 1890) et a commencé en février 2011 pour se terminer en juin de la même année. Les deux campagnes ont été menées dans les mêmes conditions, par les mêmes personnes, ont porté sur le même type d'annotation et ont utilisé le même guide d'annotation (Rosset *et al.*, 2011). La rédaction dudit guide a demandé six mois de travail en amont de la première campagne. Ces deux campagnes ont porté sur du français et les annotations produites ont été utilisées pour les évaluations des outils d'extraction d'entités nommées participant au programme Quæro, ainsi que pour celles de la campagne du projet ANR ETAPE[14]. Les corpus annotés seront librement disponibles d'ici la fin 2012 dans le catalogue d'ELDA[15].

Nous avons participé à ces campagnes sur la partie évaluation de l'annotation. Ces travaux ont donné lieu à des publications en commun avec les membres des équipes LIMSI-CNRS et LNE participant aux campagnes (Grouin *et al.*, 2011; Rosset *et al.*, 2012). Nous reprenons ici certains éléments de ces travaux, en les complétant.

5.5.1 Motivations et généralités

La tâche d'annotation de ces campagnes a consisté à annoter les entités nommées structurées présentes dans les corpus. Les entités nommées traditionnelles se limitent aux noms de lieux, de personnes et d'organisation (Grishman et Sundheim, 1996),

13. Cette analyse doit être considérée avec précaution, car calculer un accord sur une seule catégorie revient à l'isoler des autres catégories et donc à ne pas prendre en compte leur présence dans l'effort d'annotation. Dans le cas de la mesure Glozz, cela revient à ne prendre en compte que l'accord positionnel.

14. http://www.afcp-parole.org/etape.html

15. http://www.elda.org/

auxquels s'ajoutent souvent les dates, durées et montants. Diverses propositions d'extensions ont été faites, dont celle de Sekine (2004), qui propose une hiérarchie de près de 200 types, mais les entités traditionnelles restent les plus utilisées. Les entités nommées étendues et structurées proposées dans les campagnes présentées ici devraient apporter une plus grande richesse d'annotation, en permettant, par exemple, de distinguer le « garage Renault » de la « société Renault », tout en conservant un nombre de types limité grâce à la structuration.

Les entités nommées structurées des campagnes Quæro étendent les entités traditionnelles en leur ajoutant : (i) de nouveaux types (civilisations, fonctions, etc), (ii) des expressions construites autour de noms communs (*les pompiers*, par exemple).

En outre, ces entités sont hiérarchisées en types et sous-types (voir tableau 5.4) indiqués par une notation pointée (par exemple, *pers.ind*). Le jeu d'étiquettes comprend sept types (*person, function, organization, location, product, amount,* et *time*[16]) et 32 sous-types (Grouin *et al.*, 2011). Ces types et sous-types font référence à une segmentation du monde en catégories générales (voir tableau 5.4). La grande majorité des types voit son sens précisé à l'aide de sous-types qui soit marquent une opposition (individu *vs* groupe), soit ajoutent une précision (par exemple, pour les localisations : administrative, physique, etc.).

Il est important de noter que deux sous-types transverses ont été définis afin de rendre compte de l'ambiguïté intrinsèque aux entités nommées : (i) *other* pour les entités auxquelles l'annotateur veut associer un sous-type différent de ceux proposés et (ii) *unknown* lorsque l'annotateur n'arrive pas à déterminer quel sous-type utiliser.

Outre le premier niveau d'annotation que représentent les types et sous-types, les rédacteurs du guide ont défini un deuxième niveau d'annotation, constitué par les composants (*name* ou *kind*, par exemple), qui caractérisent les éléments à l'intérieur d'une entité nommée et ne sont pas utilisables en dehors d'un type ou d'un sous-type. Ces composants, qui peuvent être transverses ou spécifiques (voir tableau 5.5), peuvent être considérés comme des indices aidant l'annotateur à produire l'annotation en termes de catégories (par exemple, la présence d'un prénom, annoté *name.first*, est un bon indice de la présence d'un sous-type *pers.ind*) et en termes de frontières (un token donné est un indice de la présence d'une entité nommée et entre dans sa composition, alors que le token suivant n'est pas un bon indice et reste en dehors de l'entité).

16. Pour des raisons de facilité de diffusion, les auteurs du guide d'annotation (Sophie Rosset, Cyril Grouin et Pierre Zweigenbaum, du LIMSI-CNRS) ont choisi de donner des noms anglais aux catégories utilisées pour l'annotation.

<table>
<tr><th colspan="3">Person (Personne)</th><th colspan="3">Function (Fonction)</th></tr>
<tr><td colspan="2">pers.ind (individu)</td><td>pers.coll (groupe)</td><td colspan="2">func.ind (fonction d'un individu)</td><td>func.coll (corporation)</td></tr>
<tr><th colspan="3">Location (Localisation)</th><th colspan="3">Product (Produit)</th></tr>
<tr><td>administrative (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup)</td><td>physique (loc.phys.geo, loc.phys.hydro, loc.phys.astro)</td><td>bâtiment (loc.fac), route (loc.oro), adresse (loc.add.phys, loc.add.elec)</td><td>prod.object (objet manufacturé)
prod.doctr (doctrine)
prod.art</td><td>prod.serv (ligne de transport)
prod.rule (loi)
prod.media</td><td>prod.fin (produit financier)
prod.soft (logiciel)
prod.award (prix)</td></tr>
<tr><th colspan="3">Organization (Organisation)</th><th colspan="3">Time (Temps)</th></tr>
<tr><td colspan="2">org.adm (administration)</td><td>org.ent (services)</td><td colspan="2" rowspan="2">time.date.abs (date absolue), time.date.rel (date relative)</td><td rowspan="2">time.hour.abs (heure absolue), time.hour.rel (heure relative)</td></tr>
<tr><th colspan="3">Amount (Montant)</th></tr>
<tr><td colspan="3">amount (avec unité ou objet), y compris durée</td></tr>
</table>

Tableau 5.4: Types (cellules grisées) et sous-types (en italique) pour les campagnes 5a et 5b

<table>
<tr><th colspan="5">Composants transverses</th></tr>
<tr><td colspan="5">name (nom de l'entité), kind (hyperonyme de l'entité), qualifier (adjectif qualificatif), demonym (nom d'habitant ou de groupe ethnique), demonym.nickname (surnom d'habitant ou de groupe ethnique), val (nombre), unit (unité), extractor (élément dans une série), range-mark (intervalle entre deux valeurs), time-modifier (modificateur temporel).</td></tr>
<tr><th>pers.ind</th><th>loc.add.phys</th><th>time.date.*</th><th>amount</th><th>prod.award</th></tr>
<tr><td>name.last, name.first, name.middle, pseudonym, name.nickname, title</td><td>address-number, po-box, zip-code, other-address-component</td><td>week, day, month, year, century, millennium, reference-era</td><td>object</td><td>award-cat</td></tr>
</table>

Tableau 5.5: Composants transverses et spécifiques (campagnes 5a et 5b)

L'annotation manuelle de telles entités est par conséquent beaucoup plus complexe que celle des entités nommées « traditionnelles », non structurées.

Le premier corpus (nouvelles radio-diffusées) provient de radios variées, généralistes (France Inter, Europe 1), d'informations (France Info), et de réflexion (France Culture), ainsi que des radios marocaines. Il a été fourni par la DGA (projet ESTER2) et comprend 180 fichiers, soit plus de 1,2 millions de tokens.

Le second corpus (presse ancienne) provient de journaux publiés en décembre 1890, fournis par la Bibliothèque nationale de France (BnF). Il couvre 76 numéros issus de trois titres différents :
– *Le Temps*, 54 documents, soit 209 pages,
– *La Croix*, 21 documents, soit 84 pages,
– et *Le Figaro*, 1 document, soit 2 pages.
Ces journaux ont été numérisés par Jouve et une procédure de reconnaissance optique de caractères leur a été appliquée[17]. Après filtrage des éléments non pertinents pour l'annotation d'entités nommées, comme les publicités ou les tableaux de chiffres, (pour plus de détails sur la procédure de filtrage, voir (Galibert *et al.*, 2012)), le corpus brut final compte 1 673 460 tokens.

5.5.2 Conditions d'annotation

Ces campagnes d'annotation ont impliqué principalement le LIMSI-CNRS, en tant que gestionnaire de la campagne, le LNE, sur la partie évaluation, et ELDA, en tant que fournisseur d'annotations. Les clients sont en l'occurrence les participants aux campagnes d'évaluation de Quæro, soit le LIMSI-CNRS, Jouve et Synapse Développement. Un gestionnaire a été désigné au sein d'ELDA, qui a interagi directement avec les responsables du projet au LIMSI-CNRS (Cyril Grouin, Sophie Rosset et Pierre Zweigenbaum).

L'annotation a été réalisée à ELDA sur l'éditeur `XEmacs`[18], agrémenté d'un module spécifique, développé par Olivier Galibert, du LNE. Ce module, prévu au départ uniquement pour un usage « interne » aux gestionnaires de la campagne, a finalement été adopté par ELDA. Cet outil a le mérite de permettre une annotation très rapide, grâce à une manipulation par le clavier. Ses fonctionnalités sont cependant limitées : il ne permet pas, par exemple, l'ajout de commentaires. En outre, les balises XML posées par les raccourcis clavier sont modifiables dans le texte, ce qui peut générer des erreurs. Enfin, l'outil est permissif et laisse passer les chevauchements interdits. Pour compenser ces défauts, un outil de normalisation a été développé par Olivier Galibert pour vérifier la cohérence des balises XML. Cet outil permet également d'ajouter automatiquement les composants *name* que les annotateurs avaient l'autorisation de ne pas annoter s'ils étaient les seuls composants de l'entité annotée.

Pour les deux campagnes d'annotation, ELDA a fait travailler quatre annotateurs, encadrés par un responsable.

Au final, la première campagne (nouvelles radio-diffusées) a permis l'annotation d'environ 200 000 entités nommées et la seconde 149 055.

17. La qualité de la reconnaissance est globalement bonne : taux d'erreur de caractère de 5,09 % et taux d'erreur de mot de 36,59 % (Rosset *et al.*, 2012).
18. http://www.xemacs.org/

5.5.3 Évaluation

Le corpus de nouvelles radio-diffusées, fait rare, a été annoté intégralement en double (chaque fichier est annoté par 2 annotateurs), voire en quadruple, chaque fichier étant évalué, corrigé, puis ré-évalué.

En ce qui concerne l'annotation du corpus de presse ancienne, seuls quelques fichiers ont été annotés en double (20 % d'entre eux). Cette double annotation a été suivie d'un relevé des erreurs systématiques et des procédures de correction semi-automatiques ont ensuite été mises au point et appliquées sur l'ensemble du corpus annoté.

Dans les deux cas, ELDA a utilisé les outils d'évaluation du LNE pour aligner et évaluer les annotations (pour optimiser leur correction). Les métriques utilisées pour cette comparaison sont les suivantes : F-mesure, Slot Error Rate (avec ordre variable) et Correct (rappel).

À la fin de la première campagne, Sophie Rosset a corrigé un fichier annoté du corpus (sélectionné aléatoirement parmi les fichiers de grande taille de la campagne) plusieurs fois, jusqu'à obtenir une parfaite cohérence intra-annotateur, puis l'a utilisé comme référence pour évaluer l'annotation ELDA. Le résultat sur la campagne a été très satisfaisant, avec un SER inférieur à 5 % [19].

Par ailleurs, un sous-corpus de 400 phrases (soit environ 11 400 tokens) a été extrait de manière aléatoire du corpus brut annoté et a été ré-annoté par 4 personnes : Cyril Grouin, Sophie Rosset (LIMSI-CNRS), Sabine Barreaux (INIST-CNRS) et nous-même. Tous les fichiers ont été annotés par tous. Les résultats ont ensuite été comparés, puis corrigés et fusionnés, afin d'obtenir la meilleure référence possible (voir figure 5.3). Cette « mini référence » nous a permis de compléter la validation des annotations réalisées par ELDA. Les temps d'annotation et d'adjudication cumulés représentent près de 90 heures de travail (voir (Grouin *et al.*, 2011) pour le détail).

La même méthode de construction de mini-référence a été utilisée lors de la deuxième campagne, avec les mêmes annotateurs-experts, à l'exception de Sabine Barreaux, remplacée par Flora Badin (INIST-CNRS). L'échantillon de corpus annoté utilisé pour former cette mini-référence a cette fois été sélectionné de manière à être le plus représentatif possible [20] et comprend 12 263 tokens. Cette fois-ci, les temps d'annotation et d'adjudication cumulés ont été un peu plus élevés (une centaine d'heures en tout).

Ces deux campagnes d'annotation ont vu l'application de notre méthodologie dans son ensemble. Les annotateurs d'ELDA ont en effet été formés à l'outil et à la tâche. Une mini-référence a par ailleurs été annotée par des experts de la tâche et a servi

19. Cette expérience n'a pas pu être reproduite lors de la campagne suivante.

20. La sélection a été réalisée par Aurélien Bossard (LIMSI-CNRS) et a consisté en des tirages aléatoires dont la représentativité a été évaluée par la divergence de Kullback-Leibler (Kullback et Leibler, 1951) des distributions des taux de présence de chaque entité et de la fréquence de chaque couple (source, numéro de page) par rapport aux distributions du corpus brut.

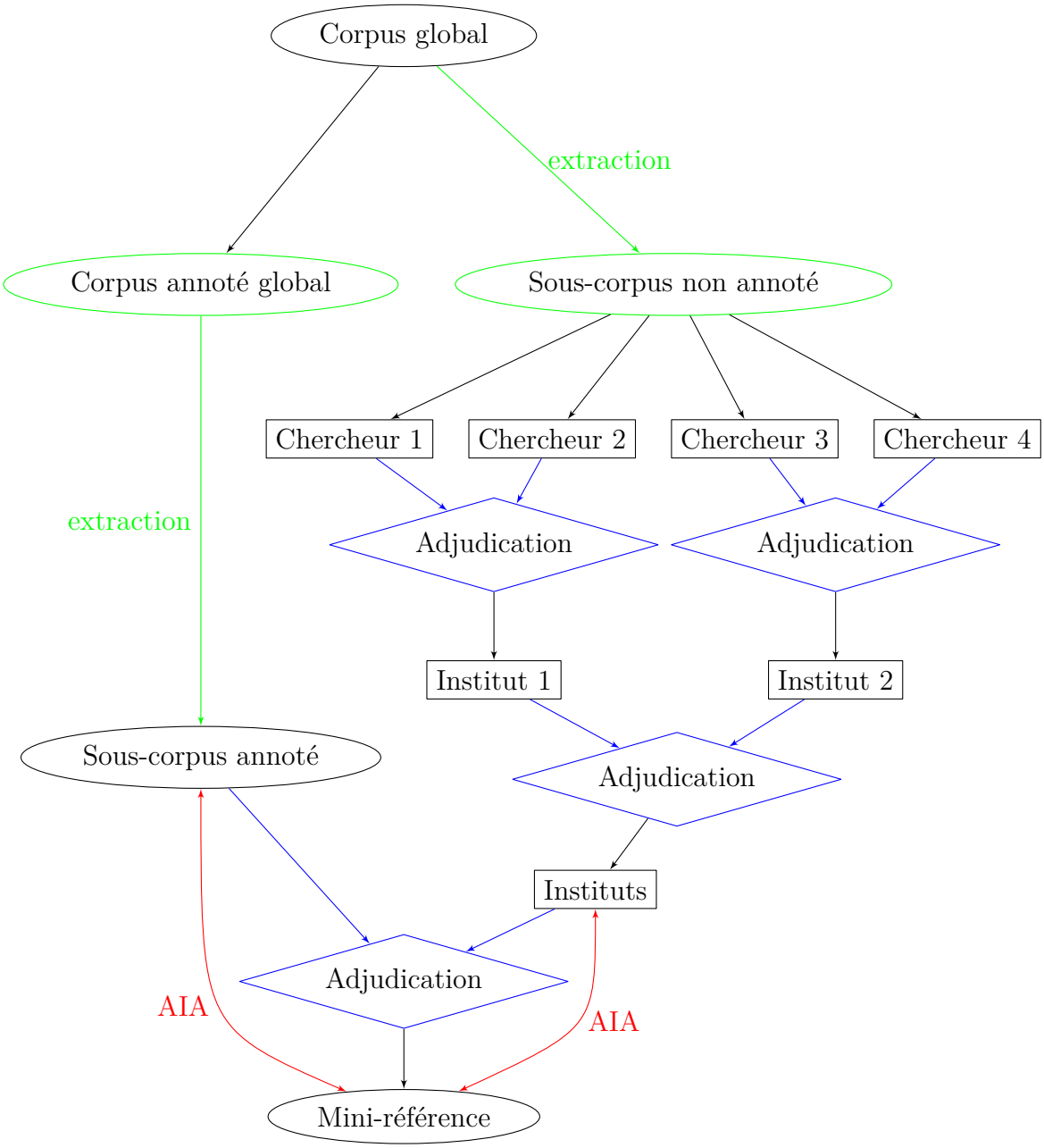


FIGURE 5.3: Création et utilisation de la « mini référence » dans les campagnes 5 (en vert, phase d'extraction, en bleu, adjudication, et en rouge, calcul de l'accord inter-annotateurs)

à l'évaluation finale des campagnes. La première campagne était déjà très avancée lorsque nous avons suggéré d'utiliser cette méthode, nous n'avons donc pas pu l'intégrer dans l'évaluation au fil de l'eau des annotations produites. Le mode d'annotation choisi

(en double, avec évaluation sur chaque fichier) a permis de compenser cette absence. Lors de la deuxième campagne, l'annotation de la mini-référence a eu lieu en parallèle de l'annotation générale, principalement pour des raisons de temps : la création de la mini-référence nous a pris une centaine d'heures et la campagne ne pouvait pas attendre. Cette seconde campagne était par ailleurs un prolongement de la première campagne et ne nécessitait donc aucune formation complémentaire des annotateurs (les mêmes, pratiquement, que lors de la première campagne). Notre grille d'analyse des dimensions de complexité de la campagne n'était alors qu'embryonnaire et nous n'avons pu l'appliquer qu'*a posteriori* (Fort *et al.*, 2012b).

5.6 Difficultés rencontrées

Les campagnes sur lesquelles nous avons travaillé ont permis de mettre en évidence certaines difficultés, de types très variés, que nous détaillons ici.

5.6.1 Problèmes liés à la gestion du projet

Un projet d'annotation manuelle de corpus est un projet comme un autre, et, à ce titre, on y rencontre les mêmes types génériques de difficultés. Parmi ceux-ci, on peut citer les questions financières (voir campagne 3), les problèmes de communication avec le client, liés à sa disponibilité (voir campagnes 3 et 4), enfin, les problèmes liés à la disponibilité du gestionnaire du projet (campagnes 3 et 4). Nous ne présenterons pas ces difficultés, trop génériques, en détails.

Il nous semble en revanche important de mettre l'accent sur les problèmes liés aux acteurs de la campagne d'annotation. Une première question concerne la sélection des annotateurs. Si nous n'avons pas eu de mal à trouver à l'INIST-CNRS des experts en microbiologie pour les campagnes 1 et 2, ou en pharmacologie, pour la campagne 4, il nous a été paradoxalement plus difficile de trouver des experts en football, et nous avons dû faire une sélection parmi les volontaires, pour éliminer les personnes dont la qualité d'annotation était trop faible. En outre, dans une structure telle que l'INIST-CNRS, une fois les annotateurs sélectionnés, encore faut-il parvenir à leur libérer du temps pour travailler sur le projet. Le fait de devoir intercaler cette tâche parmi les tâches classiques des annotateurs a été pour nous un problème lourd à gérer dans toutes les campagnes menées à l'INIST-CNRS (campagnes 1, 2, 3 et 4). Ce problème de priorité est à associer à un manque de lien hiérarchique direct entre le gestionnaire de la campagne et les annotateurs.

D'autres problèmes liés aux acteurs des campagnes d'annotation sont apparus, concernant cette fois la définition de leur rôle. Ainsi, au début de la campagne 1 (voir section 5.1), le gestionnaire de la campagne était, de fait, le client, et cette fusion des deux

rôles a eu un impact négatif sur la campagne, la communication avec les annotateurs ayant été sous-estimée. En effet, leur malaise face aux catégories définies par le client et orientées en fonction des outils d'annotation automatique a été sous-estimé, et la qualité de l'annotation en a pâti. Enfin, les outils d'annotation utilisés dans les campagnes 2 et 4 ont été choisis par les clients, or, ces outils étaient inadaptés à la tâche (le premier, `Cadixe`, ne permet pas d'annoter des relations directement et le second, `Epic`, n'est tout simplement pas un outil d'annotation) et ont posé des problèmes d'utilisation aux annotateurs, qui ont dû palier les insuffisances des outils (ajout d'identifiants pour la campagne 2 et adaptation à une interface complexe pour la campagne 4). Si le choix d'un outil n'est jamais évident (voir l'état de l'art à ce sujet en section 3.2), le gestionnaire de la campagne est en général la personne la mieux placée pour le faire.

5.6.2 Définition des catégories

La définition des catégories est, de notre point de vue, l'étape la plus difficile d'une campagne d'annotation.

Une première difficulté consiste à « faire accoucher » le client de ses besoins, en particulier lorsque ceux-ci ne sont pas clairement définis, souvent parce que le client connaît mal l'application finale et ne parvient pas à cerner ses besoins (campagnes 3 et 4).

Lorsque le client pense savoir ce dont il a besoin, le gestionnaire de la campagne doit s'assurer que les objectifs de l'annotation ne sont pas contradictoires et faire établir des catégories correspondant, dans les termes utilisés comme dans la réalité de ce qui est recherché, à l'application visée. Cette difficulté est apparue clairement dans la campagne 1 (Fort *et al.*, 2009).

Nous avons également dû faire face, dans la campagne 3, à des difficultés de définition des catégories liées aux médias sources. En effet, une partie du corpus concerne des transcriptions de commentaires de matchs. Or, ces commentaires sont très elliptiques, et comportent en particulier peu de verbes. Il nous a donc fallu définir des ancres d'annotation permettant d'annoter ces ellipses. Ainsi, l'ancre des actions (tir, hors-jeu, ...) et des relations (passe, changement de joueur, ...) est posée sur l'acteur et non sur le prédicat lui-même. Mais ces ancres ne sont pas sans poser des problèmes aux annotateurs, qui ont mis du temps à comprendre la logique de ce choix.

Enfin, lors de la seconde campagne d'annotation en entités nommées structurées (presse ancienne numérisée), nous avons constaté la présence d'erreurs de reconnaissance de caractères ainsi que la présence de césures due au format d'origine du journal. Afin de prendre en compte ces caractéristiques et de permettre aux annotateurs d'annoter les entités, même mal numérisées ou incluant une césure, un nouvel attribut (*correction*) et un nouveau composant (*noisy-entity*, entité bruitée) ont été introduits dans le schéma d'annotation. Le premier permet aux annotateurs de corriger les erreurs de reconnaissance de caractères dans les entités annotées : la correction est ajoutée

dans l'attribut *correction* (au niveau le plus général de l'annotation) et le texte d'origine n'est pas modifié, comme dans l'exemple suivant :

```
<pers.ind correction="Le Moine">
 <name.last> LE Moibte. </name.last>
</pers.ind>
```

Quant au composant *noisy-entity*, il permet aux annotateurs de signaler la présence d'une entité dans une chaîne de caractères mal segmentée, comme dans l'exemple suivant, qui présente également une correction :

```
   <loc.adm.reg  correction="EN  ALSACE-
LORRAINE">
  <noisy-entities>
   KN_ALSACE'LOBR4INE
  </noisy-entities>
 </loc.adm.reg>
```

5.6.3 Biais de la pré-annotation

Lors de la campagne 1, les annotateurs ont reconnu avoir été influencés par la pré-annotation des noms de gènes et de protéines et avoir favorisé la correction de ces pré-annotations, au détriment de la recherche de nouvelles entités. Nous avons donc essayé de prendre en compte, dans les campagnes suivantes, ce biais dû à la pré-annotation, en particulier en sensibilisant les annotateurs au problème, à la fois oralement et dans le guide d'annotation. Mais la question de la définition et de l'impact de ce biais est encore loin d'être réglée.

5.6.4 Prise en compte du contexte

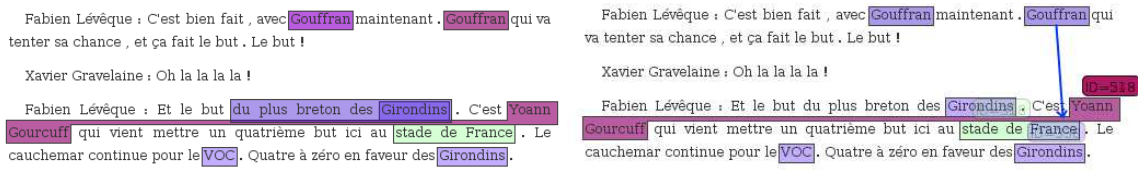
Nous avons rencontré des difficultés liées au contexte nécessaire à l'annotation dans les campagnes d'annotation de renommage de gènes (campagne 2), de matchs de football (campagne 3) et de presse ancienne (campagne 5b).

Dans la campagne 2 d'annotation de relations de renommage de gènes (voir section 5.2), les annotateurs ont souvent dû consulter l'article PubMed entier pour pouvoir en annoter le résumé, le contexte de celui-ci étant parfois insuffisant pour comprendre s'il y avait bien renommage. Or, cet accès à l'article n'était pas prévu au début de la

campagne et aurait pu bloquer celle-ci si l'INIST-CNRS ne disposait pas d'un accès aux articles complets *via* ses services Web.

Dans le même ordre d'idée, nous avons constaté tardivement, lors de la campagne 5b d'annotation d'entités nommées structurées dans un corpus de presse ancienne, que la visualisation de l'image de l'article de journal d'origine était indispensable pour ne pas commettre d'erreur d'annotation. En effet, le texte numérisé pouvait comprendre un « M. Buis » alors qu'il s'agissait en fait de « M. Buls » ou, plus grave, « touché » au lieu de « Fouché », ce qui empêche l'annotation d'une entité. Or, l'outil d'annotation utilisé avait été prévu pour la campagne 5a d'annotation de nouvelles radio-diffusées et ne permettait pas l'affichage simultané du texte numérisé et de l'image. Cela a fait perdre du temps aux annotateurs, qui ont sans doute dû renoncer à visualiser toutes les images. Par ailleurs, si les annotateurs étaient formés pour la tâche d'annotation, ce ne sont pas des experts du domaine historique et ils ont dû accéder à des sources externes pour décider, par exemple, si dans « le krach Macé », « Macé » est un nom de famille ou non, ou si le « Tonkin » était un pays ou une simple région en 1890.

Enfin, dans la campagne 3 d'annotation de matchs de football, le contexte nécessaire à l'annotation était parfois relativement large et s'étendait au-delà du paragraphe, comme le montre l'exemple présenté dans la figure 5.4. Dans cet exemple, l'annotateur 1 n'a pas vu qu'il y a une passe entre Gouffran et Gourcuff et que c'est ce dernier qui marque et non Gouffran (cet annotateur a en fait annoté deux buts), ce que l'annotateur 2, lui, a vu (la passe est indiquée par une flèche bleue et l'action de marquer par une couleur pourpre). Ces difficultés sont dues aux nombreuses ellipses caractéristiques des commentaires vidéo. Bien que l'outil Glozz, utilisé pour la campagne, soit adapté à l'annotation de relations entre éléments distants, il ne peut pas compenser un manque de concentration de l'annotateur ou un contexte trop elliptique pour être annoté.



(A) Annotation de l'annotateur 1 (B) Annotation de l'annotateur 2

FIGURE 5.4: Impact du contexte sur l'annotation dans la campagne 3 (football)

Dans notre expérience, l'importance du contexte est donc trop souvent sous-estimée et il est fondamental d'évaluer correctement les besoins des annotateurs en la matière lors de la pré-campagne, afin de prévoir de leur donner accès aux sources nécessaires lors de l'annotation.

5.6.5 Évaluation

Nous avons rencontré différentes difficultés d'évaluation dans toutes les campagnes auxquelles nous avons participé.

La première de ces difficultés a été d'évaluer une campagne *a posteriori* (voir section 5.1), autrement dit, une fois la campagne terminée, sans qu'aucune mesure n'ait été prise pour calculer un quelconque accord inter-annotateurs et sans qu'aucune référence n'existe. Nous avons, dans ce cas précis, utilisé les marqueurs d'incertitude posés par les annotateurs *a posteriori*. Les résultats de cette évaluation sont présentés dans (Fort *et al.*, 2009) et détaillés dans la section 4.4.2.

Dans les autres campagnes, les questions se sont posées plus tôt et sont liées aux caractéristiques même de l'annotation. Ainsi, dans la campagne 2, nous avons dû faire face au problème, relativement bien connu (Artstein et Poesio, 2008), mais non résolu, de la grande prévalence d'une catégorie. Nous avons analysé les accords inter-annotateurs obtenus à l'aide de plusieurs coefficients et en fonction de différents choix d'unités annotables, puis nous avons proposé un moyen de synthèse permettant de quantifier la similitude entre les catégories (Fort *et al.*, 2012a). Nous présentons ce travail plus en détails dans la sous-section 8.3.1.

Dans les campagnes 3, 4 et 5, nous avons dû évaluer des annotations dans un contexte où les unités annotables ne sont pas facilement définies (entités nommées, termes). Nous avons mené des réflexions à ce sujet dans le cadre des campagnes 3 et 5, avec, en particulier, les gestionnaires de la campagne 5 (LIMSI et LNE) et Vincent Claveau pour la campagne 3. Nous avons ainsi montré l'influence sur les accords inter-annotateurs du choix des annotables et proposé un algorithme permettant de les définir de manière optimale (voir chapitre 8).

Enfin, dans les campagnes 2 et 3, il nous a fallu évaluer des relations, plus ou moins complexes. Dans le premier cas (renommage de gènes), nous nous sommes pour cela ramenée à un cas plus simple de langage de type, avec des identifiants comme traits (Fort *et al.*, 2012a). En ce qui concerne la campagne d'annotation de matchs de football, nous avons décidé d'évaluer les relations par un Kappa, en ne permettant aucun défaut dans l'annotation, ni de localisation, ni de caractérisation (voir section 8.3.3).

L'état de l'art de l'évaluation de l'annotation manuelle, présenté en section 3.3, ne fournit pas de solution directe à ces différents problèmes, raison pour laquelle nous avons mené un travail de caractérisation des complexités de l'annotation manuelle (voir chapitre 6) que nous couplons à des propositions concernant l'évaluation de celle-ci (voir chapitre 8).

Chapitre 6

Analyser la complexité d'une campagne d'annotation

Ce chapitre reprend un article écrit en commun avec Adeline Nazarenko et Sophie Rosset (LIMSI-CNRS) (Fort *et al.*, 2012b).

Dans le travail d'annotation, l'annotateur doit déterminer ce qu'il faut annoter et comment annoter les unités qui doivent l'être. Mesurer la complexité d'une tâche d'annotation nécessite d'analyser ces opérations de localisation et de caractérisation (Widlöcher et Mathet, 2009).

Nous proposons ici d'analyser cette complexité selon six dimensions : les deux premières (la discrimination et la délimitation) sont relatives à la localisation des annotations, tandis que les trois suivantes concernent le travail de caractérisation (le pouvoir d'expression, la dimension du jeu d'étiquettes et le degré d'ambiguïté). Le sixième facteur de complexité à peser sur les décisions d'annotation est le poids du contexte : nous le présentons comme une dimension supplémentaire par souci de simplicité même s'il affecte en réalité tout à la fois la discrimination, la délimitation des frontières et la désambiguïsation.

Analyser une tâche selon ces six dimensions est artificiel au sens où les annotateurs ne décomposent jamais leurs décisions de cette manière, mais cette approche analytique est utile à la gestion de l'annotation. Elle est indépendante à la fois du volume d'annotations à ajouter et du nombre d'annotateurs impliqués : ces valeurs participent au coût d'une tâche d'annotation, mais pas à sa complexité.

6.1 Décomposition d'une tâche d'annotation

L'effort d'annotation et le coût d'une tâche d'annotation dépendent bien entendu du type des annotations attendues et de la complexité du flux de données source, mais il est important pour la planification du travail d'évaluer la complexité de la campagne au plus tôt dans le processus d'annotation.

Annoter manuellement consiste, pour un ou plusieurs annotateurs humains, à expliciter leur interprétation d'un signal source. Le guide d'annotation détaille généralement le type d'interprétation attendu et son but. En fonction de la tâche, les annotateurs ont à leur disposition un jeu d'étiquettes fermé ou peuvent ajouter les étiquettes qu'ils

jugent utiles. Ils doivent lire le document source et étiqueter tout ou partie de ses segments à l'aide d'une ou plusieurs étiquettes.

Plutôt que de décomposer les tâches d'annotation en niveaux (*levels*) ou en couches (*layers*) (Goecke *et al.*, 2010), nous proposons, pour analyser la complexité d'une tâche d'annotation, de décomposer celle-ci en tâches d'annotation élémentaires (TAE). La complexité des différentes TAE est calculée indépendamment et la complexité d'une tâche entière est la combinaison de la complexité des différentes tâches élémentaires.

Définition 2 (Tâche d'annotation élémentaire (TAE)) *Une tâche d'annotation élémentaire est une tâche d'annotation qui n'est pas décomposable. Nous considérons qu'une tâche d'annotation peut être décomposée en au moins deux TAE si le jeu d'étiquettes lui-même est décomposable en jeux d'étiquettes plus réduits et indépendants. Nous entendons ici par indépendants que les étiquettes de deux différents jeux d'étiquettes sont globalement compatibles (et ce, même si certaines combinaisons particulières sont interdites), alors que les étiquettes d'un jeu unique sont mutuellement exclusives (mis à part les besoins d'encodage de l'ambiguïté).*

Cette décomposition en TAE est formelle au sens où elle est indépendante de l'organisation pratique du travail : les annotateurs peuvent s'acquitter de différentes TAE en différentes étapes d'annotation sur le signal source ou toutes à la fois, selon la nature du travail à effectuer et des outils utilisés.

Cette décomposition en TAE ne résulte pas d'une simplification de la tâche d'origine comme c'est souvent le cas pour les Human Intelligence Tasks (HIT) effectuées par les *Turkers* (travailleurs) sur la plate-forme Amazon Mechanical Turk (voir par exemple (Cook et Stevenson, 2010)).

Pour prendre un exemple simple, l'annotation de relations de renommage de gènes peut être analysée comme une combinaison de deux TAE. La première correspond à l'identification des noms de gènes dans le signal source et la seconde repose sur ce premier niveau d'annotation et consiste à indiquer lesquels des noms de gènes entrent dans une relation de renommage. Bien entendu, les annotateurs peuvent ajouter les deux types d'annotations dans un même élan, mais les jeux d'étiquettes sont indépendants et il est plus simple, d'un point de vue formel, d'analyser cette tâche d'annotation comme une combinaison de deux TAE plutôt que comme une seule, qui serait beaucoup plus complexe.

6.2 Quoi annoter ?

Localiser l'unité à annoter est en soi une opération complexe. Nous distinguons ici la discrimination qui consiste à distinguer ce qui est à annoter de ce qui ne doit pas l'être et la délimitation des frontières qui intervient quand il faut retoucher les

frontières des parties discriminées. Nous verrons que ces deux opérations reposent sur une segmentation de référence qui fixe ce qui peut être annoté, c'est-à-dire qui détermine quelles sont les unités « annotables ».

6.2.1 Discrimination

Dans certaines expériences d'annotation, la question « quoi annoter ? » est toute résolue parce que le corpus est préparé de telle sorte que le choix des unités à annoter est déjà fait. Dans certains cas, les éléments à annoter sont faciles à repérer dans le flux textuel parce qu'ils ont été préalablement marqués, lors d'une phase de pré-annotation automatique ou d'une première passe d'annotation manuelle. Par exemple, dans l'expérience rapportée par (Weissenbacher et Nazarenko, 2005) pour la classification des occurrences du pronom *it* en impersonnelles et anaphoriques, il fallait un corpus d'entraînement et de test annoté : les occurrences du pronom étaient déjà marquées et le travail d'annotation consistait simplement à marquer chacune de ces occurrences comme impersonnelle ou anaphorique. La discrimination ne pose pas non plus de problème quand toutes les unités sont à annoter, comme dans une tâche d'étiquetage morpho-syntaxique où tous les mots du texte doivent recevoir une étiquette de partie du discours (voir par exemple (Marcus *et al.*, 1993)).

Pourtant, bien souvent, le corpus est, pour l'annotateur, une botte de paille qu'il doit fouiller pour trouver quoi annoter et la tâche d'annotation suppose de discriminer ce qui doit être annoté de ce qui ne doit pas l'être.

Identifier les unités sur lesquelles cibler le travail d'annotation est d'autant plus complexe que l'ensemble des unités à considérer est mal défini, les unités étant de taille et de nature diverses. Erk *et al.* (2003) et Widlöcher et Mathet (2009) soulignent ainsi que l'annotation des rôles sémantiques et l'annotation discursive mélangent différents niveaux de découpage et portent sur des unités qui peuvent être inférieures au mot ou supérieures à la phrase. Pour prendre un exemple simple, il est plus facile d'identifier dans un texte les adverbes connotés négativement que toutes les expressions négatives parce que ces dernières peuvent être des mots, des syntagmes, des tournures de phrase voire des passages entiers de texte. Dans le premier cas, l'ensemble des adverbes à considérer est d'autant plus facile à repérer que le texte a pu être préalablement étiqueté en parties du discours. Dans le second cas, le nombre d'unités susceptibles d'être annotées ou « annotables » est beaucoup plus élevé car plusieurs niveaux de découpage du texte sont à prendre en compte.

Nous pouvons donc définir une première échelle de difficulté.

Définition 3 (Discrimination) *Une tâche d'annotation est considérée comme d'autant plus difficile que le facteur de discrimination défini par la formule suivante est*

plus élevé :

$$Discrimination_a(F) = 1 - \frac{|A_a(F)|}{\sum_{i=1}^{n} |D_i(F)|}$$

où F *est le flux de données à annoter,* a *est une tâche d'annotation,* n *est le nombre de niveaux de segmentation potentiellement pertinents,* $|D_i(F)|$ *est le nombre d'unités obtenues lors du découpage de* F *au niveau* i *et* $|A_a(F)|$ *est le nombre d'unités à annoter dans la tâche d'annotation considérée.*

De manière intuitive, cette mesure indique que le poids de discrimination est d'autant plus élevé que les unités à annoter sont « noyées » au milieu des autres et donc que la proportion de ce qui est à annoter ($|A_a(F)|$) par rapport à ce qui est annotable ($\sum_{i=1}^{n} |D_i(F)$) est plus faible. Le facteur de discrimination vaut 0 quand toutes les unités du texte sont à annoter, et tend vers 1 quand peu d'unités sont à annoter alors que beaucoup sont annotables.

Pour la classification des occurrences de pronoms, le facteur de discrimination est à 0 si les occurrences à annoter ont toutes été identifiées au préalable. En revanche, pour le renommage de noms de gènes (voir section 5.2), le facteur de discrimination est plus élevé parce qu'une faible proportion de couples de noms de gènes entrent dans une relation de renommage et sont à annoter comme tels [1].

On peut généralement estimer facilement le nombre d'éléments « à annoter » à partir de la définition de la tâche d'annotation, sur un échantillon de corpus annoté ou par référence à des tâches comparables, mais il est souvent plus difficile d'estimer ce qui est annotable parce que cela suppose de choisir une segmentation de référence qui découpe le texte à annoter en unités, certaines devant être annotées et d'autres pas. Il y a plusieurs manières de choisir cette segmentation de référence :

1. La solution la plus simple consiste à choisir la segmentation qui s'approche le plus de l'annotation à faire, du moment qu'elle peut être pré-calculée automatiquement ou qu'elle est intuitive pour l'annotateur, quitte à retoucher les frontières des unités à annoter qui ne correspondent pas à ce découpage. Dans l'exemple des entités nommées, partir d'un découpage en mots amène à considérer les entités nommées composées comme des unités « retouchées ». Cette stratégie réduit le poids de discrimination aux dépens de la délimitation des frontières (voir section 6.2.2 ci-après).
2. Quand les unités à annoter sont trop hétérogènes, on peut considérer plusieurs segmentations de référence ($n > 1$ dans la formule ci-dessus) : cela augmente d'autant le nombre des annotables mais évite d'avoir beaucoup de frontières

1. On suppose ici que les noms de gènes sont pré-annotés (ou identifiés dans une tâche d'annotation élémentaire différente) et que tout couple de nom de gènes figurant dans un même résumé est annotable, c'est-à-dire susceptible de traduire une relation de renommage.

à retoucher. On privilégie cette approche si elle paraît moins coûteuse que la précédente. Dans le cas des entités nommées, cela reviendrait à considérer tous les mots et tous les syntagmes comme annotables.

3. On peut enfin décomposer la tâche d'annotation en différentes couches correspondant à des tâches d'annotation élémentaires distinctes, ce qui permet d'analyser le poids de discrimination de chaque couche de manière indépendante, chacune ayant sa propre segmentation de référence. Une telle décomposition en couches ne s'impose cependant que lorsque les différents types d'annotables qui résultent des différentes segmentations s'annotent différemment. À l'inverse, il serait artificiel d'analyser le cas des entités nommées comme différentes couches d'annotation dès lors que c'est le même jeu d'étiquettes qui sert pour les mots et les syntagmes.

6.2.2 Délimitation des frontières

Identifier des points d'intérêt dans le flux textuel ne suffit pas parce que les éléments à annoter sont souvent des « segments de données » (Habert, 2005). Pour savoir quoi annoter, il faut également délimiter les frontières du segment à annoter.

Là encore la tâche est simple lorsqu'une segmentation de référence fiable est calculable. Une pré-annotation permet alors de fixer les frontières des unités identifiées comme étant à annoter sans que la charge de la délimitation soit laissée à l'annotateur. Les segmentations calculables automatiquement sont cependant souvent assez approximatives et l'annotateur doit retoucher localement les frontières des unités discriminées : pour annoter des entités nommées ou des termes, on peut partir d'un découpage en mots mais il faut corriger cette segmentation pour toutes les unités polylexicales.

Dans la majorité des cas, délimiter les frontières consiste à étendre ou rétrécir l'unité discriminée sur la base de la segmentation de référence mais il peut aussi s'agir de décomposer une unité discriminée en plusieurs éléments ou de regrouper plusieurs unités discriminées contiguës en une seule annotation. Il existe des cas, plus rares, où les unités à étiqueter sont discontinues (un exemple bien connu est celui de la négation en français). La solution consiste généralement soit à prendre l'unité englobante, soit à étiqueter indépendamment les différents constituants de l'unité à annoter, soit encore à établir une relation entre ces différents composants. Ces solutions permettent toutes d'éviter de considérer des unités discontinues en tant que telles.

Définition 4 (Délimitation) *La délimitation des frontières représente un deuxième facteur de complexité, $Délimitation_a(F)$, qui se calcule de la manière suivante :*

$$Délimitation_a(F) = min(\frac{S + I + D}{|A_a(F)|}, 1)$$

où $|A_a(F)|$ est le nombre d'unités discriminées au final, I *le nombre d'unités ajoutées, obtenues par décomposition d'une unité initiale,* D *le nombre d'unités supprimées par*

regroupement d'unités initiales, et S *est le nombre de substitutions, c'est-à-dire d'unités discriminées qui ont vu leurs frontières changer en dehors des cas de décomposition et regroupement précédents.*

La délimitation se calcule donc en comparant la segmentation obtenue à l'issue de l'annotation avec la segmentation de référence. Nous nous inspirons pour cela du Slot Error Rate (Makhoul *et al.*, 1999) qui est une mesure usuellement utilisée en évaluation des systèmes de reconnaissance et de classification des entités nommées, qui tient compte des frontières (voir sous-section 3.3.4).

Le facteur de délimitation vaut 0 quand aucune unité discriminée n'a été modifiée et il augmente avec le nombre de décompositions, regroupements et modifications de frontières que l'annotateur fait par rapport à la segmentation de référence. Comme le dénominateur est donné par le nombre d'unités discriminées au final, le rapport $\frac{S+I+D}{|A_a(F)|}$ peut théoriquement être supérieur à 1, d'où la borne introduite dans la formule.

Le coût de délimitation se mesure *a posteriori*, mais il peut être estimé sur la base d'un échantillon du corpus annoté à produire ou par comparaison avec une tâche similaire.

6.3 Comment annoter ?

Une fois les unités discriminées et délimitées, une deuxième question se pose : comment les annoter ? Après les avoir localisées, il faut caractériser les annotations et la complexité de cette opération dépend elle aussi de plusieurs critères. La difficulté de la tâche de caractérisation croît avec la richesse du langage d'annotation utilisé, qui est elle-même fonction du degré d'expressivité (sous-section 6.3.1) et de la taille du vocabulaire (la « dimension du jeu d'étiquettes », sous-section 6.3.2) de ce dernier. La difficulté de la tâche d'annotation dépend aussi du degré d'ambiguïté des unités à annoter (sous-section 6.3.3).

6.3.1 Expressivité du langage d'annotation

Les annotations à apposer sur les unités qui ont été discriminées et délimitées peuvent être de différents types. Nous distinguons trois types de langage, auxquels correspondent différents degrés d'expressivité. La complexité induite par l'expressivité du langage d'annotation forme naturellement une échelle qualitative ordinale, mais par souci d'homogénéité avec les facteurs précédents qui se décrivent par des valeurs numériques, nous associons les différents niveaux d'expressivité à des graduations sur une échelle numérique allant de 0 à 1.

Définition 5 (Degrés d'expressivité du langage d'annotation) *Les degrés d'expressivité du langage d'annotation sont les suivants :*
- $0,25$ *: correspond aux langages de types,*
- $0,5$ *et* $0,75$ *: correspondent aux langages relationnels, respectivement d'arité 2 et d'arité supérieure à 2,*
- 1 *: réservé aux langages d'ordre supérieur.*

Dans le cas le plus simple, le langage d'annotation est un langage de type : annoter consiste à associer un type à un segment de données, c'est-à-dire à l'étiqueter. Beaucoup de tâches d'annotation reposent sur cette catégorie de langage : on associe aux mots d'un texte des parties du discours, aux tours de parole des interlocuteurs ou des fonctions rhétoriques, à des syntagmes des types d'entités nommées ; on relève les phrases ou des passages pertinents, les occurrences de pronoms anaphoriques, etc. Dans certains cas, l'étiquette utilisée est elle-même structurée comme dans les étiquettes morpho-syntaxiques associant une partie du discours à un lemme et des traits morpho-syntaxiques mais cela augmente la dimension du jeu d'étiquettes (voir section 6.3.2) sans changer le degré d'expressivité du langage d'annotation qui reste un langage de types.

Établir des relations entre des unités est également une tâche assez courante mais elle est plus complexe. Relier des pronoms anaphoriques à leurs antécédents, marquer des couples de noms de gènes comme étant des renommages, identifier des relations de dépendances syntaxiques ou expliciter des relations temporelles entre verbes nécessite en effet de mettre en relation des segments de données différents qui sont souvent eux-mêmes typés. Ces relations sont en outre souvent typées et orientées. On essaye généralement de se ramener à des relations binaires qui sont plus faciles à appréhender mais des relations d'arité supérieure sont parfois nécessaires : on a notamment des langages d'annotations relationnels d'arité supérieure à 2 dans les tâches d'annotation dédiées à l'extraction d'information où il faut repérer les différents acteurs et circonstants d'un événement (qui a acheté quoi, quand, à qui et à quel prix ?). Il ne faut pas négliger la complexité des annotations relationnelles : même si les annotateurs ne procèdent pas toujours ainsi, tout se passe comme s'il fallait une première passe d'annotation pour repérer et typer les arguments de la relation puis discriminer les couples, triplets et plus généralement les n-uplets de segments à annoter parmi l'ensemble des n-uplets annotables et enfin typer la relation qu'entretiennent les éléments du n-uplet.

On fait appel à un langage d'ordre supérieur dès lors qu'on appose des annotations sur d'autres annotations mais la complexité de ce type de langage difficile à formaliser et à manipuler fait qu'on se ramène souvent à un autre cas de figure. S'il s'agit de qualifier une annotation (par exemple, la marquer comme incertaine), on préfère en général augmenter la dimension du jeu d'étiquettes en considérant le qualificatif comme un attribut associé à un type principal. La décomposition d'une tâche d'annotation en

couches permet aussi de qualifier ou de relier dans une passe ultérieure des annotations apposées dans une passe antérieure : on a alors plusieurs couches d'annotation mais chacune fait appel à un langage d'ordre 1.

Cette représentation, même si elle est grossière et en partie arbitraire, permet de représenter les différentes dimensions de la complexité d'une tâche d'annotation de manière similaire.

6.3.2 Dimension du jeu d'étiquettes

Que l'étiquette soit un type ou une relation, qu'elle soit d'ordre 1 ou d'ordre supérieur, l'annotateur a toujours à choisir la valeur d'une annotation dans un jeu prédéfini d'étiquettes, ce qui constitue un nouveau facteur de complexité : le choix est d'autant plus difficile qu'il est plus ouvert.

Dans le cas le plus simple, le choix est booléen et l'annotation revient à répartir les unités discriminées en deux catégories [2] : on peut marquer des phrases comme pertinentes ou non, les occurrences du pronom *it* comme anaphoriques ou impersonnelles, une relation anaphorique comme fiable ou incertaine.

Le choix est cependant souvent plus ouvert, notamment quand il s'agit de représenter finement la diversité des unités morpho-syntaxiques (parties du discours et traits morpho-syntaxiques), d'annoter des dépendances syntaxiques ou de typer des entités nommées. Ce jeu d'étiquettes est généralement défini dans le guide d'annotation. Pour les annotations les plus riches, on propose souvent des étiquettes structurées : tout se passe comme si l'annotateur posait plusieurs étiquettes sur une même unité, la combinaison de ces étiquettes formant son annotation (Dandapat *et al.*, 2009).

Il existe enfin des tâches d'annotation pour lesquelles le choix d'une étiquette est totalement ouvert pour l'annotateur, comme pour la transcription de la parole où il peut y avoir autant d'étiquettes que de mots de vocabulaire. Nous considérons dans ces cas-là que l'on a un jeu d'étiquettes de très grande taille, même si l'effort d'annotation est probablement de nature un peu différente pour l'annotateur.

Si une annotation A est formée d'une séquence de m étiquettes ($A = E_1 E_2 \dots E_m$) et que chaque étiquette E_i peut prendre n_i valeurs différentes, le jeu d'étiquettes complet comporte théoriquement n étiquettes différentes, avec $n = n_1 * n_2 * \dots * n_m$. En pratique cependant, des contraintes sont définies qui réduisent cette combinatoire : l'annotateur

2. Cette tâche d'annotation booléenne peut se confondre avec la tâche de discrimination qui revient elle aussi à faire une classification binaire pour identifier les unités à annoter de celles qui ne le sont pas dans l'ensemble des annotables mais il s'agit ici de caractériser les unités à annoter par une étiquette et non pas de les repérer. A noter qu'une tâche de discrimination peut toujours s'analyser comme un première tâche d'annotation élémentaire avec une discrimination nulle (tous les annotables sont à annoter) et un jeu d'étiquettes binaire (*à annoter* ou *à ne pas annoter*).

n'a pas à choisir 1 étiquette parmi n en une fois mais une étiquette parmi n_1 dans un premier temps, puis 1 parmi au plus n_2, etc. jusqu'à en choisir 1 parmi au plus n_m. La dimension du jeu d'étiquettes ne dépend donc pas du nombre total d'étiquettes possibles mais des degrés de liberté des choix successifs que l'annotateur a à faire.

Définition 6 (Degré de liberté) *Le degré de liberté global ν pour le choix d'une étiquette composée de* m *sous-étiquettes est donné par la formule suivante :*

$$\nu \leq \nu_1 + \nu_2 + \ldots + \nu_m$$

où ν_i est le degré de liberté maximal que l'annotateur a dans le choix de la i^{eme} sous-étiquette ($\nu_i = n_i - 1$)[3].

Le jeu d'étiquettes pour l'annotation en morpho-syntaxe du *Penn Treebank* contient 36 étiquettes (Santorini, 1990), on a donc théoriquement un degré de liberté $\nu = 35$, mais comme certaines étiquettes sont des sous-types d'autres étiquettes (comme JJR et JJS pour JJ, l'adjectif) et que le nombre de choix maximal pour ces sous-types est de 6 (pour les verbes), on peut considérer que $\nu = 20 + 5 = 25$.

A partir de là, on peut définir un quatrième facteur de complexité, la dimension du jeu d'étiquettes.

Définition 7 (Dimension du jeu d'étiquettes) *Cette dimension est calculée selon la formule suivante :*

$$Dimension_a(F) = min(\frac{\nu}{\tau}, 1)$$

où ν est le degré de liberté global que l'annotateur a dans le choix d'une étiquette simple ou complexe pour une tâche d'annotation a sur un flux de données F, et τ est le seuil à partir duquel on considère le jeu d'étiquettes comme étant arbitrairement grand. Dans les expériences détaillées ci-après, τ vaut 50, ce qui, d'après les retours des annotateurs, fait déjà beaucoup.

La dimension du jeu d'étiquettes vaut 0 pour les jeux d'étiquettes binaires présentant un degré de liberté de 1 et elle croît avec la taille du jeu d'étiquettes et donc le degré de liberté. Elle vaut $0,5$ pour la tâche d'annotation morpho-syntaxique du *Penn Treebank* ($0,7$, si on considère le jeu d'étiquettes comme non hiérarchique). Elle plafonne à 1.

Les tâches d'annotation avec de très grands jeux d'étiquettes ($\nu \geq \tau$) sont très difficiles à gérer.

3. La formule donne une borne haute du degré de liberté global parce que, le choix de la i^{eme} étiquette étant souvent contraint par les étiquettes déjà posées, l'annotateur a en pratique un degré de liberté moindre que $(n_i - 1)$, si n_i est le nombre d'étiquettes disponibles à ce point de choix.

6.3.3 Degré d'ambiguïté

La nécessité de désambiguïser les unités lors de l'annotation introduit un cinquième facteur de difficulté. Ce dernier est cependant plus difficile à estimer que les précédents : le rôle de l'annotateur étant précisément de résoudre les ambiguïtés lorsque cela est possible, les ambiguïtés sont difficilement observables. On peut cependant approcher le degré d'ambiguïté que l'annotateur a à résoudre pour une tâche donnée de deux manières.

La première méthode consiste à mesurer le degré d'ambiguïté résiduelle en observant les traces que l'annotateur a laissé de son travail de désambiguïsation : à supposer que le protocole d'annotation l'y autorise, l'annotateur peut annoter avec plusieurs étiquettes faute de savoir choisir l'une d'entre elles, utiliser une étiquette sous-déterminée quand il ne sait pas caractériser de manière précise (par exemple, ne pas noter de trait de genre pour un adjectif épicène ambigu) ou marquer son hésitation sous la forme d'un attribut d'incertitude associé à l'étiquette choisie. Cela permet de mesurer le degré d'ambiguïté résiduelle :

Définition 8 (Degré d'ambiguïté résiduelle) *Le degré d'ambiguïté résiduelle d'une tâche d'annotation se mesure de la manière suivante :*

$$Ambiguité_{Res,a}(F) = \frac{|Annot_a|}{|Annot|}$$

où a *et* F *sont la tâche d'annotation et le flux de données à considérer et où* $|Annot_a|$ *et* $|Annot|$ *sont respectivement le nombre d'annotations portant une marque d'ambiguïté et le nombre total d'annotations posées sur* F*.*

Par définition, ce degré d'ambiguïté résiduelle ne peut se mesurer qu'*a posteriori*, une fois l'annotation effectuée ou à partir d'un échantillon de celle-ci.

Ce degré d'ambiguïté résiduelle vaut 0 quand aucune marque d'ambiguïté n'a été posée par l'annotateur et vaudrait 1 dans le cas (en réalité absurde) où toutes les annotations seraient marquées comme ambiguës d'une manière ou d'une autre.

Il est évident qu'en fonction du type de traces utilisé pour le calcul (traits d'incertitudes, étiquette *autre* ou *inconnu*, etc.) et des recommandations données aux annotateurs, cette mesure d'ambiguïté peut être plus ou moins fiable et il faut si possible l'associer à des résultats obtenus avec une autre méthode de calcul. En outre, les annotateurs n'utilisent en général ce type de traces que dans les cas le plus problématiques [4]. Ainsi, au cours de la campagne 1 (voir section 5.1) les annotateurs ont

4. Le cas de l'évaluation a *posteriori* dans la campagne 1 est particulier, puisqu'il était demandé explicitement aux annotateurs de se concentrer sur les ambiguïtés.

sous-utilisé les attributs d'incertitude qui leur étaient proposés. Cette mesure donne donc une estimation basse du degré d'ambiguïté réel.

Une seconde méthode consiste à mesurer le degré d'ambiguïté théorique pour les tâches où on annote plusieurs occurrences des mêmes unités de vocabulaire ou vocables : cela s'applique à l'étiquetage morpho-syntaxique ou la désambiguïsation sémantique mais pas à l'analyse des tours de parole ou au renommage de noms de gènes.

Cette mesure repose sur l'idée que les vocables ambigus ont des occurrences annotées différemment à différents endroits du flux de données. Il suffit alors de mesurer la proportion d'unités à annoter correspondant à des vocables ambigus. L'ambiguïté théorique peut se mesurer à partir d'un dictionnaire qui liste les étiquettes possibles pour tous les vocables quand l'annotation repose sur un tel dictionnaire ou, directement, sur un échantillon de texte annoté. Elle dépend aussi de la fréquence des vocables ambigus dans le flux de données à annoter.

Définition 9 (Degré d'ambiguïté théorique) *Le degré d'ambiguïté théorique se mesure de la manière suivante :*

$$Ambiguité_{Th,a}(F) = \frac{\sum_{i=1}^{|Voc(F)|} (Ambig_a(i) * freq(i,F))}{|Unités_a(F)|}$$

avec

$$Ambig_a(i) = \begin{cases} 1 & si \quad |Étiquettes_a(i)| > 1 \\ 0 & sinon \end{cases}$$

où Voc *est le vocabulaire des unités du flux de données* F, $|Voc(F)|$ *la taille de ce vocabulaire,* $freq(i,F)$ *la fréquence du vocable* i *dans* F, $|Unités_a(F)|$ *le nombre d'unités à annoter dans* F *et* $|Étiquettes_a(i)|$ *le nombre d'étiquettes que peut prendre le vocable* i *pour la tâche d'annotation* a.

Lorsqu'aucun vocable n'est ambigu, $|Étiquettes_a(i)|$ vaut 1 et $Ambig_a(i)$ vaut 0 pour tout i, la tâche d'annotation est triviale et peut être automatisée car il suffirait de projeter sur le texte un vocabulaire établissant la correspondance entre les vocables et leurs étiquettes. Dans ce cas $Ambiguité_{Th,a}(F)$ vaut 0. A l'inverse, si tous les vocables sont ambigus, $Ambiguité_{Th,a}(F)$ vaut 1. On notera que le poids d'un vocable ambigu compte d'autant plus dans le degré d'ambiguïté théorique qu'il est plus fréquent.

L'ambiguïté théorique tend à surestimer le poids de l'ambiguïté pour l'annotateur car on peut supposer que la résolution de certaines ambiguïtés est triviale. On a donc en général, pour une tâche d'annotation a et un flux de données F :

$$Ambiguité_{res,a}(F) < Ambiguité_a(F) < Ambiguité_{Th,a}(F)$$

dans la mesure où les deux types d'ambiguïtés peuvent être calculés.

6.4 Le poids du contexte

Le poids du contexte est un sixième facteur de complexité dans l'analyse des tâches d'annotation. Même si ce n'est pas réellement un facteur indépendant des précédents comme ces derniers l'étaient entre eux (la nécessité de prendre en compte le contexte vient compliquer les tâches de discrimination, délimitation et de désambiguïsation pour l'annotateur), nous le représentons comme tel, par souci de simplicité.

La complexité d'une tâche d'annotation augmente avec la taille de la fenêtre de texte à prendre en compte autour de l'unité à annoter et avec le nombre de connaissances à mobiliser. Même s'il est difficile de déterminer le nombre de mots qui entrent dans la résolution d'une tâche d'annotation et, plus encore, le nombre de connaissances en jeu [5], on peut repérer deux échelles qualitatives :

- La taille de l'empan de texte autour de l'unité à annoter : l'unité toute seule, son environnement immédiat (quelques mots avant et après), la phrase ou le paragraphe, voire le texte dans son entier.
- Le degré d'accessibilité des sources de connaissances qui sont consultées : l'annotateur peut ne consulter aucune source extérieure, il peut utiliser le matériel d'annotation (par exemple le guide d'annotation) ou des sources prédéfinies comme des nomenclatures, mais il peut aussi rechercher de nouvelles sources de connaissances (par exemple consulter des tiers ou faire des recherches sur Internet).

Définition 10 (Poids du contexte) *Le poids du contexte dépend de la taille de l'empan de texte à prendre en compte et du degré d'accessibilité des sources à consulter. Par souci d'homogénéité avec les critères de complexité précédents, nous traduisons ces deux échelles qualitatives dans une même échelle discrète allant de* 0 *à* 1*, représentée sur la figure 6.1.*

Les cas distingués sont les suivants :

- *La valeur* 0 *correspond aux cas où aucun empan de texte autour de l'unité à annoter et aucune connaissance ne sont mis en jeu.*
- *La valeur* 1 *est appliquée aux cas les plus complexes où l'éclairage du texte dans son ensemble et la consultation de sources de connaissances extérieures (par exemple Wikipedia) sont nécessaires pour annoter les unités.*
- *La valeur* $0,25$ *correspond à deux cas de figure : 1) si l'annotation ne dépend que de l'environnement immédiat de l'unité à annoter 2) ou s'il faut consulter des sources mises à disposition comme le guide d'annotation.*
- *La valeur* $0,5$ *est appliquée dans trois cas : 1) si les deux difficultés précédentes se conjuguent, 2) si un empan de texte plus large (la phrase, dans toute sa complexité)*

5. Encore faudrait-il en effet que ces connaissances soient dénombrables, ce qui n'est évidemment pas toujours le cas.

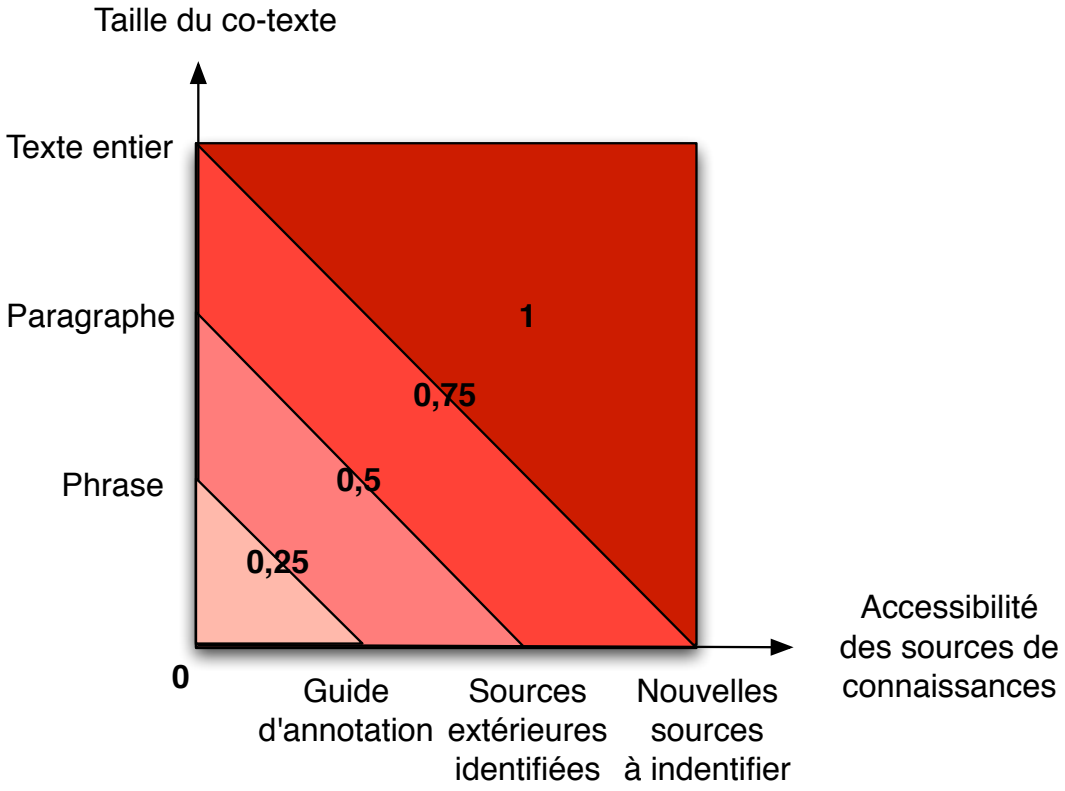


FIGURE 6.1: Poids du contexte

doit être pris en compte, 3) ou s'il faut consulter une source de connaissances extérieures préalablement bien identifiée (une nomenclature, comme la base UniProt[6]*, par exemple).*

– *Enfin, on estime à* $0,75$ *la difficulté dans trois cas : 1) si l'annotateur doit à la fois regarder toute la phrase et consulter des ressources extérieures pour poser des annotations, 2) s'il doit prendre en compte le texte dans son ensemble, 3) ou s'il doit rechercher de nouvelles sources de connaissances.*

La valeur 0 correspond à un cas de figure théorique, car une tâche d'annotation dans laquelle le contexte n'entrerait aucunement en compte devrait être automatisée. À l'inverse, l'annotation des relations de renommage entre noms de gènes correspond à la valeur 1, parce que la relation de renommage, qui est souvent peu marquée dans le texte, se confond souvent avec une relation d'isotopie (ressemblance) ou d'appartenance à une même famille. Il faut alors lire le résumé entier pour déterminer la sémantique de la relation et il arrive que les annotateurs consultent une source extérieure pour mieux comprendre les propriétés des gènes et éclairer leur décision.

6. http://www.uniprot.org/uniprot/

Cette échelle est évidemment simplificatrice, mais il est important de prendre ce facteur en compte lors de la préparation d'une campagne d'annotation et les critères présentés ici ont pour but de guider l'analyse et de faciliter la comparaison entre dimensions de complexité.

6.5 Synthèse

Les six facteurs de complexité étant normalisés sur une même échelle, il est facile, une fois les dimensions de complexité de différentes tâches analysées et calculées, de représenter ces différentes dimensions sur un diagramme radar (ou en toile d'araignée) [7].

Considérons par exemple la tâche simple, déjà mentionnée, qui consiste à classer les occurrences de pronoms comme étant impersonnelles ou anaphoriques (Weissenbacher et Nazarenko, 2005). Les pronoms étant pré-annotés, la discrimination et la délimitation valent 0. Le jeu d'étiquettes n'en contient que deux, cette dimension est donc également à 0 et l'expressivité du langage d'annotation est de $0,25$ (langage de types). Cependant, le degré d'ambiguïté est élevé (1) car toutes les occurrences de pronoms sont ambiguës. Notre expérience personnelle de ce type d'annotation nous enseigne que le contexte vaut $0,5$ (voire davantage), car il faut prendre en compte toute la phrase pour comprendre le rôle des pronoms. La complexité de cette tâche est représentée graphiquement sur la figure 6.2.

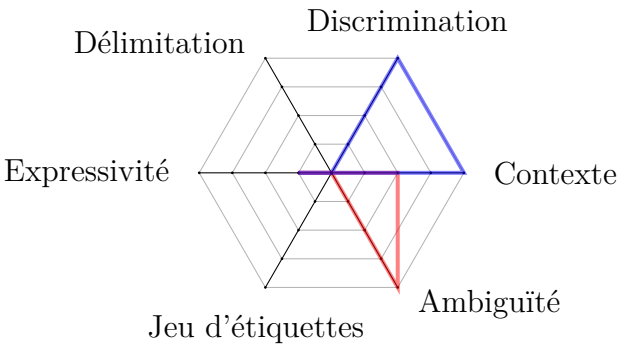


FIGURE 6.2: Synthèse des dimensions de complexité des campagnes de classification des pronoms (rouge) et d'annotation de noms de gènes (bleu)

Le cas du renommage de gènes (Jourde *et al.*, 2011) est plus complexe et est plus facilement analysé comme une combinaison de deux TAE. Représenter chaque TAE sur un graphique séparé ne permet pas de visualiser la complexité de la tâche dans son ensemble. Nous proposons donc de représenter toutes les TAE d'une tâche sur un

7. Les résultats présentés sous forme de diagrammes radars dans cette section sont arrondis à l'entier le plus proche.

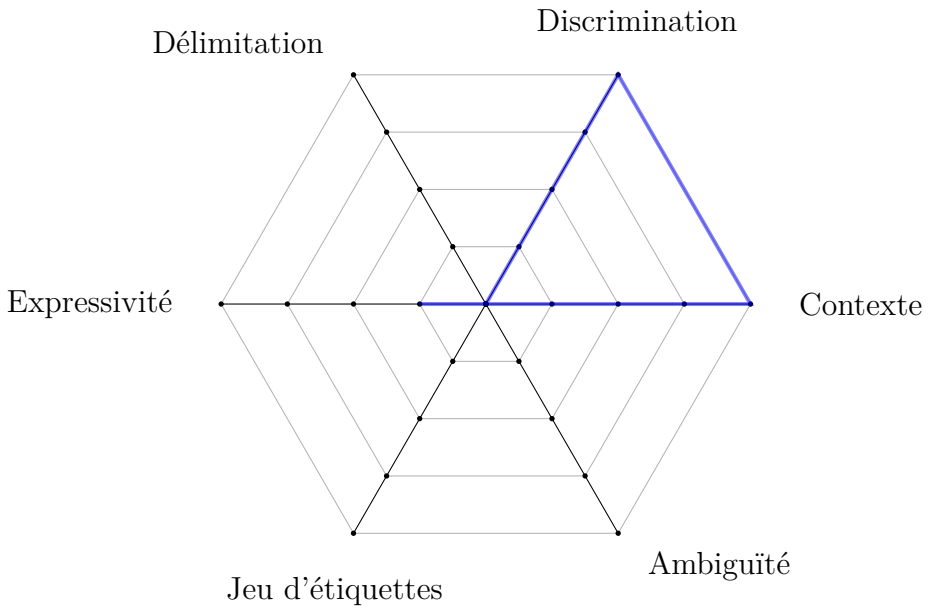


FIGURE 6.3: Synthèse des dimensions de complexité de la campagne de renommage de gènes (2 TAE, échelle x2)

graphique unique, dont la taille doit être proportionnellement élargie, comme montré dans la figure 6.3.

Dans le cas qui nous occupe, la première TAE est l'étiquetage de noms de gènes dans la séquence de mots qui compose le texte. La discrimination est élevée $(0,9)$, ce qui s'explique par le fait que peu de mots sont en fait des noms de gènes. La délimitation est de 0, car les noms de gènes, dans notre campagne, ne sont que des tokens simples. A l'opposé, les facteurs de caractérisation reçoivent des valeurs peu élevées : le jeu d'étiquettes est booléen ($Dimension = 0$), on utilise un langage de types ($Expressivité = 0,25$) et l'ambiguïté est très faible, puisque peu de noms de gènes sont également des noms communs (l'ambiguïté théorique est approximativement de $0,01$ [8] et l'ambiguïté résiduelle est en moyenne pour les deux annotateurs de $0,04$ [9]). Le contexte est ici le facteur de complexité le plus important, car les annotateurs ont souvent eu besoin de lire tout le résumé PubMed pour comprendre le rôle d'une entité (Fort *et al.*, 2010) et ils ont dû consulter des ressources externes, le poids du contexte est donc de 1.

Cette première TAE est représentée sur le même graphique que la campagne de classification des pronoms, ce qui permet de les comparer (voir figure 6.2). Si les deux

8. Ce résultat a été obtenu en comptabilisant les noms de gènes trouvés dans nos textes qui sont également présents dans le corpus Brown.

9. Ce résultat, qui contredit l'inégalité $Ambiguité_{res,a}(F) < Ambiguité_{Th,a}(F)$, est dû à une surannotation des incertitudes par l'un des annotateurs, qui manquait de confiance en lui sur la campagne.

campagnes montrent peu de complexité sur trois des six dimensions (la délimitation, l'expressivité du langage et la dimension du jeu d'étiquettes), la première (classification des pronoms) présente une ambiguïté élevée et aucun problème de discrimination, alors que la seconde (noms de gènes) pose des problèmes de discrimination et de contexte, et pas d'ambiguïté. Les solutions pour alléger le coût de ces campagnes doivent donc être adaptées (par exemple, pré-annotation et accès facilité au contexte pour les noms de gènes et documentation bien conçue pour la classification des pronoms).

Le renommage de gènes consiste non seulement à identifier les noms de gènes, mais également à relier entre elles les occurrences de ceux-ci qui portent une relation de renommage. La tâche comprend donc une seconde TAE, qui consiste à marquer les relations de renommage ou leur absence sur tous les couples de noms de gènes apparaissant dans un même résumé. Dans notre cas, la discrimination est élevée $(0,95)$, car peu de couples de noms de gènes sont effectivement des renommages. La délimitation est nulle, puisque les noms de gènes sont déjà annotés. Le jeu d'étiquettes est composé de trois étiquettes puisque la relation de renommage est orientée (lien de gauche à droite, lien de droite à gauche, absence de lien), sa dimension est donc de $0,04$. Même si les annotations sont porteuses d'informations relationnelles, le langage est un langage de types (un couple de noms de gènes est étiqueté par une catégorie qui exprime la direction de la relation). L'ambiguïté est très faible (l'ambiguïté résiduelle est en moyenne pour les deux annotateurs de $0,02$), mais le poids du contexte est élevé, comme pour la première TAE de la campagne.

La figure 6.3 montre comment la combinaison des deux TAE permet d'obtenir une visualisation globale de la complexité de la tâche, qui porte principalement sur la discrimination et le contexte.

6.6 Conclusion

Nos expériences de gestion et de participation à diverses campagnes d'annotation nous ont montré que l'amélioration de l'efficacité de l'annotation manuelle de corpus passe par une meilleure connaissance du processus d'annotation. Les décompositions d'une campagne d'annotation en phases et en dimensions de complexité que nous proposons, et que nous avons décrites dans cette partie, posent un cadre qui devrait permettre de mieux maîtriser la gestion d'une campagne d'annotation.

Nous présentons dans la suite de la thèse différents outils pour faciliter le travail du gestionnaire de campagne, en complétant et en instanciant certaines des solutions méthodologiques proposées ici.

Troisième partie

Outiller le gestionnaire

Chapitre 7

Pré-annoter ou ne pas ?

Des outils de TAL sont depuis longtemps utilisés pour améliorer l'efficacité de l'annotation manuelle, notamment par la pré-annotation automatique de corpus (voir, par exemple (Marcus *et al.*, 1993)). Nous avons vu en section 3.4.1 que si un certain nombre d'études se sont penchées sur les effets de cette pré-annotation automatique, peu en ont évalué les biais. Par ailleurs, il restait à établir le seuil de qualité de l'outil utilisé pour cette pré-annotation, à partir duquel on peut prévoir que celle-ci sera bénéfique.

Nous avons donc travaillé, en collaboration avec Benoît Sagot (INRIA/Paris 7), sur l'évaluation précise du gain en temps et en qualité (*via* l'exactitude et l'accord inter-annotateurs) obtenu sur le travail purement manuel, en utilisant différentes qualités d'outils de pré-annotation (en l'occurrence, des étiqueteurs morpho-syntaxique) (Fort et Sagot, 2010). Nous avons en particulier mis au point le protocole expérimental décrit en section 7.1. Nous reprenons ici le contenu de ce travail en le complétant sur certains sujets, notamment la courbe d'apprentissage des annotateurs.

Nous présentons ensuite rapidement en section 7.3 certaines expériences complémentaires que nous avons menées dans le cadre d'une autre campagne d'annotation, concernant du français parlé spontané (Benzitoun *et al.*, 2012).

7.1 Protocole expérimental

Afin de maîtriser tous les paramètres de notre expérience, nous avions besoin d'une référence, sur laquelle entraîner notre outil et comparer nos annotations. Nous avons pour cela utilisé le *Penn Treebank* (Marcus *et al.*, 1993), dont les sections 2 à 21 ont servi à entraîner des outils de pré-annotation de différente qualité et la section 23 a été utilisée pour l'annotation manuelle et la correction. Nous avons donc entraîné notre étiqueteur morpho-syntaxique sur des sous-parties du *Penn Treebank* de tailles croissantes et avons pré-annoté la section 23 avec les différents étiqueteurs ainsi obtenus. Nous avons ensuite annoté ou corrigé manuellement des parties de la section 23 dans différentes conditions expérimentales que nous décrivons dans cette section.

7.1.1 Création des étiqueteurs morpho-syntaxiques

Nous avons choisi d'utiliser l'étiqueteur morpho-syntaxique MElt (Denis et Sagot, 2009), un système librement disponible [1] fondé sur le maximum d'entropie capable de prendre en compte des informations extraites à la fois d'un corpus d'entrainement et d'un lexique morphologique externe. Entrainé sur les sections 2 à 21 du *Penn Treebank* ($\text{MElt}_{\text{en}}^{\text{ALL}}$), sans lexique externe, et évalué sur la section 23, MElt fait preuve d'une exactitude de 96,4 %, ce qui est proche du niveau de l'état de l'art pour l'anglais (Spoustová *et al.* (2009) ont annoncé 97,4 % et Ratnaparkhi (1996) 96,6 %, sans lexique).

Benoît Sagot a entraîné MElt sur des sous-parties de tailles croissantes du *Penn Treebank* annoté en morpho-syntaxe, créant ainsi différents étiqueteurs de différents niveaux de qualité (exactitude obtenue sur la référence). Ces étiqueteurs, $\text{MElt}_{\text{en}}^{i}$, entrainés sur les i premières phrases des sections 2 à 21 du *Penn Treebank*, sont présentés dans le tableau 7.1. Benoît Sagot a ensuite pré-annoté toute la section 23 avec chacun de ces étiqueteurs, ce qui nous a permis d'obtenir, pour chaque phrase, un ensemble de pré-annotations.

Étiqueteur	Nb phrases (entrainement)	Tokens	Exactitude (%)
$\text{MElt}_{\text{en}}^{10}$	10	189	66,5
$\text{MElt}_{\text{en}}^{50}$	50	1 254	81,6
$\text{MElt}_{\text{en}}^{100}$	100	2 774	86,7
$\text{MElt}_{\text{en}}^{500}$	500	12 630	92,1
$\text{MElt}_{\text{en}}^{1000}$	1 000	25 994	93,6
$\text{MElt}_{\text{en}}^{5000}$	5 000	126 376	95,8
$\text{MElt}_{\text{en}}^{10000}$	10 000	252 416	96,2
$\text{MElt}_{\text{en}}^{\text{ALL}}$	37 990	944 859	96,4

TABLEAU 7.1: Exactitude obtenue par les étiqueteurs sur la section 23 du *Penn Treebank*

7.1.2 Expériences

Nous avons mis au point différentes expériences afin d'évaluer l'impact de la pré-annotation et en particulier de sa qualité, sur la qualité du corpus corrigé final.

1. MElt est disponible sous licence LGPL sur le site du projet http://gforge.inria.fr/projects/lingwb/.

Conditions d'annotation

Benoît Sagot et nous-même avons été les annotateurs dans ces expériences. Nous considérons que nous avons une bonne connaissance de la langue, mais que nous avons peu de connaissance de l'annotation morpho-syntaxique du *Penn Treebank*, même si Benoît Sagot avait une expérience préalable de l'annotation morpho-syntaxique. Il est aussi important de noter que bien que nous parlions relativement couramment anglais, ce n'est pas notre langue maternelle. Par ailleurs, nous estimons que les conditions expérimentales étaient suffisamment rigoureuses pour limiter l'influence de notre connaissance de celles-ci sur les résultats (voir sous-sections suivantes 7.1.2 et 7.1.2).

Nous avons noté nos temps d'annotation (ou de correction) par série de 10 phrases. Nous avons également décidé de n'utiliser qu'un simple éditeur de texte, sans macro ou fonctionnalité particulière susceptible d'accélérer l'annotation, en dehors des habituels `Rechercher-Remplacer`. Aucune interface ne nous permettait donc d'alléger la charge cognitive que représentent les 36 étiquettes et les nombreux cas particuliers décrits dans le guide d'annotation (Santorini, 1990).

Courbe d'apprentissage

Comme nous l'avons déjà mentionné, la formation des annotateurs influence de manière significative la qualité de l'annotation et permet un gain de temps important (Marcus *et al.*, 1993; Dandapat *et al.*, 2009; Mikulová et Štěpánek, 2009). Ainsi, Marcus *et al.* (1993) ont observé que les annotateurs avaient besoin d'un mois pour être pleinement efficaces sur la tâche de correction d'annotations morpho-syntaxiques du *Penn Treebank*, atteignant une vitesse de correction de 20 minutes pour 1 000 mots.

Notre vitesse de correction ne peut pas être comparée à la leur, puisque nos expériences n'ont porté que sur des échantillons du corpus. Pour autant, notre vitesse et notre exactitude se sont améliorées avec la pratique. La figure 7.1 montre notre courbe d'apprentissage, présentant en abscisse l'identifiant des phrases corrigées et en ordonnée le temps de correction. Nous y constatons clairement l'influence de la formation sur le temps de correction.

Nous avons par conséquent pris soin d'éviter les problèmes liés à la phase d'apprentissage rencontrés par Rehbein *et al.* (2009) en mélangeant les phrases pré-annotées par les différents outils de pré-annotation (de $\text{MElt}_{\text{en}}^{10}$ à $\text{MElt}_{\text{en}}^{\text{ALL}}$).

Expériences

Les sous-parties de la section 23 utilisées pour ces expériences sont identifiées par les numéros de phrases correspondant. Ainsi, 1–100 désigne les 100 premières phrases de la section 23. Lors de chaque expérience, les phrases ont été annotées (ou corrigées) de

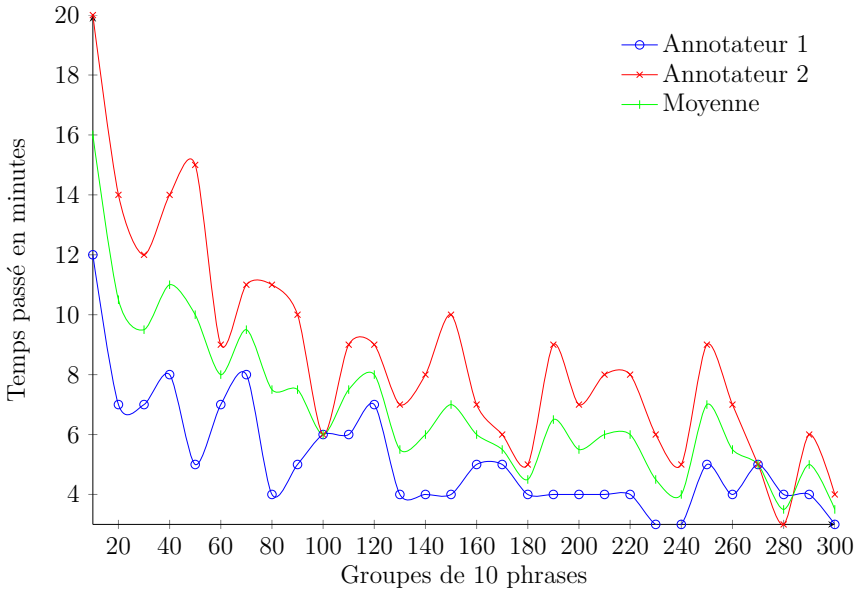


FIGURE 7.1: Courbe d'apprentissage pour l'annotation morpho-syntaxique du *Penn Treebank*

manière séquentielle. Elles ont été menées dans l'ordre décrit ci-dessous. Ainsi, nous avons commencé notre première tâche de correction d'annotations (phrases 1–100) par la phrase 1.

Nous avons mené les expériences suivantes :

1. **Impact de la qualité de la pré-annotation sur l'exactitude et l'accord inter-annotateurs :** nous avons utilisé pour cette expérience les phrases 1–400. Pour chaque phrase, une pré-annotation a été sélectionnée au hasard parmi les pré-annotations possibles (une par instance d'étiqueteur). Le but de cette randomisation est d'éliminer le biais dû à la courbe d'apprentissage des annotateurs (voir figure 7.1). Le temps de correction pour chaque série de 10 phrases consécutives est noté, de même que l'exactitude par rapport à la référence et l'accord inter-annotateurs (nous avons tous les deux corrigé les phrases 1–100 et 301–400, mais un seul d'entre nous a corrigé la série 101–200 quand l'autre se chargeait de la série 201–300).
2. **Impact de la qualité de la pré-annotation sur le temps d'annotation :** cette expérience a été réalisée sur les phrases 601–760, avec pré-annotation. Nous les avons divisées en séries en 10 phrases. Pour chaque série, une pré-annotation a été sélectionnée parmi les huit produites par les étiqueteurs, de manière à ce que chaque pré-annotation soit utilisée pour deux séries. Nous avons mesuré les temps de correction pour chaque série et pour chaque annotateur.

3. **Biais introduit par la pré-annotation :** pour cette expérience, nous avons annoté tous les deux totalement manuellement les phrases 451–500 [2]. Par la suite, nous avons corrigé les phrases 451–475 avec pré-annotation par $\text{MElt}_{\text{en}}^{\text{ALL}}$ (le meilleur étiqueteur) et les phrases 476–500 avec pré-annotation par $\text{MElt}_{\text{en}}^{50}$ (le deuxième moins bon étiqueteur). Nous avons ensuite comparé les annotations réalisées totalement manuellement avec celles correspondants à des corrections de pré-annotations pour vérifier si et comment elles divergent de la référence. Nous avons également comparé les temps d'annotation, afin de confirmer et quantifier précisément le gain de temps observé lors des expériences précédemment rapportées.

7.2 Résultats

7.2.1 Impact de la qualité de la pré-annotation sur l'exactitude et l'accord inter-annotateurs

La qualité des annotations produites lors de l'expérience 1 a été évaluée selon deux méthodes. La première a consisté à prendre les annotations d'origine du *Penn Treebank* comme référence et à calculer une simple exactitude par rapport à celle-ci. La figure 7.2 [3] présente les résultats obtenus par cette méthode.

Cette mesure est en elle-même insuffisante pour évaluer la qualité de l'annotation, la référence n'étant pas parfaite [4]. Nous avons donc cherché à évaluer la fiabilité de l'annotation en calculant l'accord inter-annotateurs entre nous sur les séries de 100 phrases que nous avions corrigées en parallèle. Nous avons pour cela utilisé le coefficient π, ou Kappa de Carletta (Carletta, 1996) (pour plus de détails à ce sujet, voir la sous-section 3.3.2). Les résultats sont présentés dans le tableau 7.2.

Les résultats montrent un accord très élevé, puisque π est toujours supérieur à $0,9$. Il est par ailleurs rassurant de constater que cet accord s'améliore légèrement avec le temps (de 0,955 au début à 0,963 à la fin).

Nous avons également calculé π sur le corpus utilisé pour évaluer le biais de la pré-annotation (expérience 3). Les résultats obtenus sont présentés dans le tableau 7.3.

2. Nous avons remarqué que, pendant cette phrase d'annotation manuelle sans pré-annotation, nous avons utilisé les fonctionnalités `Rechercher-Remplacer partout` de nos éditeurs de texte respectifs pour accélérer l'annotation de certains tokens faciles à annoter comme « the » ou « Corp. », ce qui explique en partie pourquoi les premiers groupes de 10 phrases ont été plus longs à annoter que les autres. En outre, en l'absence d'interface adaptée, un petit nombre d'erreurs typographiques se sont glissées dans l'annotation, telles que *DET* à la place de *DT*.

3. Notons que l'échelle n'est pas régulière en abscisse.

4. Nous avons d'ailleurs pu constater que l'un de nous a fait mieux que la référence sur au moins l'un des sous-corpus (voir sous-section 7.2.3).

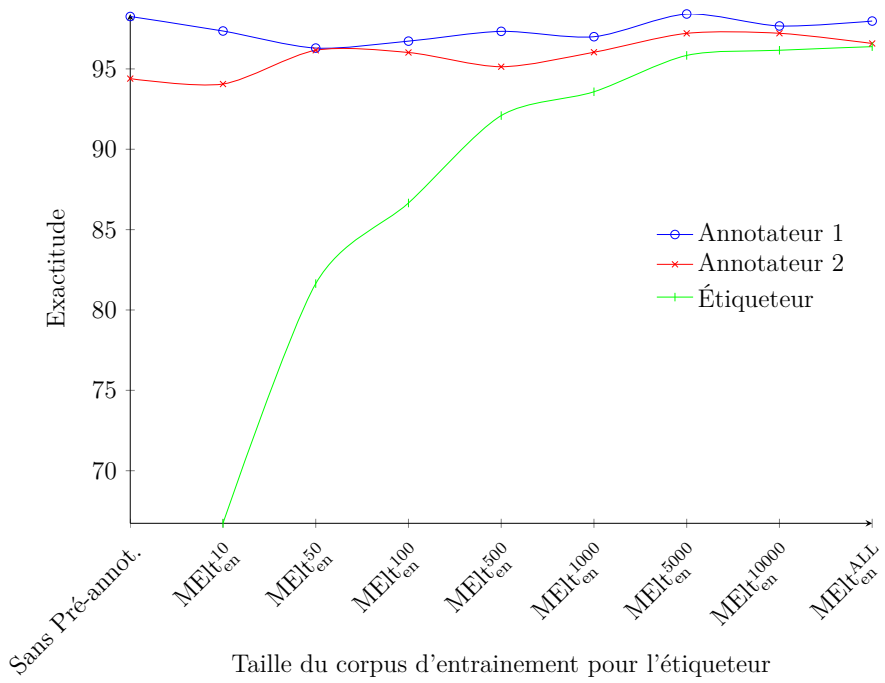


FIGURE 7.2: Exactitude de l'annotation

Sous-corpus	π
1-100	0,955
301-400	0,963

TABLEAU 7.2: Accord inter-annotateurs sur les sous-corpus du *Penn Treebank* annotés en parallèle.

Sous-corpus	Nb phrases	π
Sans pré-annotation	50	0,947
MElt$^{50}_{en}$	25	0,944
MElt$^{ALL}_{en}$	25	0,983

TABLEAU 7.3: Accord inter-annotateurs sur les sous-corpus utilisés pour évaluer le biais de la pré-annotation.

Là encore, les résultats obtenus sont très élevés, bien qu'un peu moins qu'au début de la session d'annotation décrite précédemment. Ils sont presque de 100 % avec la pré-annotation réalisée avec MElt$^{ALL}_{en}$.

Enfin, nous avons calculé π tout au long de l'expérience 2. Les résultats sont présentés

dans la figure 7.3 et, mis à part un pic inexpliqué avec $\text{MElt}_{\text{en}}^{50}$, ils montrent une progression régulière de l'exactitude et de l'accord inter-annotateurs (π), qui sont corrélés. Quant au pic qui apparait avec $\text{MElt}_{\text{en}}^{50}$, il est absent sur la figure 7.2, nous l'interprétons donc comme étant un artéfact.

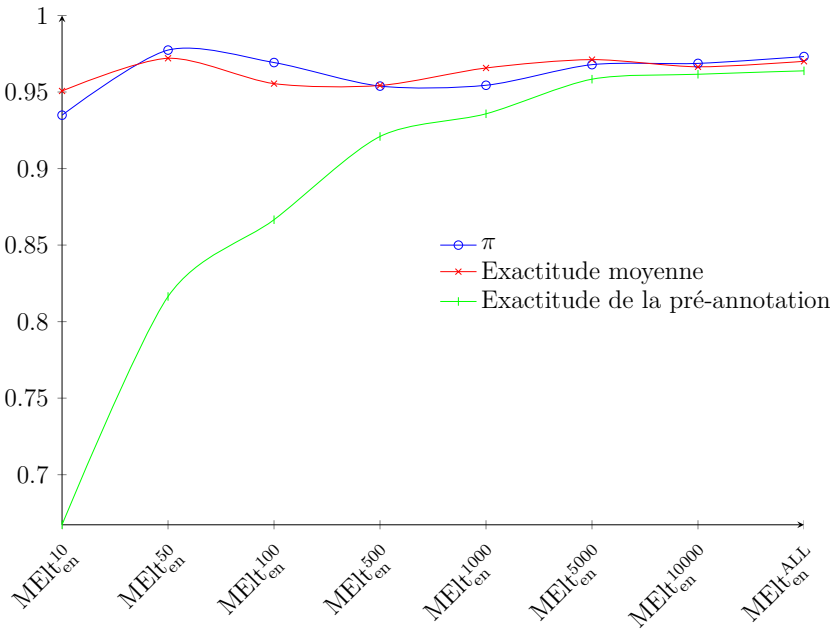


FIGURE 7.3: Exactitude de l'annotation et π en fonction du type de pré-annotation

7.2.2 Impact de la qualité de la pré-annotation sur le temps d'annotation

Notons, avant de présenter les résultats de l'expérience 2, que les mesures du temps d'annotation relevées lors de l'expérience 3 confirment que l'utilisation d'une pré-annotation automatique de bonne qualité (comme celle de $\text{MElt}_{\text{en}}^{\text{ALL}}$) permet de réduire fortement le temps d'annotation par rapport à une annotation totalement manuelle. Ainsi, l'annotateur 1 a eu besoin d'un temps moyen d'approximativement 7,5 minutes pour annoter 10 phrases non pré-annotées (expérience 3) alors que ce temps passe à approximativement 2,5 minutes de correction en cas de pré-annotation avec $\text{MElt}_{\text{en}}^{\text{ALL}}$. Pour l'annotateur 2, les temps correspondant sont respectivement de 11,5 et 2,5 minutes.

La figure 7.4 montre l'impact de la qualité de la pré-annotation sur les temps d'annotation. Nous avons été surpris de constater que seul l'étiqueteur de moins bonne qualité

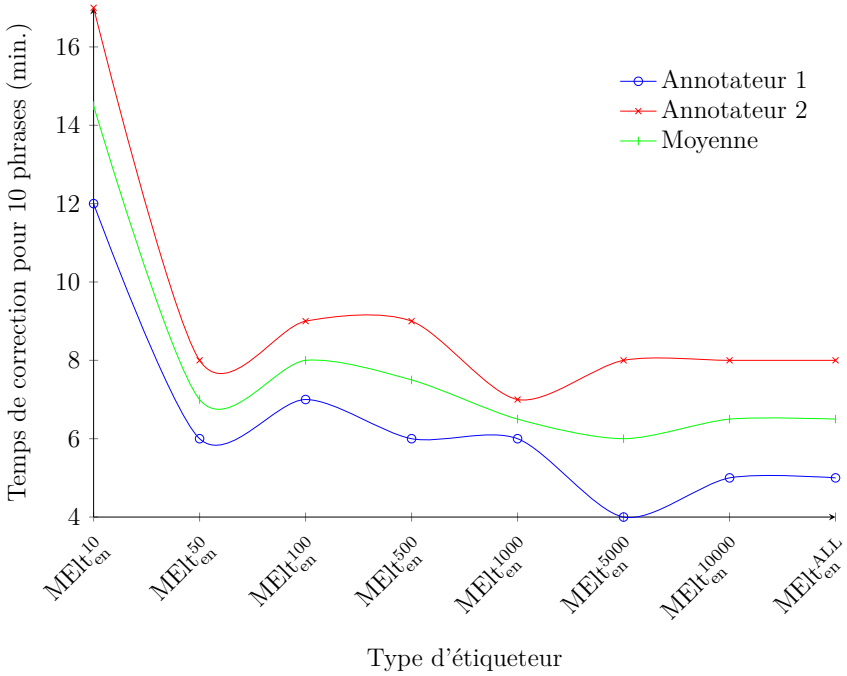


FIGURE 7.4: Temps d'annotation en fonction de la qualité de la pré-annotation

(MElt$^{10}_{en}$) produit des pré-annotations ralentissant l'annotation. En d'autres termes, une pré-annotation automatique présentant une exactitude de 96,4 % n'accélère pas de manière significative le processus d'annotation par rapport à une pré-annotation produisant 81,6 % d'exactitude. Ce résultat est extrêmement intéressant car il pourrait signifier que le développement d'un corpus annoté en morpho-syntaxe pour une langue qui n'en possède pas encore pourrait être largement accéléré par une pré-annotation, même de qualité limitée. Annoter approximativement 50 phrases pourrait en effet être suffisant pour entrainer un étiqueteur morpho-syntaxique tel que MElt et l'utiliser pour pré-annoter le reste du corpus, même si la qualité de l'outil est encore loin d'être satisfaisante.

Une explication de ces résultats pourrait être la suivante : la correction de pré-annotations implique (i) de lire la phrase pré-annotée et (ii) de remplacer une ou plusieurs étiquettes incorrectes. La lecture de la phrase nécessite un temps qui ne dépend pas de la qualité de la pré-annotation. Par contre, la correction d'étiquettes demande un temps que l'on pourrait qualifier de linéaire par rapport au nombre d'erreurs de pré-annotation. Par conséquent, lorsque ce nombre est inférieur à un certain niveau, la correction prend significativement moins de temps que la lecture. De ce fait, sous ce seuil, les variations dans le taux d'erreurs ne provoquent pas de temps de correction significativement plus élevé. Or, ce seuil semble correspondre à une exactitude de la pré-annotation entre 66,5 % et 81,6 %, exactitude qui peut être atteinte grâce à un corpus d'entrainement de taille étonnamment petite.

7.2.3 Biais introduits par la pré-annotation

Nous avons évalué les biais introduits par la pré-annotation avec le meilleur étiqueteur, $\text{MElt}_{\text{en}}^{\text{ALL}}$, et ceux introduits par l'utilisation de l'un des moins bons étiqueteurs, $\text{MElt}_{\text{en}}^{50}$. Les résultats de l'expérience sont présentés respectivement dans les tableaux 7.4 et 7.5.

Ceux-ci montrent un biais très différent en fonction de l'annotateur. Ainsi, l'exactitude de l'annotateur 2 passe de 94,6 % à 95,2 % avec une pré-annotation par un étiqueteur à 81,6 % d'exactitude ($\text{MElt}_{\text{en}}^{50}$) et de 94,1 % à 97,1 % avec un étiqueteur à 96,4 % d'exactitude ($\text{MElt}_{\text{en}}^{\text{ALL}}$). Par conséquent, l'annotateur 2, dont l'exactitude est toujours inférieure à celle de l'annotateur 1 (voir figure 7.2), semble positivement influencé par la pré-annotation, qu'elle soit de bonne ou de mauvaise qualité. Le gain est cependant plus saillant avec la meilleure pré-annotation (+ 3 points).

Quant à l'annotateur 1, qui est le meilleur des annotateurs (voir figure 7.2), ses résultats sont plus surprenants puisqu'ils montrent une dégradation significative de son exactitude, de 98,1 % sans pré-annotation à 95,8 % avec le moins bon étiqueteur ($\text{MElt}_{\text{en}}^{50}$). Un examen plus approfondi de ces résultats montre (i) que la version annotée sans pré-annotation par l'annotateur 1 est de meilleure qualité que la référence et (ii) que les erreurs présentes dans la version pré-annotée avec $\text{MElt}_{\text{en}}^{50}$ sont d'une évidence telle qu'elles ne peuvent être dues qu'à un manque de concentration.

Par contre, les résultats obtenus avec la meilleure pré-annotation restent stables (passant de 98,4 % à 98,2 %), ce qui tend à confirmer ceux présentés dans Dandapat *et al.* (2009), qui montrent que les annotateurs les mieux formés sont moins influencés par la pré-annotation et que la qualité de leurs annotations reste constante.

Annotateur	Sans pré-annotation	Avec $\text{MElt}_{\text{en}}^{\text{ALL}}$
Annotateur 1	98,4	98,2
Annotateur 2	94,1	97,1

TABLEAU 7.4: Exactitude avec ou sans pré-annotation avec $\text{MElt}_{\text{en}}^{\text{ALL}}$ (phrases 451–475)

Annotateur	Sans pré-annotation	Avec $\text{MElt}_{\text{en}}^{50}$
Annotateur 1	98,1	95,8
Annotateur 2	94,6	95,2

TABLEAU 7.5: Exactitude avec ou sans pré-annotation avec $\text{MElt}_{\text{en}}^{50}$ (phrases 476–500)

En tant qu'annotateurs, nous avons remarqué que notre concentration était plus intense lorsque nous annotions totalement manuellement. Il semble donc que les rela-

tivement bon résultats obtenus par les étiqueteurs ont eu pour conséquence un relâchement de notre concentration, d'autant plus que la tâche est répétitive et pénible. Néanmoins, nous avons constaté que l'annotation totalement manuelle pouvait être davantage source d'erreurs.

Ces impressions sont confirmées par la comparaison des matrices de confusion, comme le montrent les tableaux 7.6, 7.7 et 7.8. Dans ces tableaux, les lignes correspondent aux étiquettes provenant de l'annotation et les colonnes aux étiquettes de la référence, et seules les lignes contenant au moins une cellule avec deux erreurs ou plus sont affichées, avec toutes les colonnes correspondantes. On peut ainsi voir que l'annotateur 1 fait davantage d'erreurs aléatoires sans pré-annotation et plus d'erreurs systématiques en présence de pré-annotations avec $\text{MElt}_{\text{en}}^{\text{ALL}}$ (par exemple, *JJ* au lieu de *VBN*, c'est-à-dire adjectif au lieu de participe passé, ce qui correspond à une erreur caractéristique de $\text{MElt}_{\text{en}}^{\text{ALL}}$).

	JJ	VBN
JJ	36	4

(Annotateur 1)

	JJ	NN	NNP	NNPS	VB	VBN
JJ	36					4
NN	1	68			2	
NNP			24	2		

(Annotateur 2)

Tableau 7.6: Extraits des matrices de confusion pour les phrases 451–457 (512 tokens) avec pré-annotation avec $\text{MElt}_{\text{en}}^{\text{ALL}}$

7.2.4 Conclusions

Nos expériences montrent que 50 phrases annotées manuellement sans pré-annotation, ce qui prend environ 40 minutes, permettent de construire un outil de pré-annotation tel que la vitesse de l'annotation manuelle par un expert est quasiment identique à ce que l'on obtient avec un outil de pré-annotation de niveau état-de-l'art, c'est-à-dire que l'on peut construire un corpus complet de taille standard (10 000 phrases) en environ 6 000 minutes (100 heures). Il faut cependant être très attentifs aux biais introduits par cette technique, qui doivent être identifiés et notifiés aux annotateurs dans le guide d'annotation.

Quant au biais introduit par la pré-annotation, les résultats obtenus sont moins clairs. Le moins bon des annotateurs est positivement influencé par la pré-annotation, quelle

	IN	JJ	NN	NNP	NNS	RB	VBD	VBN
JJ		30	2					2
NNS			1	2	40			
RB	2					16		
VBD	1						17	2
WDT	2							

(Annotateur 1)

	JJ	NN	RB	VBN
JJ	28	3		
NN	2	75	1	
RB	2		16	
VBN	2			10

(Annotateur 2)

Tableau 7.7: Extraits des matrices de confusion pour les phrases 476–500 (523 tokens) avec pré-annotation avec $\text{MElt}^{50}_{\text{en}}$

	CD	DT	JJ	NN	NNP	NNS
CD	30			2		
JJ		2	72			
NN			2	148		
NNS					3	68

(Annotateur 1)

	CD	DT	IN	JJ	JJR	NN	NNP	NNS	RB	VBN
IN			104						2	
JJ		2		61		2			1	9
NN	1			4		145				
NNPS							2			
NNS						1	2	68		
RBR					2					

(Annotateur 2)

Tableau 7.8: Extraits des matrices de confusion pour les phrases 450–500 (1 035 tokens) sans pré-annotation

que soit sa qualité, alors que le meilleur semble négativement influencé par une pré-annotation de piètre qualité. Le caractère répétitif de la tâche semble donc peser particulièrement sur la concentration de cet annotateur.

7.3 Confirmation des résultats sur une autre campagne d'annotation

7.3.1 Présentation de la campagne

Nous avons été contactée par Christophe Benzitoun (ATILF, Université Nancy 2) pour du conseil concernant l'organisation d'une campagne d'annotation en morpho-syntaxe d'un corpus de français parlé. Nous avons donc mis en place avec lui la campagne d'annotation, en fonction de ses impératifs (la formation des étudiants L3 et M2 en Sciences du Langage de Nancy 2 étant l'un d'eux), tout en essayant de mettre à profit les résultats obtenus lors de nos expériences précédentes en ce qui concerne la pré-annotation (voir ci-dessus). Nous avons également assuré toute la partie technique de la campagne (calcul d'accords inter-annotateurs, fusion des annotations par les étudiants, etc).

La campagne d'annotation, ainsi que les résultats obtenus, sont décrits en détails dans (Benzitoun *et al.*, 2012). Nous n'en reprenons ici que certains points précis.

Le corpus d'origine qui a été annoté est celui du projet TCOF (Canut *et al.*, 2010), librement disponible sur le site du CNRTL [5]. Le corpus final annoté en morpho-syntaxe, TCOF-POS, est disponible sur le site du CNRTL [6] sous licence Creative Commons BY-NC-SA 2.0 [7], héritée du corpus TCOF. Il contient un peu plus de 100 000 tokens, dont un peu plus de 20 000 tokens annotés dès le début de la campagne par deux experts (C. Benzitoun et L. Bérard) et dit « de référence », et 80 000 tokens obtenus grâce à la double-annotation (par les deux étudiantes recrutées) puis adjudication par un expert linguiste (C. Benzitoun).

L'annotation et les corrections d'annotations ont été réalisées sur un simple tableur, la liste des étiquettes étant contrainte par l'affichage. Les fichiers obtenus sont de type tabulé, avec trois champs : forme, étiquette morpho-syntaxique et lemme.

7.3.2 Méthodologie appliquée

Pour des raisons de difficulté d'utilisation de MElt, qui est encore peu documenté, nous avons dû commencer la campagne en utilisant un étiqueteur morpho-syntaxique bien connu, mais moins performant, `TreeTagger` (Schmid, 1997).

Par ailleurs, l'un des objectifs de la campagne était de montrer l'influence du type de corpus (en l'occurrence, du français parlé « spontané ») sur la qualité de l'étiquetage.

5. http ://cnrtl.fr/corpus/tcof/
6. http ://cnrtl.fr/corpus/perceo/
7. http ://creativecommons.org/licenses/by-nc-sa/2.0/fr/

La méthodologie que nous avons finalement appliquée est la suivante :

1. Création d'un corpus de référence C_{ref} de 22 240 tokens par correction, par deux experts linguistes, d'une pré-annotation automatique :
   – Les 10 000 premiers tokens de C_{ref} ont été pré-annotés avec la version standard de `TreeTagger` ;
   – Les 12 240 tokens suivants de C_{ref} ont été pré-annotés avec une version de `TreeTagger` entraînée sur les 10 000 premiers ;
2. Ré-annotation par deux étudiantes d'environ 7 500 tokens du corpus de référence C_{ref} (suivie d'une phase d'adjudication), afin d'évaluer la qualité des annotations dans deux configurations distinctes :
   – environ 6 000 tokens ont été pré-annotés par `TreeTagger` (version standard) ;
   – environ 1 500 tokens ont été pré-annotés avec une version de `TreeTagger` entraînée sur les 16 312 premiers tokens de C_{ref} ;
   L'objectif était ici de mesurer l'impact de la différence de qualité entre pré-annotateurs en termes de vitesse d'annotation et de précision du résultat de l'étape manuelle ;
3. Application de cette méthodologie à un plus grand nombre d'étudiants pour en valider la robustesse ;
4. Annotation par deux étudiantes d'un corpus additionnel C_{add} de 80 000 nouveaux tokens, pré-annotés avec la version de `TreeTagger` entraînée sur la totalité de C_{ref}.

Enfin, nous avons mené des expériences complémentaires, afin de déterminer à partir de quelle taille de corpus d'entraînement l'étiqueteur obtenu peut être utilisé comme pré-annotateur dans une campagne d'annotation manuelle. Pour cela, Benoît Sagot et Christophe Benzitoun ont entrainé MElt et `TreeTagger` sur 10 sous-corpus successifs du corpus de référence C_{ref}, dont la taille croît de 2 000 à 20 000 tokens.

7.3.3 Résultats

Les résultats obtenus dans le cadre de cette campagne sont, en ce qui concerne la correction de pré-annotations automatiques, cohérents avec ceux que nous avions obtenus lors de nos expériences précédentes (décrites en section 7.1). Nous ne les présentons pas ici en détails, mais ils sont donnés dans (Benzitoun *et al.*, 2012) et nous en faisons ici une synthèse.

D'une part, plus la qualité de l'outil de pré-annotation est élevée, plus les deux étudiantes comme le groupe ont vu leur temps de correction diminuer, en particulier après le ré-entrainement de l'étiqueteur sur le corpus C_{ref}. En 60 heures, les deux étudiantes ont corrigé 80 000 tokens chacune, ce qui fait une moyenne d'un peu plus de 21 minutes par fichier de 500 tokens. Elles sont passées, sur les 15 premiers fichiers, de 107 minutes de correction à 31. Le groupe d'étudiants est passé quant à lui de 110 minutes

pour le premier fichier à moins de 40 pour le sixième. Il faut cependant noter que dans ces expériences la courbe d'apprentissage n'a pas pu être prise en compte pour des raisons pratiques et qu'elle influence donc sans doute largement les résultats.

En ce qui concerne la qualité obtenue, outre une augmentation de l'accord inter-annotateurs, nous avons également constaté une importante augmentation de l'exactitude en moyenne pour chaque étudiant. Pour les deux étudiantes, l'accord inter-annotateurs (Kappa de Cohen (Cohen, 1960)) est passé en moyenne de 0,89 à 0,98 et l'exactitude moyenne de 84,3 % à 94,5 %. Quant au groupe d'étudiants, leur taux d'exactitude moyen est passé de 81,49 % à 94,16 %. Le Kappa moyen est quant à lui passé de 0,74 à 0,93.

Nous avons par ailleurs mis en évidence un seuil de 6 000 tokens comme base de départ optimale pour ré-entraîner l'étiqueteur. A ce stade, on obtient de bons résultats (94,3 % pour MElt et 93,6 % pour `TreeTagger`) et la précision progresse de manière moins marquée.

7.4 Conclusion

Nous avons démontré, grâce à ces différentes expériences, que, dans le cas de l'étiquetage morpho-syntaxique, une pré-annotation, même de qualité limitée et donc développable à faible coût, permet d'améliorer très largement le temps et la qualité des annotations manuelles.

Le même type d'expérience devrait selon nous être mené sur l'apprentissage actif (voir section 3.4.1), avec en ligne de mire une comparaison entre celui-ci et une pré-annotation automatique, voire un mélange des deux. Cette expérience devrait prendre en compte non seulement les gains en termes de qualité et de vitesse d'annotation, mais également la facilité de mise en place et d'application du système.

Par ailleurs, notre expérience, à travers les diverses campagnes d'annotation auxquelles nous avons participé, nous montre que les annotateurs, en particulier les plus experts, peuvent au cours de la campagne devenir capables de produire par eux-mêmes des règles utiles à l'annotation. Ces règles leur permettent de réduire leur temps d'annotation, en réalisant tout ou partie du travail automatiquement. Cela va du `Rechercher-Remplacer` utilisant des expressions régulières simples, à la création de véritables grammaires locales, qui pourraient par la suite être directement utilisables par les outils de TAL. L'outil d'annotation `SYNC3` tient d'ailleurs déjà compte de cette richesse présente dans l'activité d'annotation et propose automatiquement aux annotateurs des expressions régulières déduites de leur travail, qu'ils peuvent modifier avant de les appliquer au reste du corpus.

Ces deux extrêmes du continuum de l'annotation assistée par ordinateur représentent des pistes de recherche prometteuses que nous comptons développer à l'avenir.

Chapitre 8

Évaluer l'annotation manuelle

Comme on peut le voir dans la figure 4.3, la place de l'évaluation est centrale dans toute campagne d'annotation manuelle. Elle est en effet le fil d'Ariane qui permet de garder le lien avec l'interprétation recherchée et d'éviter ainsi l'échec de la campagne. Elle est mise en place par le gestionnaire de la campagne, qui doit s'assurer de sélectionner les mesures appropriées à la campagne et de les calculer de manière cohérente. Nous donnons dans ce chapitre des clés pour y parvenir.

8.1 Motivations

Différentes comparaisons des annotations réalisées (par les annotateurs et les experts) doivent être effectuées pendant la campagne. Ces comparaisons ont deux objectifs, l'amélioration de l'annotation et son évaluation.

8.1.1 Amélioration de l'annotation

Nous l'avons vu dans le chapitre 4, une comparaison des annotations réalisées en parallèle par les annotateurs est réalisée très régulièrement lors de la création de la mini-référence et lors du rodage, afin d'améliorer progressivement le guide d'annotation et de finaliser la formation des annotateurs.

Cette comparaison n'ayant pas pour but premier de produire un coefficient, il n'est pas indispensable de calculer un accord inter-annotateurs *stricto sensu*, surtout dans les cas pour lesquels les coefficients habituels sont moins adaptés (grande prévalence d'une catégorie, annotations de types hétérogènes, etc.). Un accord observé (voir section 3.3) peut alors être suffisant, en complément d'une analyse précise des disparités entre les annotateurs.

Par ailleurs, le fait que les annotateurs soient d'accord ne signifie pas forcément, surtout en début de campagne, qu'ils ont correctement assimilé le guide ou que celui-ci est suffisamment complet ou clair. Ils peuvent avoir fait des interprétations cohérentes entre eux, mais non désirées. Ainsi, lors de la création du corpus arboré du polonais, les auteurs de (Woliński *et al.*, 2011) ont réalisé une vérification au hasard sur 100 phrases, parmi les annotations sur lesquelles les annotateurs étaient en accord et ont trouvé parmi celles-ci 18 arbres erronés. Il est donc important de vérifier régulièrement lors de la campagne que les annotations correspondent aux consignes.

8.1.2 Évaluation de l'annotation

Il est donc important d'évaluer la conformité de l'annotation par rapport à la mini-référence et de calculer un accord inter-annotateurs, au moins en début et en fin de campagne, afin d'avoir une vision claire de la cohérence de l'annotation produite.

Il est en outre utile de calculer un accord intra-annotateur à des moments clés de la campagne. Cet accord intra-annotateur permet de mettre au jour d'éventuelles difficultés chez un annotateur. Ainsi, dans notre campagne 2, un annotateur a vu ses performances nettement diminuer au cours de la campagne. En discutant avec lui, il est apparu qu'il faisait face à des problèmes personnels qui l'empêchaient de se concentrer sur la tâche. Nous avons donc, avec son accord, transféré ses fichiers à annoter à l'autre annotateur.

L'analyse de conformité peut être réalisée régulièrement au cours de la campagne, afin de vérifier que les annotateurs ne régressent pas et restent cohérents par rapport à la mini-référence (donc avec les experts).

Les accords inter-annotateurs calculés pendant une campagne permettent de fournir aux utilisateurs du corpus annoté une mesure de la qualité finale de celui-ci. Les coefficients obtenus peuvent également servir à pondérer les résultats obtenus par les outils entraînés sur ce corpus annoté.

Les coefficients proposés dans la littérature, en particulier ceux de la famille des Kappas, sont aujourd'hui beaucoup utilisés. Cependant, la définition des annotables nécessaire à leur calcul n'est pas toujours très explicite ou très adaptée à la campagne. Or, les annotables peuvent être très différents d'une campagne à l'autre, étant données la diversité et la complexité des annotations réalisées (voir section 2.4).

8.2 Analyse de conformité

Nous reprenons ici le principe d'évaluations régulières sur référence mis en place dans le jeu `Phrase Detectives` (Chamberlain *et al.*, 2008b). Dans ce jeu, les joueurs « formés » se voient régulièrement re-proposer des extraits de la référence (*gold-standard*) utilisés pour la formation à annoter, afin qu'ils puissent corriger d'éventuelles mauvaises habitudes prises au cours du jeu et pour évaluer leur niveau. Nous appelons cette forme d'évaluation sur référence l'analyse de conformité, afin de la distinguer des autres formes d'évaluations que nous utilisons.

Nous proposons, pour évaluer la conformité, d'utiliser des extraits de la mini-référence. Bien entendu, si l'annotation est précédée d'une pré-annotation automatique, celle-ci doit également être appliquée à l'extrait de la mini-référence proposé pour l'évaluation, afin que tout reste transparent pour l'annotateur.

Les annotations réalisées par l'annotateur sont comparées à la mini-référence et les désaccords sont analysés par le gestionnaire. Cette comparaison donne lieu à une mesure d'accord, qui peut être une simple exactitude ou un accord inter-annotateur.

Ces évaluations régulières peuvent mettre au jour des erreurs dans la mini-référence et/ou des incohérences du guide d'annotation, qui devront alors être corrigés. Les corrections s'appliqueront dans ce cas également au corpus déjà annoté (voir section 4.4). Les évaluations peuvent également permettre au gestionnaire de constater des erreurs récurrentes chez un annotateur et de lui demander de rectifier sa manière d'annoter.

8.3 Mesure des accords inter- et intra-annotateur

Les accords inter- et intra-annotateur sont complémentaires de l'analyse de la conformité sur mini-référence (qui peut correspondre à un accord entre l'annotateur et les experts), en ce sens qu'ils permettent de valider l'assimilation de la formation et du guide de manière directe, et de vérifier la cohérence de l'annotation réalisée.

Nous reprenons ici les bonnes pratiques préconisées par (Bonneau-Maynard *et al.*, 2005) et proposons de calculer l'accord inter-annotateurs au plus tôt dans la campagne, puis régulièrement tout au long de celle-ci. Cela implique de faire annoter en parallèle le même extrait du corpus à tous les annotateurs et donc de le prévoir lors de la répartition des fichiers, si tous les annotateurs n'annotent pas tous les fichiers.

Calculer cet accord inter-annotateurs implique de bien connaître la campagne, afin de déterminer le type de mesure adapté (voir section 8.3.2). En cas d'utilisation d'un Kappa, il faut également identifier la segmentation de référence qui définit les « annotables » (voir section 8.3.3). La construction de la mini-référence permet d'obtenir des indices à ce sujet dès le début de la campagne, par le biais du calcul des dimensions de complexité.

Le calcul de l'accord intra-annotateur (de l'annotateur avec lui-même, plus tard dans la campagne) peut être effectué une ou deux fois pendant la campagne.

Nous recommandons d'utiliser pour ce calcul d'accord un extrait différent de la mini-référence, afin de diversifier les sources de test. De même, pour éviter la monotonie, source d'erreur des annotateurs, nous suggérons de ne pas utiliser le même extrait pour le calcul de l'accord inter- et intra-annotateur.

8.3.1 Synthétiser les données

Le gestionnaire de la campagne a besoin d'avoir une vision synthétique et qualitative des données, pas seulement d'une mesure finale. Par ailleurs, une meilleure évaluation des campagnes d'annotation passe par une documentation systématique des choix

réalisés lors de cette synthèse des données. En effet, il nous paraît indispensable de définir ce que l'on considère pertinent pour l'évaluation. De ce point de vue, les représentations de cette synthèse doivent être non seulement fournies, mais leur création détaillée. Nous présentons ici les modes de représentation les plus utilisés, auxquels nous ajoutons un nouveau tableau de synthèse.

Matrice de confusion

La matrice de confusion (ou tableau de contingence) est une représentation détaillée, sous forme de tableau, des résultats quantitatifs d'une évaluation. Nous présentons dans le tableau 8.1 un exemple de matrice de confusion, réalisée pour notre campagne d'annotation de relations de renommage de gènes. Dans ce tableau, A1 est l'annotateur 1, A2 l'annotateur 2, *Former* et *New* sont les catégories utilisées pour marquer le renommage de gènes, et *Rien* est la catégorie correspondant à ce qui n'a pas été annoté dans le texte (pour plus de détails, voir (Fort *et al.*, 2010)).

		A1			
		Former	New	Rien	Total
A2	Former	71	13	23	107
	New	8	69	15	92
	Rien	7	8	18 840	18 855
	Total	86	90	18 878	19 054

Tableau 8.1: Matrice de confusion de la campagne d'annotation de relations de renommage de gènes (campagne 2)

Ce type de représentation permet non seulement de visualiser l'accord entre annotateurs en un coup d'œil (diagonale du tableau), mais également d'avoir rapidement une idée des problèmes que pose la campagne, comme ici la prévalence d'une catégorie, la catégorie *Rien* correspondant aux annotables, qui présente un accord sur 18 878 tokens.

De ce point de vue, nous sommes totalement en accord avec Hripcsak et Heitjan (2002), qui soulignent :

> « showing the two-by-two contingency table with its marginal totals is probably as informative as any measure ».

Cela étant, la matrice de confusion seule ne suffit pas, car sa réalisation implique de faire des choix, qui doivent être explicités. Ainsi, nous aurions tout aussi bien pu, pour notre campagne 2, proposer la matrice 8.2 suivante, dont les annotables ne correspondent plus au nombre de tokens, mais au nombre de noms de gènes identifiés dans le corpus.

		A1			
		Former	New	Rien	Total Noms gènes
A2	Former	71	13	23	107
	New	8	69	15	92
	Rien	7	8	951	966
	Total Noms Gènes	86	90	989	1 165

TABLEAU 8.2: Matrice de confusion calculée à partir des noms de gènes (campagne 2)

Inutile de préciser que l'utilisation de cette matrice de confusion plutôt que la précédente induit des coefficients d'accord inter-annotateurs différents (Fort *et al.*, 2010).

Malheureusement, cette matrice, en deux dimensions, ne peut être utilisée que pour des évaluations menées en parallèle avec deux annotateurs au plus.

Tableau de Krippendorff

Dans les cas d'une évaluation à plus de deux annotateurs, seul le tableau de Krippendorff (Krippendorff, 2004) permettait jusqu'à présent de représenter les données.

Ce type de tableau présente, dans ses lignes, les différentes unités (empans de texte) du corpus et, dans ses colonnes, les annotateurs, les cellules prenant la valeur de la catégorie assignée par l'annotateur à l'unité. Ainsi, pour 3 annotateurs (A1, A2, A3), 5 unités numérotées (par exemple, 5 tokens) et 3 catégories (a, b, c), on pourrait obtenir :

	A1	**A2**	**A3**
token 1	a	a	c
token 2	b	b	a
token 3	b	a	a
token 4	a	a	c
token 5	a	b	c

TABLEAU 8.3: Exemple de tableau de Krippendorff

La plupart des corpus annotés dans les campagnes d'annotation sont d'une taille relativement importante (au moins plusieurs milliers de tokens). Par conséquent, ce tableau est généralement d'une taille telle qu'il ne permet pas d'avoir une vue synthétique de l'évaluation.

Tableau de similarité entre catégories

Ce travail a été mené en collaboration avec Claire François et Maha Ghribi, de l'INIST-CNRS, et Olivier Galibert, du LNE (Fort *et al.*, 2012a). Nous proposons un tableau qui permet de visualiser de manière synthétique les similarités entre les catégories et peut donc être utilisé dans les cas d'évaluations impliquant plus de deux annotateurs. Nous évaluons la similarité entre deux catégories en fonction de la difficulté qu'ont les annotateurs à répartir les éléments annotés entre l'une ou l'autre des catégories.

La création de ce tableau se fait en deux temps. Dans un premier temps, on calcule l'ensemble des probabilités (de type $P(C_2|C_1)$) qu'un annotateur affecte un élément à une catégorie (C_2) sachant qu'un autre l'affecte à une autre catégorie (C_1). Ce calcul est réalisé de la manière suivante :

$$P(C_2|C_1) = \frac{n_{1C_1,2C_2} + n_{2C_1,1C_2}}{n_{C_1}}$$

où $n_{1C_1,2C_2}$ représente le nombre d'éléments assignés par l'annotateur 1 à la catégorie C_1 alors que l'annotateur 2 les assigne à la catégorie C_2 et n_{C_1} représente la somme des éléments assignés à la catégorie C_1 par les deux annotateurs. Lorsque cette probabilité est faible, cela signifie que la catégorie C_2 est peu similaire à C_1 et que le risque d'obtenir une annotation ambiguë entre les deux est faible.

Ces probabilités sont cependant asymétriques, nous ne pouvons donc pas utiliser cette formule telle quelle. Nous considérons que les éléments sont distribués entre les catégories par les annotateurs de manière semblable et définissons chaque similarité entre catégories comme étant la moyenne des similarités orientées :

$$Sim(C_1, C_2) = \frac{P(C_2|C_1) + P(C_1|C_2)}{2}$$

Une fois toutes les combinaisons de similarités calculées, nous pouvons créer le tableau final, qui permet de visualiser facilement les problèmes d'ambiguïté entre catégories. Pour notre campagne d'annotation de relations de renommage de gènes (voir section 5.2), nous avons ainsi obtenu le tableau 8.4.

	Sim
Sim(Former,New)	0,112096
Sim(Former,Rien)	0,078117
Sim(New,Rien)	0,063491

Tableau 8.4: Similarités entre catégories pour la campagne renommage de gènes (campagne 2)

Ce tableau permet de voir que $Sim(Former, New)$ a une valeur plus élevée $(0, 11)$ que $Sim(Former, Rien)$ et $Sim(New, Rien)$, ce qui signifie que *Former* et *New* sont plus ambiguës entre elles qu'avec *Rien*.

Cette nouvelle vue sur les données, centrée sur les catégories, constitue selon nous un nouvel outil très utile, qui permet, à la différence du tableau de Krippendorff (Krippendorff, 2004) d'avoir une vue synthétique des données qui reste lisible, même pour les campagnes impliquant plus de deux annotateurs.

8.3.2 Sélectionner les mesures d'accords à utiliser en fonction de la campagne

Les résultats présentés dans cette sous-section et la sous-section 8.3.4 s'inscrivent dans le cadre d'un groupe de travail sur les accords inter-annotateurs que nous avons initié et qui regroupe Sophie Rosset, Pierre Zweigenbaum, Cyril Grouin (LIMSI-CNRS), Olivier Galibert, Juliette Kahn (LNE), Claire François (INIST-CNRS), Yann Mathet et Antoine Widlöcher (GREYC-CNRS).

Nous sommes à l'origine de la réflexion qui sous-tend cette sous-section. Nous l'avons également formalisée et finalisée, mais les discussions menées dans le cadre du groupe de travail nous y ont largement aidée.

La proposition de sélection des mesures d'accords en fonction de la campagne que nous présentons ici est le reflet de la grille de complexité que nous avons présentée dans le chapitre 6. Nous distinguons l'évaluation d'annotation d'unités (continues ou non) et de relations (orientées ou non).

Évaluation d'annotation d'unités

La méthode d'évaluation de l'annotation d'unités doit, selon nous, se faire en fonction de deux critères, recouvrant en partie les dimensions de discrimination et de délimitation présentées dans le chapitre 6 :

1. la « couverture » de l'annotation : couvre-t-elle tout le texte (étiquetage morpho-syntaxique) ou est-elle dispersée dans celui-ci (renommage de gènes) ?
2. l'homogénéité de la segmentation : est-elle homogène (cas des tokens) ou hétérogène (cas des entités nommées) ?

Ces critères de couverture et d'homogénéité permettent de distinguer quatre cas de figure différents, que nous détaillons ici.

Dans tous les cas, on distingue les jeux d'étiquettes simples des jeux structurés. Il faut également prendre en considération les effets de la prévalence. La prévalence est un phénomène commun, puisqu'elle apparaît dès lors que toutes les catégories ne sont pas

uniformément représentées. Or, une annotation réalisée au hasard peut tirer parti de cet état de fait, en particulier dans les cas de grande prévalence. La prise en compte du hasard permet de réduire ce risque.

Pavage complet avec segmentation simple Dans ce type de campagne d'annotation, tout le flux de données doit être annoté et la segmentation des unités doit être homogène (principalement des tokens).

L'exemple prototypique de ce type de campagne est l'annotation morpho-syntaxique, dans laquelle tous les tokens sont annotés avec une étiquette à choisir parmi un jeu prédéfini. Ce jeu d'étiquettes peut être simple, comme pour le *Penn Treebank* (Marcus *et al.*, 1993), ou structuré, comme dans le cas du bangla (Dandapat *et al.*, 2009).

Dans ce cas précis de campagne, toutes les mesures d'accord (voir section 3.3) peuvent s'appliquer. Le gestionnaire de la campagne a donc tout intérêt à choisir une mesure à la fois simple (car ne nécessitant pas de définir une quelconque distance) et prenant en compte le hasard, comme un Kappa de Cohen (Cohen, 1960) ou de Carletta (Carletta, 1996). En effet, on a dans ce cas autant d'annotables que d'unités annotées et aucune estimation de ceux-ci n'est nécessaire. Il est en outre conseillé de calculer les deux types de Kappa, afin de vérifier qu'il n'existe pas de biais entre annotateurs. Enfin, la prise en compte du hasard permet de réduire l'impact de la prévalence sur les mesures.

Par ailleurs, l'avantage des Kappas, par rapport à la mesure Glozz (Mathet et Widlöcher, 2011a) par exemple, est de fournir un résultat qui puisse être comparé avec ceux d'autres campagnes du même type menées précédemment.

En cas de jeu d'étiquettes structuré, il peut être intéressant d'utiliser une mesure introduisant de la souplesse dans la prise en compte des désaccords, et de considérer par exemple qu'une erreur de sous-type est moins importante qu'une erreur de type. Une mesure comme le Kappa pondéré (Cohen, 1968) ou Alpha (Krippendorff, 1980, 2004) serait alors conseillée.

Pavage partiel avec segmentation simple Dans ce type de campagne, à la différence du précédent, tout le flux de données n'est pas annoté. La segmentation reste cependant homogène.

Un exemple de ce type de campagne est le typage des *« it »* en anglais en anaphoriques ou impersonnels (Weissenbacher et Nazarenko, 2005) [1].

Dans ce cas, il est possible de se ramener au cas précédent en ajoutant une catégorie vide au jeu d'étiquettes et de considérer que tous les tokens non annotés le sont avec cette catégorie.

1. Cette campagne comprenait en fait deux TAE, la première concernant le typage et la seconde l'identification des antécédents des *« it »* anaphoriques.

En cas de très forte prévalence de cette catégorie vide (annotation très dispersée), un Kappa se rapprocherait de la F-mesure (Hripcsak et Rothschild, 2005). Il est alors possible de se limiter à cette mesure.

Pavage complet avec segmentation hétérogène Ce cas regroupe les campagnes dans lesquelles tout le flux de données est annoté, mais pas suivant une segmentation homogène.

Un exemple type est l'annotation syntaxique en constituants. Dans ce cas précis, l'annotation elle-même est structurée et il est important de prendre en compte cette structure dans l'évaluation et de considérer par exemple qu'un *Nom* mal identifié à l'intérieur d'un *Groupe Nominal*, lui bien identifié, est une erreur moins grave que l'inverse. Là encore, une mesure comme le Kappa pondéré ou Alpha serait conseillée. La mesure Glozz peut bien entendu être utilisée, mais elle reste encore trop peu connue pour permettre une comparaison de campagne à campagne.

Dans le cas d'une annotation de ce type, mais non structurée (comme de l'annotation morpho-syntaxique prenant en compte les unités polylexicales), un Kappa peut être suffisant, mais il nécessite d'estimer les annotables, en utilisant par exemple la méthode décrite en sous-section 8.3.3.

Pavage partiel avec segmentation hétérogène Ce type de campagne est de plus en plus courant avec le développement des annotations sémantiques où seules certaines parties du flux de données sont annotées, qui plus est avec une segmentation hétérogène.

Les campagnes d'annotation en microbiologie (campagne 1, voir section 5.1) et en entités nommées (campagnes 5, voir section 5.5 et campagne 3 dans sa partie non relationnelle, voir section 5.3), auxquelles nous avons participé, sont de ce type. Dans les premier (microbiologie) et dernier cas (football), l'annotation était peu ou pas structurée, alors que dans le cas des campagnes 5, les entités nommées étaient structurées.

Dans le cas d'une annotation de ce type, peu ou pas structurée, un Kappa peut être calculé mais il faut alors estimer les annotables (voir section 8.3.3). C'est par exemple ce que nous avons fait dans le cadre de la campagne 3 (football). Le pavage étant partiel, il faut en revanche être attentif aux effets de prévalence, au cas où les annotables seraient trop nombreux par rapport aux annotés (grande dispersion des annotations). Dans ce cas, on peut se limiter à la F-mesure.

Dans le cas d'une annotation structurée, il vaut mieux choisir une mesure permettant de donner des poids différents aux erreurs. Les solutions sont alors les mêmes que citées précédemment : Kappa pondéré, Alpha ou mesure Glozz.

Cas particulier 1 : annotations concurrentes Un premier cas particulier est celui d'annotations concurrentes sur la même unité, quelle que soit sa longueur.

Ces annotations peuvent correspondre à plusieurs couches d'annotation, comme dans le cas de notre campagne d'annotation de matchs de football (campagne 3), où les annotateurs devaient annoter les noms de joueurs et éventuellement ancrer les actions de ceux-ci sur leur nom. Dans ce cas, l'accord inter-annotateurs doit être calculé par couche, sans quoi celui-ci n'aurait guère de sens.

Les annotations concurrentes peuvent également correspondre à des ambiguïtés notées comme telles par les annotateurs. Cette ambiguïté peut soit être réelle (et identifiée comme telle) et, dans ce cas, il peut être intéressant de la conserver (et donc de la prendre en compte comme une nouvelle étiquette, fusion des étiquettes ambigües), soit être le résultat d'un doute d'un annotateur, auquel cas il faut fusionner pour pouvoir considérer cette ambiguïté comme une erreur (ce qui permettra de la voir apparaître et d'éventuellement apporter des précisions dans le guide).

Cas particulier 2 : chevauchements d'annotations Les chevauchements d'annotations (*overlap*) sont des annotations multitokens dans lesquelles un au moins des tokens de l'une des annotations fait également partie d'une autre annotation. Un exemple (théorique) de ce type d'annotation pourrait être, dans un segment de phrase comme « un beau tir de Benzema, le joueur de l'équipe des bleus », d'annoter « beau tir de Benzema » d'une part (comme une action de jeu) et « Benzema, le joueur de l'équipe des bleus » d'autre part (comme une entité nommée « étendue »). Ce type d'annotation n'est pas possible dans du XML en ligne, mais les formats déportés le permettent.

Dans ce cas particulier, comme dans le précédent, l'accord inter-annotateurs doit être calculé de manière différenciée pour les deux couches d'annotation représentées ici.

Les chevauchement d'annotations peuvent également être du même type, comme dans le cas d'une segmentation ambigüe (« pomme de terre cuite »). Dans ce cas, ils représentent une ambiguïté et doivent être traités comme tels (voir ci-dessus).

Évaluation d'annotation de relations

L'évaluation d'annotation de relations reste un domaine encore peu exploré. La tâche est en effet complexe et les relations sont de natures très variées. Le travail que nous menons à ce sujet avec le groupe cité en introduction est d'ailleurs encore en cours.

Une relation se définit en fonction de deux critères, son orientation et son arité. On obtient ainsi la typologie présentée dans le tableau 8.5. Nous distinguons dans ce tableau les chaînes de co-références, dans lesquelles les éléments co-référents sont reliés entre eux comme un ensemble et chaînes « anaphoriques », dans lesquelles le premier

antécédent est identifié comme tel et définit la source de la chaîne (donc son orientation). Cela étant, ces annotations sont souvent considérées comme une somme de relations binaires orientées.

	Binaire	**N-aire**
Orientée	renommage de gènes	chaînes « anaphoriques »
Non orientée	interaction entre gènes	chaînes de co-références

TABLEAU 8.5: Typologie des relations annotées manuellement

Sévérité... Lors de notre campagne d'annotation de matchs de football (voir sous-section 5.6.5) nous avons utilisé, pour l'évaluation de relations (principalement binaires et orientées), un simple Kappa, en prenant en compte de la même manière tous les éléments de la relation (même ancrage, de la même catégorie, même orientation), sans permettre la moindre erreur.

Ce choix, effectué pour des raisons pratiques de simplicité du calcul et d'homogénéité des résultats, nous a permis d'obtenir des accords inter-annotateurs « plancher » (entre 0,4 et 0,57 selon le média source), qui ne sont pas véritablement représentatifs de la qualité de l'accord.

Ainsi, les annotations présentées en figure 8.1 ont été considérées comme en désaccord, alors que seul l'occurrence du nom de joueur est différente.

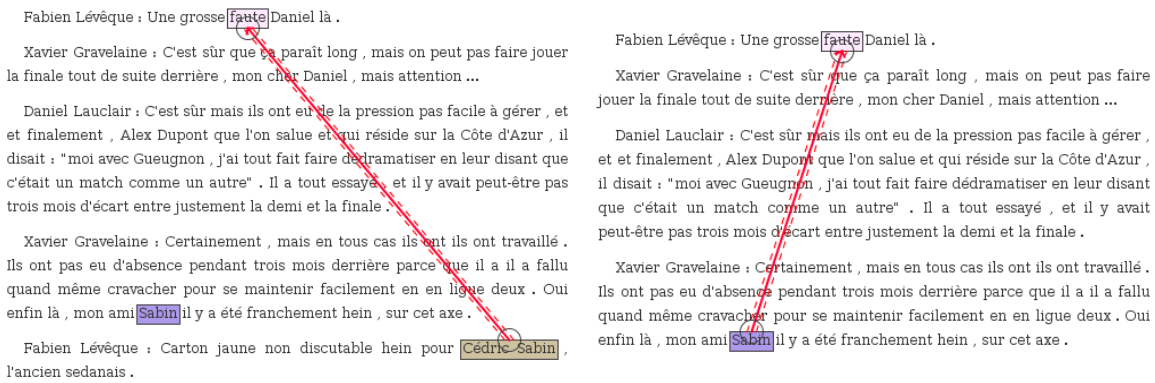


FIGURE 8.1: Exemple de désaccord sur l'ancrage d'une relation *FaireFauteSurJoueur* dans la campagne 3

De même, les erreurs d'orientation de la relation ont été comptabilisées au même titre que les erreurs d'ancrage.

... ou souplesse Une solution pour apporter de la souplesse à la mesure est d'utiliser un coefficient pondéré. Ainsi, Passonneau (2004), puis Poesio et Artstein (2005) proposent d'utiliser Alpha pour l'évaluation d'annotation de chaînes de co-référence (relations n-aires non orientées), car il permet la prise en compte d'un accord partiel (manque d'un élément de la chaîne, par exemple). Se pose alors la question de la définition de la distance nécessaire au calcul de cette mesure. Passonneau (2004) propose de définir cette distance en fonction du nombre d'éléments communs dans l'ensemble que constitue la chaîne de co-référence.

Si cette définition peut être satisfaisante pour ce type de campagne, elle serait sans doute insuffisante pour des campagnes impliquant des relations orientées, dans lesquelles il serait souhaitable de sanctionner plus faiblement une erreur d'orientation par rapport à une erreur d'ancrage, par exemple.

Par ailleurs, si dans la relation de co-référence présentée, l'absence d'un élément dans l'ensemble annoté a un impact réduit, le manque d'un élément dans l'annotation d'une relation binaire correspond à du silence et doit évidemment être sanctionné de manière importante.

La définition de la distance utilisée pour le calcul d'une mesure utilisant une distance, comme Alpha, ou des poids, comme dans le Kappa pondéré, devrait prendre en compte ces critères. Il paraît cependant difficile de les définir objectivement et se pose alors le problème de l'utilisation de l'intuition plutôt que de la vérité terrain.

Les solutions pour l'évaluation de relations restent donc, aujourd'hui encore, partielles. L'utilisation de mesures plus souples nous semble justifiée, malgré leurs limites, mais il faut alors détailler précisément les choix effectués pour définir les distances ou les poids utilisés.

8.3.3 Identifier les annotables

Nous avons vu dans la section 3.3 que les coefficients de la famille des Kappas (*pi* de Scott, Kappa de Cohen) sont très utilisés. Cependant, ces coefficients nécessitent de connaître le nombre d'*annotables* (ou *marquables*). Ce nombre d'annotables est évident ou connu *a priori* pour certaines tâches, comme l'étiquetage morphosyntaxique, où tous les tokens sont annotables, mais ne peut être qu'estimé *a posteriori* pour des tâches où la discrimination n'est pas évidente (voir sous-section 6.2.1).

Impact du choix des annotables sur les mesures

Dans le cadre de la première campagne d'annotation en entités nommées structurées (campagne 5a, voir section 5.5), nous avons mené une réflexion sur la mesure de l'accord inter-annotateurs en collaboration avec Sophie Rosset, Pierre Zweigenbaum,

Cyril Grouin (LIMSI-CNRS), Olivier Galibert et Ludovic Quintard (LNE). Les résultats présentés ici proviennent d'une publication commune (Grouin *et al.*, 2011).

Les entités nommées font partie des nombreux types d'annotation pour lesquels les annotables ne sont pas évidents (Alex *et al.*, 2010). Nous avons donc cherché à voir dans quelle mesure les accords inter-annotateurs obtenus diffèrent selon la méthode choisie pour sélectionner ces annotables.

Dans le cas des entités nommées structurées, on peut par exemple considérer que potentiellement tous les syntagmes nominaux peuvent être annotés (ligne U3 dans le tableau 8.6 [2]), calculé à partir de la campagne PASSAGE (Vilnat *et al.*, 2010)). Bien entendu, il s'agit d'une approximation, les entités nommées n'étant pas nécessairement des syntagmes nominaux (par exemple, « hier » n'est pas un syntagme nominal).

On peut également considérer tous les n-grams de tokens du corpus (ligne U1) ou tous les groupes contigus de tokens de taille inférieure à 6, nous obtenons les résultats présentés en ligne U2. Tout ceci est bien sûr artificiel, car l'annotation des entités nommées ne doit en réalité rien au hasard.

Pour obtenir des résultats plus réalistes, on peut utiliser des données provenant d'autres campagnes d'annotation en entités nommées (la ligne U4 donne les résultats calculés à partir des annotations de la campagne ESTER II (Galliano *et al.*, 2009)), cependant, notre campagne concerne des entités nommées différentes, beaucoup plus étendues et les résultats sont donc moins réalistes qu'on pourrait le croire.

Une autre solution consiste à considérer comme annotables toutes les unités annotées par au moins un annotateur (ligne U5). Dans ce cas, les unités annotées par aucun annotateur sont ignorées (silence).

L'accord inter-annotateurs le plus faible est celui utilisant la dernière méthode de sélection des annotables (ligne U5 du tableau 8.6), le plus élevé correspond à la F-mesure (il est calculé à partir des tous les annotables et est présenté en ligne U1). Les deux premières solutions (U1 et U2) sont trop éloignées de la réalité pour que les résultats obtenus soient pris en compte : les entités nommées structurées restent des annotations dispersées et considérer tous les tokens comme annotables n'est pas adapté. Les trois autres méthodes semblent plus pertinentes, car fondées sur des données réelles. Néanmoins, la solution utilisant les syntagmes (U3) surestime le nombre d'annotables, puisque tous les syntagmes ne sont pas des entités nommées. À l'opposé, l'utilisation des entités ESTER comme annotables sous-estime leur nombre, car cette tâche produit 16,3 fois moins d'annotations que la notre. Enfin, la solution de la ligne U5 (un annotateur minimum a annoté l'unité) donne un résultat plancher pour l'estimation

2. Les accords inter-annotateurs présentés dans ce tableau ont été calculés en utilisant comme référence les annotations des annotateurs ou celles correspondant à la fusion des résultats des deux instituts.

Annotables	Annotateurs	Deux instituts (LIMSI+INIST)
N/A	F = 0,84522	F = 0,91123
U1 : n-grams	κ = 0,84522 π = 0,81687	κ = 0,91123 π = 0,90258
U2 : n-grams $\leqslant$ 6	κ = 0,84519 π = 0,81685	κ = 0,91121 π = 0,90257
U3 : syntagmes	κ = 0,84458 π = 0,81628	κ = 0,91084 π = 0,90219
U4 : entités ESTER	κ = 0,71300 π = 0,71210	κ = 0,82607 π = 0,82598
U5 : un annotateur mini.	κ = 0,71300 π = 0,71210	κ = 0,82607 π = 0,82598

TABLEAU 8.6: Accords inter-annotateurs pour la campagne 5a (κ pour le Kappa de Cohen, π pour le Pi de Scott (Kappa de Carletta), et F pour la F-mesure).

du κ, ce qui représente une information intéressante mais qui ne rend pas tout à fait justice à la qualité de l'annotation produite réellement.

Notre cas de figure confirme la réflexion présentée dans (Hripcsak et Rothschild, 2005) : κ tend vers la F-mesure lorsque le nombre d'annotables tend vers l'infini. Nos résultats montrent surtout qu'il est difficile de bâtir une hypothèse justifiable quant au nombre d'annotables (qui est plus grand que le nombre d'annotations) tout en maintenant κ en-deçà de la F-mesure. Pour autant, ne faire aucune hypothèse revient à sous-estimer le κ.

Proposition d'algorithme d'estimation des annotables

Dans certains cas, les annotables sont déjà identifiés, par le biais d'une pré-annotation automatique, par exemple. Mais il n'existe à notre connaissance aucune solution générique permettant d'estimer le nombre d'annotables dans les autres cas (majoritaires).

Dans le cadre de notre campagne d'annotation de matchs de football (voir section 5.3), nous avons travaillé en collaboration avec Vincent Claveau (IRISA/CNRS) sur le calcul de l'accord inter-annotateurs, et en particulier sur l'estimation des annotables. Nous proposons une estimation originale du nombre d'*annotables* $\mathcal{M}$, basée sur une procédure EM (*Expectation-Maximization*) décrite dans l'algorithme 1, présenté dans (Fort et Claveau, 2012a). Celui-ci énumère itérativement le nombre d'annotables δ (étape de *Maximization*) en utilisant la probabilité γ (estimée itérativement) que tous les annotateurs aient manqué le même annotable, elle-même calculée comme le produit de la probabilité qu'un annotateur A_j manque un annotable (étape d'*expectation*).

Algorithm 1 *Algorithme EM*

Entrées : $\{M_j\}_j$ (ensembles des éléments annotés par les annotateurs A_j)

$\delta_0 = \left| \bigcup_j M_j \right|$

for (i=1 ; changements dans δ ; i++) **do**

expectation : $\gamma_i = \prod_j P(A_j \text{ manque un annotable}) = \prod_j \frac{\delta_{i-1} - |M_j|}{\delta_{i-1}}$

maximization : $\delta_i = \frac{\delta_0}{1 - \gamma_i}$

end for

return δ

Autrement dit, le nombre d'annotables est estimé à partir de l'union des unités annotées par les deux annotateurs, qui représente une sous-estimation du nombre des annotables, et il est peu à peu ajusté pour tenir compte de celles qui seraient oubliées.

Dans l'implémentation réalisée pour notre campagne et qui concernait des unités disparates (entités nommées, actions, relations), nous avons décidé d'être extrêmement stricts et de ne considérer comme annotés par les deux annotateurs que les éléments pour lesquels tout est rigoureusement identique : position, type et éventuellement sens de la relation. Nous avons fait ce choix car nous disposions également de la mesure Glozz (voir sous-section 3.3.3), qui prend en compte les glissements de positions et qui nous donnait donc un résultat « plafond ». Cette définition peut donc être assouplie en fonction d'éventuelles autres mesures disponibles pour la comparaison ou du type de campagne lui-même.

Ainsi, dans le cas de notre campagne d'annotation de matchs de football (voir section 5.3), l'accord inter-annotateurs sur les unités et les actions dans les compte-rendus est de $0,76$ avec la mesure Glozz (résultat « plafond ») et de $0,59$ avec les Kappas, dont le nombre d'annotables est calculé selon la méthode présentée ici. Si l'on considère que tous les tokens sont annotables, on obtient des Kappas de $0,94$. Cette différence considérable s'explique principalement par la prévalence de la catégorie vide (les tokens sans annotation), qui influe largement sur l'accord inter-annotateurs.

Il existe donc de nombreuses manières d'estimer les annotables, qui sont fonction du type de campagne d'annotation, mais également du contexte de la campagne (autres mesures d'évaluation disponibles, par exemple). Nous avons montré que les choix effectués à ce sujet ne sont pas sans conséquence et doivent être détaillés, puis nous avons proposé une solution générique permettant d'estimer le nombre d'annotables pour tous les types d'annotations.

8.3.4 Faciliter l'interprétation des mesures d'accord

Si toute la réflexion sur la signification des accords inter-annotateurs et l'interprétation des résultats a été réalisée en commun dans le cadre du groupe de travail cité précédemment, l'idée d'origine de ce qui suit et l'implémentation de celle-ci reviennent à Yann Mathet et Antoine Widlöcher. Ce travail a donné lieu à une publication (Mathet *et al.*, 2012), dont nous reprenons certaines parties en les complétant, notamment en ce qui concerne les expériences menées sur le corpus TCOF-POS.

Nous avons montré dans la sous-section 3.3.5 que l'interprétation des mesures d'accord inter-annotateurs pose problème. Il est en effet difficile de savoir si un Kappa de $0,7$ représente un bon résultat. Il est en outre pratiquement impossible de comparer des mesures aussi différentes que le Kappa et la mesure Glozz entre elles.

L'idée proposée par Yann Mathet et Antoine Widlöcher est de « retourner » le problème et d'analyser ces mesures sur des corpus portant des annotations dégradées de manière contrôlée.

Effet « Richter »

L'effet « Richter » a été appliqué pour la première fois à une mesure d'accord inter-annotateurs dans (Mathet et Widlöcher, 2011a), mais il est dérivé de recherches menées dans le cadre de la segmentation thématique décrites dans (Pevzner et Hearst, 2002), puis dans (Bestgen, 2009).

Le principe est de générer de manière statistiquement contrôlée des corpus dégradés (par exemple, un segment sur deux est « oublié » par les annotateurs) à partir d'un corpus annoté de référence. Plusieurs corpus sont ainsi générés, correspondant à différentes valeurs d'un paramètre détérioré (par exemple, la longueur moyenne des segments). Cette méthode a permis à Yves Bestgen de démontrer que la distance de Hamming généralisée est plus robuste aux variations en taille de segments que WindowDiff (Bestgen, 2009). Ces premières applications n'utilisent cependant qu'un seul niveau de détérioration, alors que dans l'implémentation réalisée pour (Mathet et Widlöcher, 2011a), toute une série de dégradations est appliquée.

Ce même principe a été repris et généralisé pour notre travail en commun. Yann Mathet et Antoine Widlöcher ont développé dans ce cadre un outil de génération de « secousses » (dégradations) que nous avons appliqué sur différents corpus annotés de référence. Il est à noter cependant que ce travail est à l'heure actuelle limité aux annotations consistant à identifier (discriminer) puis délimiter un empan de texte et à l'étiqueter. Une extension vers les relations et les ensembles est évidemment prévue dans le cadre de notre groupe de travail.

Dispositif

Les erreurs produites par les annotateurs sont de types variés et concernent différents paradigmes. Chaque unité annotée peut en effet diverger de ce qu'elle devrait être (une « référence », forcément imparfaite) d'une ou plusieurs manières différentes, notamment :

- la délimitation de l'unité est incorrecte (les frontières d'une unité ne correspondent pas exactement à la référence) ;
- la catégorisation de l'unité n'est pas correcte (mauvaise catégorie ou mauvaise valeur de trait) ;
- la discrimination de l'unité est incorrecte : l'annotation n'est pas dans la référence (faux positif) ;
- ou, au contraire, une unité de la référence manque (faux négatif).

Toutes ces causes d'erreurs dans l'annotation doivent être prises en compte dans les mesures d'accord inter-annotateurs.

Notre but est donc d'appliquer chaque mesure d'accord à un corpus, dont chaque fichier contient des erreurs d'un ou plusieurs types (ou paradigmes) dans une proportion contrôlée (magnitude). Plus la magnitude est élevée, plus les erreurs sont graves. Les résultats obtenus doivent permettre d'observer le comportement des mesures en fonction des différents paradigmes d'erreurs et selon tout un éventail de magnitudes. Cela permettrait non seulement de comparer précisément les mesures entre elles (pour une magnitude donnée, il sera possible de comparer les scores obtenus par toutes les mesures), mais également d'interpréter les scores de manière tangible (un score donné pour une mesure donnée correspond à une certaine magnitude, dont on connait les effets).

La méthode utilisée consiste donc à simuler le comportement d'un annotateur humain, pour un ou plusieurs paradigmes d'erreurs, grâce à un algorithme qui dégrade progressivement les annotations de l'annotateur idéal (celui qui annote exactement comme la référence), au pire des annotateurs (celui qui annote sans même lire le texte). Cet algorithme prend en paramètre la magnitude, dont la valeur se situe entre 0 (annotateur parfait) et 1 (pire annotateur).

Résultats expérimentaux

Nous ne détaillerons pas ici les protocoles mis en place par Yann Mathet et Antoine Widlöcher, mais ils le sont dans (Mathet *et al.*, 2012). Les expériences menées jusqu'ici ont impliqué principalement des annotations générées automatiquement (ou artificielles) à partir d'un modèle statistique (décrivant les distributions positionnelle et catégorielle des annotables), mais nous avons aussi commencé à tester l'outil sur des corpus annotés réels.

L'outil, dans son état actuel, ne permet pas la prise en compte d'une combinaison de paradigmes, nous avons donc mené des expériences sur la segmentation et la catégorisation de manière isolée. Dans ces expériences, le nombre d'annotateurs a été fixé à trois. Une première expérience concerne le paradigme de segmentation seul et reprend l'expérience de Bestgen (2009) comparant la distance de Hamming généralisée et WindowDiff, en y ajoutant la mesure Glozz. Nous ne détaillerons pas cette expérience ici, car elle concerne des mesures qui ne correspondent pas réellement à des accords inter-annotateurs (WindowDiff et la distance de Hamming nécessitent une référence).

Catégorisation : comparaison de mesures sur des annotations artificielles Les résultats présentés ici ne concernent que la catégorisation. La situation simulée est donc celle d'une campagne d'annotation dans laquelle les unités à annoter sont déjà localisées. Nous avons pour cela créé quatre jeux d'annotations, incluant ou non des cas de prévalence et de proximité entre catégories (une erreur portant sur une sous-catégorie, par exemple, doit être considérée comme moins grave qu'une erreur de catégorie).

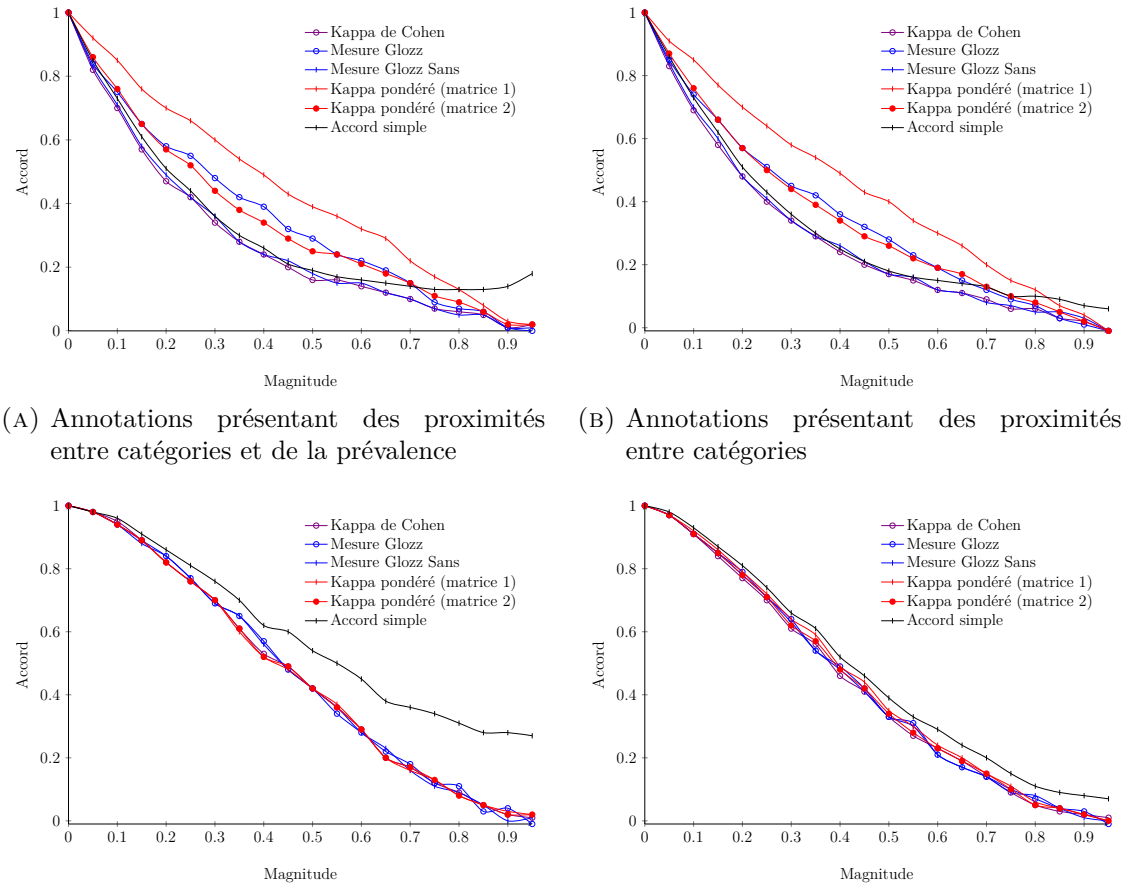
Yann Mathet et Antoine Widlöcher ont ensuite appliqué l'outil Richter sur ces annotations, afin de comparer les mesures suivantes : le Kappa de Cohen (Cohen, 1960), le Kappa pondéré (Cohen, 1968) avec deux matrices de poids différentes (la première étant bien plus indulgente que la seconde), et la mesure Glozz, avec ou sans capacité de gestion des proximités entre catégories. Un accord simple (pourcentage d'accord strict des trois annotateurs) est en outre utilisé comme point de comparaison (*baseline*). Les résultats de ces expériences sont présentés dans la figure 8.2.

Ces résultats montrent avant tout qu'en l'absence de cas de prévalence ou de proximité entre catégories (figure 8.2d), toutes les mesures comparées ont un comportement semblable, y compris l'accord simple (même s'il surestime un peu l'accord à mesure que la magnitude augmente, du fait qu'il ne tient pas compte du hasard).

En cas de prévalence (figure 8.2c), toutes les mesures continuent à se comporter de manière équivalente, à l'exception de l'accord simple, qui surestime de plus en plus l'accord, jusqu'à près de $0,25$. Le hasard semble donc avoir un impact significatif ici.

La prise en compte de la proximité entre catégories oppose clairement le Kappa pondéré et la mesure Glozz aux autres. Cette prise en compte, qu'elle soit couplée à de la prévalence ou non (figure 8.2a et 8.2b), respectivement), provoque d'importantes différences, jusqu'à $0,15$ pour la mesure Glozz, et jusqu'à $0,25$ pour le Kappa pondéré. Cela s'explique par le fait que ces mesures utilisent une matrice décrivant justement les proximités entre catégories (qu'elle soit issue des données, comme dans la mesure Glozz, ou fournit par l'utilisateur, comme dans le Kappa pondéré).

Il est en outre intéressant de constater qu'en appliquant ces deux mesures à des données sans proximité entre catégorie (figures du bas), elles se comportent presque exactement



(A) Annotations présentant des proximités entre catégories et de la prévalence

(B) Annotations présentant des proximités entre catégories

(C) Annotations présentant de la prévalence

(D) Annotations ne présentant ni proximité entre catégories, ni prévalence

FIGURE 8.2: Comparaison des comportements des mesures sur le paradigme de catégorisation

comme leurs versions simples ne prenant pas le phénomène en compte. Ces versions des mesures ne sont donc pas biaisées, quel que soit le corpus. L'accord simple est plus proche des autres mesures dans les cas avec proximité que dans les cas sans, probablement du fait que dans les magnitudes moyennes la proximité prend le pas sur la prévalence, affectant les autres mesures d'une manière semblable. Cependant, au-dessus d'une magnitude de $0,6$, l'accord simple surestime de nouveau l'accord.

Ces résultats sont prometteurs, mais ils doivent être confirmés sur des corpus réels, avec par exemple des variations dans le nombre d'annotateurs et le volume des données.

Application à un corpus réel Nous voulions appliquer notre outil sur de véritables annotations, mais nous étions limités par l'impossibilité de combiner les paradigmes. Nous avons donc choisi de nous concentrer sur les catégories seules et avons cherché

un corpus annoté librement disponible, dont les dimensions de discrimination et délimitation sont nulles. TCOF-POS (Benzitoun *et al.*, 2012) nous a semblé de ce point de vue idéal, les annotateurs de ce corpus annoté en morpho-syntaxe n'ayant pas eu à effectuer de segmentation.

Nous n'avons par ailleurs pas observé de phénomène de prévalence sur cette campagne. Par contre, le jeu d'étiquettes utilisé contient une hiérarchie de types (*PRO :ind*, *PRO :dem*, *PRO :cls*, etc.), il faut donc prendre en compte la proximité des catégories.

Les résultats présentés dans la figure 8.3 confirment ceux obtenus précédemment. L'accord simple, du fait qu'il ne prend pas en compte le hasard, sous-estime l'accord. Le Kappa pondéré semble être la mesure qui sous-estime le moins l'accord dans ce cas.

Il faut cependant noter que cette mesure a été calculée à partir d'une matrice de pondération définie, non à partir des données, mais déduite du guide d'annotation. Les poids de cette matrice sont en effet fonction de l'appartenance ou non à une même sur-catégorie. Ainsi, un poids de $0,5$ est associé à une erreur de sous-catégorie de la même sur-catégorie (par exemple, *Verb-PPRES* et *Verb-FUTUR*) et de 1 dans le cas d'une erreur impliquant deux catégories bien distinctes (par exemple, *Verb-PPRES* et *Noun*).

L'accord inter-annotateurs sur ce corpus, calculé en utilisant le Kappa de Cohen (Cohen, 1960), avait atteint $0,96$. Cela correspond, sur l'échelle Richter obtenue grâce à notre outil et présentée dans la figure 8.3, à une magnitude de $0,1$, donc à une très faible détérioration. Nous pouvons en déduire sans plus aucun doute que le corpus est annoté de manière très cohérente.

Nous n'avons malheureusement pas encore pu utiliser cet outil pour évaluer les résultats obtenus lors des campagnes d'annotation de matchs de football et de brevets en pharmacologie. L'analyse de ces corpus nécessiteraient en effet l'application d'une combinaison de paradigmes (position et catégorisation), qui n'est pas encore disponible.

8.4 Conclusion

Nos travaux ont conduit à la création d'une nouvelle représentation synthétique des accords inter-annotateurs qui permet de visualiser très simplement les ambiguïtés entre catégories, et ce, quel que soit le nombre d'annotateurs. Ils ont également permis de poser de manière claire le problème de l'identification des annotables, auquel nous proposons une solution. Nous proposons en outre ici une aide au choix de la mesure d'évaluation en fonction de la campagne qui devrait être très utile aux campagnes à venir.

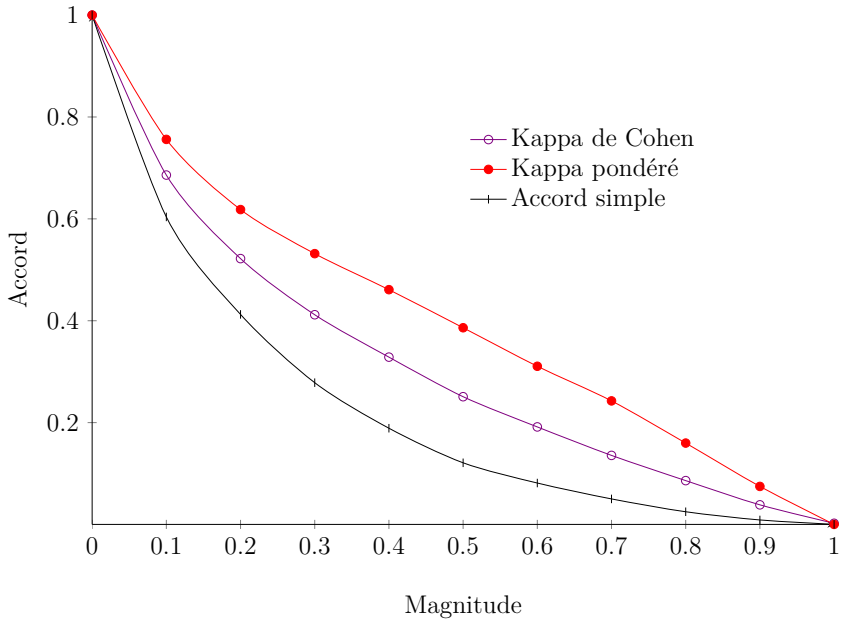


FIGURE 8.3: Comparaison des comportements de différentes mesures sur le corpus TCOF-POS (pas de prévalence, mais prise en compte de la proximité entre catégories)

Enfin, le travail auquel nous avons participé sur l'outil Richter a d'ors et déjà apporté une avancée majeure dans le domaine, en fournissant les moyens de donner une véritable signification aux différentes mesures d'accord utilisées. Ces travaux continuent et devraient permettre d'élargir les résultats déjà obtenus à d'autres mesures d'accord et à d'autres configurations de campagne, ainsi qu'aux relations.

Chapitre 9

Processus et outils des campagnes d'annotation

Nous avons vu dans la section 3.2 que de nombreux outils d'aide à l'annotation existent mais qu'aucun ne fait l'unanimité. Certains de ces outils sont plus complets que d'autres et permettent notamment de gérer une campagne (voir Annexe A.4), mais ils sont souvent difficiles à installer et à paramétrer, donc inadaptés aux petites campagnes d'annotation. Par ailleurs, les outils d'annotation proposent nombre de fonctionnalités intéressantes (notamment de versionnage du jeu d'étiquettes et de l'annotation), mais ils ne prennent pas toujours bien en compte les besoins des annotateurs, en particulier en ce qui concerne le lien avec la documentation, la communication entre annotateurs ou les fonctionnalités de recherche dans le corpus. Enfin, seuls deux d'entre eux (`Slate` et `GATE Teamware`) distinguent différents acteurs, donc différents accès à l'outil.

Surtout, aucun de ces outils ne prend en compte les besoins en termes de préparation de l'annotation, et en particulier de pré-campagne. La gestion proposée est donc amputée d'une phase fondamentale au bon déroulement de la campagne (voir figure 9.1). Enfin, il n'existe à ce jour aucune analyse de la totalité des processus en œuvre lors d'une campagne d'annotation.

Or, le gestionnaire, qui est au cœur de la campagne, a besoin d'en avoir une vision d'ensemble pour la gérer efficacement. Il doit pour cela avoir à sa disposition un outil de pilotage de la campagne, lui permettant d'en définir le cadre (le protocole), d'en gérer les flux et les points de contrôle. Cet outil fondamental est transversal, il intervient pratiquement tout au long de la campagne.

Nous présentons dans ce chapitre différents scénarios de réalisation de pré-campagne d'annotation, puis nous montrons les processus en œuvre lors d'une campagne d'annotation, qui nous amèneront à définir les différents modules nécessaires, en particulier le module de pilotage de la campagne.

9.1 Quelques scénarios de pré-campagne

Nous présentons ici quelques scénarios de pré-campagnes pour analyser les processus de manière générique. Ces scénarios sont inspirés des campagnes d'annotation auxquelles nous avons participé et qui sont décrites dans le chapitre 5, ou de campagnes à venir.

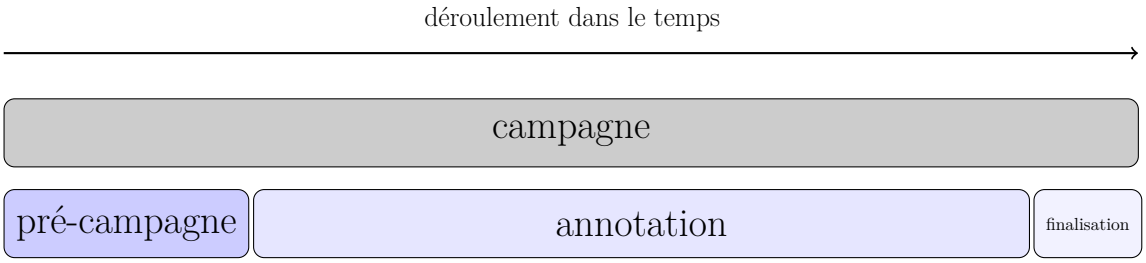


FIGURE 9.1: Organisation simplifiée d'une campagne d'annotation.

9.1.1 Scénario 1 : création de la mini-référence lors du travail préparatoire avec le client

Ce scénario est inspiré de la campagne 4 d'annotation de brevets en pharmacologie (voir section 5.4), en corrigeant ce qui nous semble avoir été responsable des retards dans la campagne. Dans ce scénario, il n'y a pas d'expert annotateur en tant que tel, mais le gestionnaire tient un rôle d'encadrement et d'adjudication (il a alors le rôle d'expert) et le client connaît le domaine.

Le client n'étant pas très clair dans ses besoins, le gestionnaire de la campagne lui a proposé d'annoter avec lui un échantillon du corpus considéré. Cette annotation a été réalisée à deux, sur une machine, en discutant des problèmes au fur et à mesure qu'ils apparaissaient (Demirsahin *et al.*, 2012). Lors de cette phase d'annotation avec le client, un simple logiciel de traitement de texte avec coloration des éléments annotés a été utilisé. Le client s'est rendu compte des difficultés rencontrées et de la nécessité d'avoir un logiciel adapté. À la suite de cette annotation (plusieurs rencontres se sont révélées nécessaires pour finaliser cette annotation), l'échantillon de corpus annoté a été considéré comme une mini-référence.

Le guide d'annotation a ensuite été créé par le gestionnaire, avec validation par le client.

Les dimensions de complexité de la campagne ont été calculées par le gestionnaire à partir de ses connaissances de la campagne et de l'échantillon annoté. Le type de mesure d'accord inter-annotateurs à utiliser a également été choisi à ce moment-là.

La mini-référence a ensuite été utilisée lors de la formation des annotateurs sur l'outil et l'annotation. Leurs annotations ont été comparées directement à celles du client (analyse de conformité) et entre annotateurs (mesure de l'accord inter-annotateurs). Lors de la formation, des points problématiques sont apparus, que le gestionnaire n'a pas pu trancher et qui ont été soumis au client. Le guide d'annotation et la mini-référence ont ensuite été mis à jour par le gestionnaire en fonction des décisions prises.

Ce scénario peut se résumer ainsi :

1. le gestionnaire annote un échantillon du corpus (future mini-référence) et crée le jeu d'étiquettes ;
2. le gestionnaire et le client valident la mini-référence ;
3. le gestionnaire rédige le guide d'annotation ;
4. le gestionnaire évalue la complexité de la campagne ;
5. le gestionnaire forme les annotateurs :
   a) les annotateurs annotent une sous-partie de la mini-référence non annotée ;
   b) le gestionnaire évalue les annotations (par rapport à la mini-référence créée avec le client et entre elles) ;
6. le gestionnaire effectue l'adjudication (suite aux remarques des annotateurs) de la mini-référence ;
7. le gestionnaire modifie le guide d'annotation.

9.1.2 Scénario 2 : création de la mini-référence lors de la formation des annotateurs

Ce scénario reprend en la simplifiant notre campagne 3 d'annotation de matchs de football (voir section 5.3). Dans ce scénario, les phases de formation et de création de la mini-référence sont fusionnées.

Le gestionnaire et le client, également experts du domaine, ont, lors du travail préparatoire, défini l'application (le résumé automatique de matchs de football), un jeu d'étiquettes et écrit un premier guide d'annotation. S'ils ont pris quelques exemples dans le corpus pour étayer leurs choix, ils ne sont pas allés jusqu'à annoter tout un échantillon (par exemple, un compte-rendu de match). Ils ont choisi un outil d'annotation qui leur semblait adapté (`Glozz`).

Une fois les annotateurs choisis, ceux-ci ont annoté un compte-rendu de match court. Cette annotation a permis de soulever de nombreux points problématiques, qui ont été remontés au client par le gestionnaire. Ces deux derniers ont alors décidé d'annoter, chacun de leur côté, le même compte-rendu que les annotateurs. Ils ont comparé ensuite toutes les annotations produites, se sont mis d'accord, puis le gestionnaire a réalisé une version corrigée qui est devenue la mini-référence. Le guide d'annotation a été modifié en conséquence.

Une nouvelle formation des annotateurs, prenant le nouveau guide comme base, a eu lieu. Leurs annotations ont été comparées à celles du gestionnaire et du client (analyse de conformité) et entre elles (mesure de l'accord inter-annotateurs). Lorsqu'il restait des points à trancher, ils l'ont été par le gestionnaire avec l'aide du client, le guide et la mini-référence étant mis à jour en conséquence.

Le gestionnaire a ensuite pu calculer les dimensions de complexité de la campagne et décider d'appliquer des outils de pré-annotation automatique (sur les entités nommées, en l'occurrence).

Ce scénario peut se résumer comme suit :

1. le gestionnaire rédige le guide d'annotation et crée le jeu d'étiquettes ;
2. le gestionnaire forme les annotateurs :
   a) les annotateurs annotent une sous partie du corpus ;
   b) le gestionnaire évalue les annotations produites entre elles ;
3. le gestionnaire et le client annotent la même sous partie du corpus ;
4. le gestionnaire évalue toutes les annotations entre elles ;
5. le gestionnaire effectue l'adjudication en accord avec le client et ce sous-corpus devient la mini-référence ;
6. le gestionnaire modifie le guide d'annotation ;
7. le gestionnaire forme les annotateurs :
   a) les annotateurs annotent une sous partie de la mini-référence non annotée ;
   b) le gestionnaire évalue les annotations (par rapport à la mini-référence créée avec le client et entre elles) ;
8. le gestionnaire effectue l'adjudication de la mini-référence en accord avec le client ;
9. le gestionnaire modifie le guide d'annotation ;
10. le gestionnaire évalue la complexité de la campagne.

9.1.3 Scénario 3 : création de la mini-référence pour un jeu (GWAP)

Ce scénario correspond à la pré-campagne d'une campagne d'annotation en syntaxe de dépendance utilisant un jeu de type GWAP (jeu ayant un but) [1].

Dans ce scénario, le guide d'annotation est pré-existant, car il a été défini lors d'une précédente campagne d'annotation. Le corpus est pré-annoté automatiquement. Plusieurs annotateurs sont disponibles, dont un est également gestionnaire de la campagne. Le jeu constitue le module d'annotation et d'adjudication. Le *backend* du jeu est le module de pilotage.

La pré-campagne se déroule comme suit :

1. des experts annotent (corrigent) des phrases pré-annotées,

1. Ce scénario correspond à une campagne réelle, que nous venons de démarrer au sein de l'équipe Sémagramme du Loria.

2. des annotateurs non-experts corrigent eux aussi ces phrases pré-annotées,
3. les experts créent une mini-référence à partir de ces deux sources d'annotation,
4. le gestionnaire calcule les dimensions de complexité de la campagne,
5. le gestionnaire utilise ces dimensions de complexité pour créer le jeu et les difficultés rencontrées par les annotateurs non-experts sont prises en compte pour adapter celui-ci (notamment en décomposant ou en simplifiant la tâche).

Dans ce scénario, la mini-référence va servir ensuite à la formation des annotateurs-joueurs. La tâche sera probablement décomposée en deux tâches d'annotation élémentaires (TAE) : validation de la pré-annotation automatique et ré-annotation des phrases invalidées. Par ailleurs, des simplifications devront sans doute être imaginées pour les phrases les plus complexes.

9.2 Processus en œuvre lors de la pré-campagne

Nous avons vu dans le chapitre 4 qu'une campagne d'annotation se décompose en plusieurs phases (voir figure 9.1), dont une de pré-campagne, qui permet de créer une mini-référence, de former les annotateurs et de mettre au point le protocole de la campagne.

Cette pré-campagne met en œuvre des processus complexes, que nous décrivons ici.

9.2.1 Création de la mini-référence

La première étape de la pré-campagne consiste à créer une mini-référence, qui sera utilisée tout au long de la campagne. Les différents processus mis en œuvre pour créer la mini-référence sont présentés dans la figure 9.2. Les trois principaux correspondent à la création de la mini-référence, la création du guide d'annotation et l'évaluation.

Dans cette figure, les outils sont représentés par des carrés de couleur bleue, plus ou moins foncée en fonction de leur disponibilité : en bleu foncé, les outils existants (même imparfaits), et en bleu clair, les outils à créer. Les processus qui ne sont pas totalement automatisables sont accompagnés d'un acteur humain. Le module de pilotage de la campagne n'apparaît pas sur la figure parce qu'il est transversal et intervient pratiquement partout. Les paramètres qu'il transmet et les informations qu'il récupère sont représentés par des flèches vertes.

La création de la mini-référence prend en entrée un corpus brut (non annoté), ainsi qu'une version élémentaire du guide d'annotation créée par le gestionnaire de la campagne.

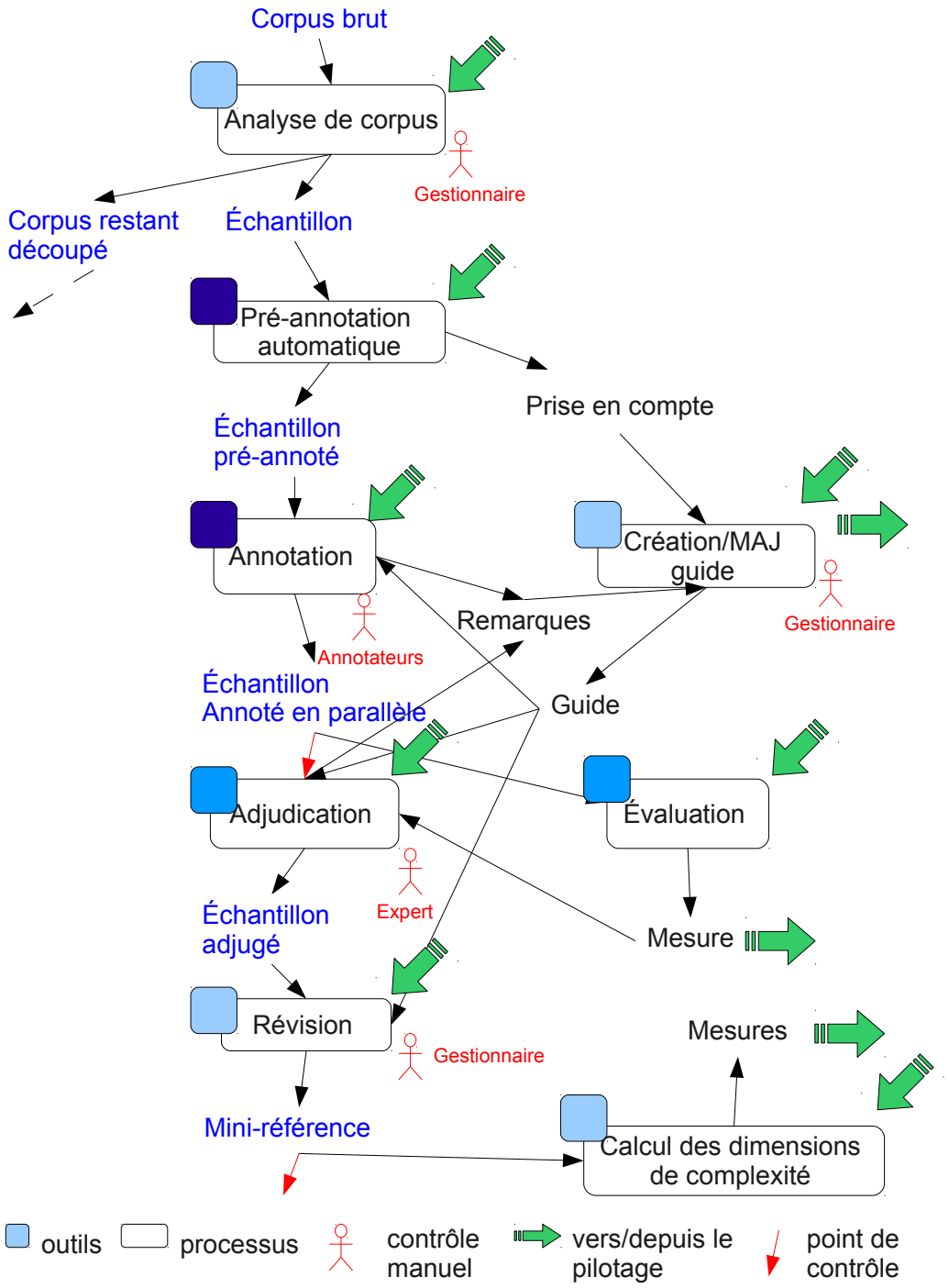


FIGURE 9.2: Processus de création de la mini-référence

Le corpus brut passe par un premier processus d'analyse, d'où il sort découpé de manière représentative et équilibrée en différents blocs. Une partie de ces blocs va servir

d'échantillon et former après annotation la mini-référence. Le reste sera dispatché aux annotateurs et annoté lors de la phase d'annotation (flèche en pointillés sur la figure 9.2).

L'échantillon subit ensuite une éventuelle pré-annotation automatique.

Une fois pré-annoté, cet échantillon est réparti entre les annotateurs-experts pour être annoté en parallèle par au moins deux d'entre eux, qui réalisent cette annotation en suivant les directives fournies dans le guide d'annotation élémentaire. Ils font des remarques qui doivent être prises en compte par le gestionnaire de la campagne. Ce dernier met à jour le guide en fonction des réponses à apporter : simple ajout d'explication ou d'exemple, ou modification plus importante du guide, en particulier du modèle de données.

Une fois que l'échantillon a été annoté par au moins deux annotateurs en parallèle, il subit une évaluation. Le processus d'évaluation produit une ou plusieurs mesures. Dans le cas de la création de la mini-référence, les mesures produites sont des accords inter-annotateurs. Les résultats sont transmis au module de pilotage. Si le gestionnaire de la campagne les juge insuffisants (point de contrôle sur la figure 9.2), il peut décider d'arrêter la campagne ou de tout reprendre, en accord avec le client. Dans le cas contraire, les échantillons annotés vont subir une phase d'adjudication par un expert.

L'échantillon adjugé fait ensuite éventuellement l'objet d'une révision automatique pilotée par le gestionnaire afin de garantir sa conformité avec les dernières mises à jour du guide. L'échantillon ainsi révisé constitue la mini-référence.

Cette mini-référence va être soumise au client pour approbation. Si celui-ci estime que la mini-référence ne correspond pas à ses attentes, tout le processus doit être repris. Dans le cas contraire, la mini-référence va être utilisée par le module de calcul des dimensions de complexité de la campagne, qui va produire les différentes mesures de complexité.

Outre la mini-référence, l'ensemble des processus mis en œuvre lors de la création de la mini-référence produit une mesure d'évaluation intermédiaire, un guide d'annotation mis à jour et des mesures de complexité. Ces mesures de complexité sont des outils d'analyse utilisés par le gestionnaire pour en déduire certains paramètres de la campagne.

Les paramètres nécessaires à la campagne sont en effet produits au fur et à mesure de la création de la mini-référence. Ces paramètres sont les suivants :
– le corpus d'entrée : volume, caractéristiques (mises à jour par le module d'analyse) ;
– le découpage du corpus d'entrée ;
– l'outil de pré-annotation ;
– l'outil d'annotation (lui-même paramétré avec le modèle de données) ;
– l'outil d'adjudication ;
– le nombre d'annotateurs ;

– les annotateurs eux-mêmes (nom, statut) ;
– le type de mesure d'évaluation à appliquer, à quel moment de la campagne ;
– la base utilisée pour l'évaluation (proportion d'annotations à mener en parallèle).

Le protocole à mettre en place pour l'annotation elle-même peut donc être déduit de cette phase de création de la mini-référence.

9.2.2 Formation des annotateurs

La pré-campagne comprend, outre la création de la mini-référence, une phase de formation des annotateurs. Cette phase permet de vérifier la « faisabilité » de la tâche dans le cadre prévu (annotateurs, outils, guide d'annotation).

Les processus mis en œuvre lors de cette phase sont présentés dans la figure 9.3. Cette phase n'utilise que la mini-référence, ce qui représente un volume de données limité, et comprend des itérations relativement courtes.

La formation des annotateurs consiste à leur faire annoter en parallèle une partie de la mini-référence, éventuellement pré-annotée automatiquement (si une telle pré-annotation est prévue pour le reste de la campagne). Lors de ce processus, ils peuvent faire des remarques sur le guide d'annotation qui sont prises en compte par le gestionnaire de la campagne. Le guide est mis à jour en fonction des décisions prises par celui-ci. Le gestionnaire peut parfois être amené à faire modifier la mini-référence afin que celle-ci reste cohérente avec le guide.

Les annotateurs utilisent un outil d'aide à l'annotation. Leur formation comprend donc de fait également l'apprentissage de cet outil. Si celui-ci se révèle difficile à utiliser ou inadapté à la tâche, le gestionnaire peut en sélectionner un autre.

Une fois l'échantillon annoté, il est évalué par rapport à la mini-référence (mesure de conformité) en fonction de paramètres fournis par le module de pilotage. La ou les mesures produites sont envoyées au module de pilotage et le gestionnaire les utilise pour vérifier si l'annotation réalisée est suffisamment correcte pour que la production de l'annotateur puisse être prise en compte. Dans certains cas, plusieurs passes d'annotation peuvent être nécessaires à l'annotateur pour atteindre un niveau suffisant sur la tâche. Il faut parfois remettre en cause sa participation à la campagne (de manière explicite ou implicite). Ce n'est donc qu'à l'issue de cette phase que le choix des annotateurs (nombre et identité) est définitivement fixé.

Lors de cette phase de formation, les annotateurs doivent voir les mesures d'évaluation les concernant.

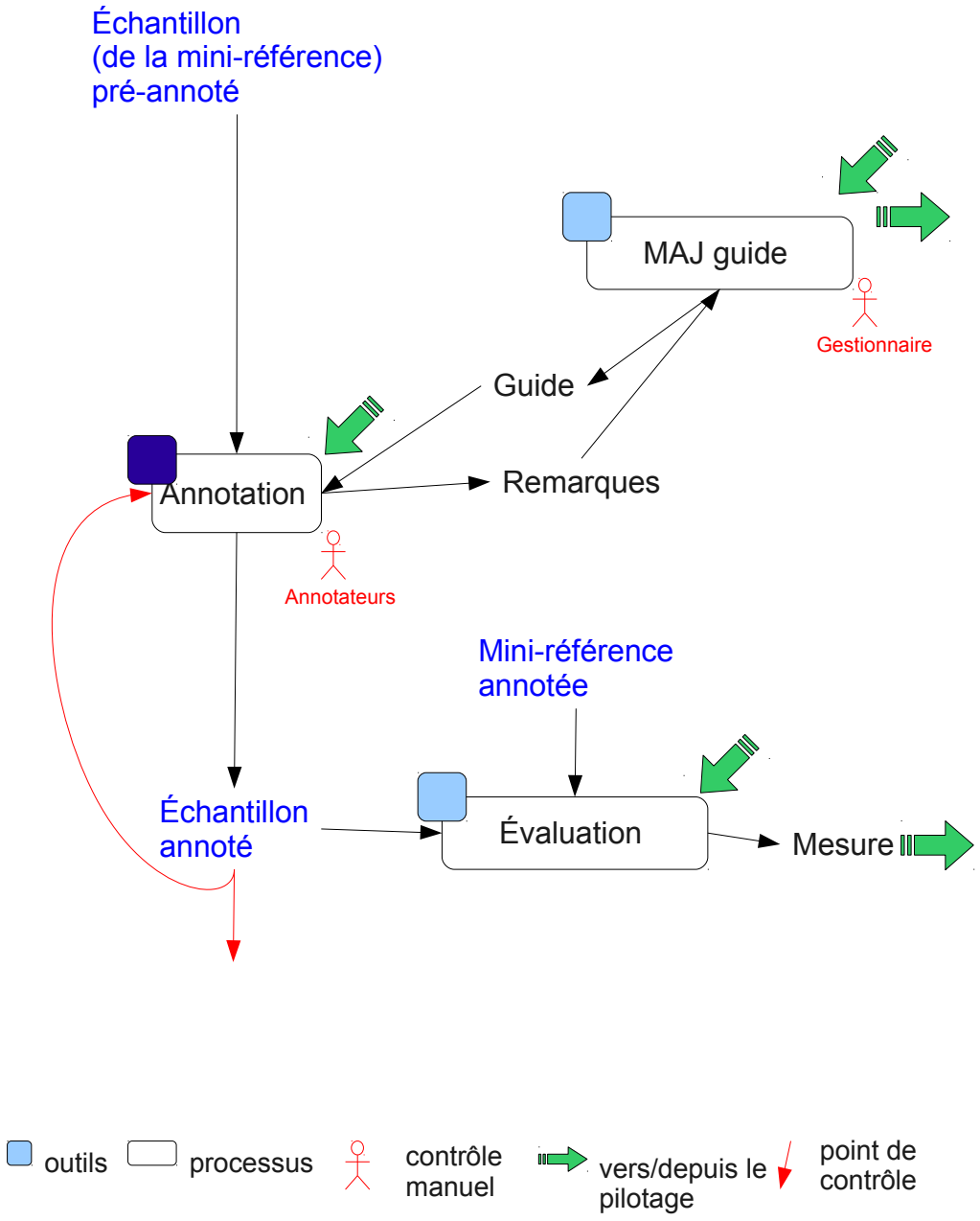


FIGURE 9.3: Processus de formation des annotateurs

9.3 Processus en œuvre lors de l'annotation

Une fois la pré-campagne terminée, la phase d'annotation peut démarrer. Cette phase commence par une étape de rodage, pendant laquelle les annotateurs continuent à monter en compétence et à remettre en cause le guide d'annotation. Le rodage se termine lorsque le guide est stabilisé et les annotateurs efficaces en termes de qualité produite et de vitesse d'annotation (leur courbe d'apprentissage se rapproche de l'asymptote).

La figure 9.4 montre que l'annotation elle-même met en œuvre moins de processus que la création de la mini-référence. Cette phase concerne cependant un volume de données très supérieur. La figure 9.4 reprend les même codes que la figure 9.2, en y ajoutant la couleur magenta pour représenter les processus participant uniquement à la phase de rodage.

L'annotation prend en entrée un corpus découpé en blocs élémentaires représentatifs et équilibrés, ainsi que le guide d'annotation et la mini-référence annotée.

Le corpus découpé en blocs est éventuellement pré-annoté, en fonction du protocole défini pour la campagne. Ces blocs pré-annotés sont ensuite répartis entre les annotateurs en fonction des paramètres fournis par le module de pilotage, puis annotés par les annotateurs.

Pendant la phase de rodage, les annotateurs annotent les blocs de corpus qui leur sont attribués (et qui sont différents de ceux utilisés pour former la mini-référence) et peuvent faire des remarques sur le guide d'annotation. Le gestionnaire va éventuellement mettre à jour le guide et faire corriger la mini-référence. Cette phase correspond généralement à la montée en compétence des annotateurs.

Lorsque les modifications deviennent rares et que les annotateurs sont efficaces, le gestionnaire peut déclarer la fin de la période de rodage et « fixer » le guide, qui ne devra plus subir que des modifications de forme (ajout d'exemples, d'explications, par exemple).

Les blocs de corpus annotés qui peuvent l'être sont évalués, par rapport à d'autres annotations faites sur le même bloc (mini-référence ou annotations d'un autre annotateur). En fonction des résultats, le gestionnaire peut décider de faire ré-annoter tout ou partie de ce qui a été annoté ou de faire réaliser une correction complète lors de la phase d'adjudication, voire de renoncer.

Une fois tout le corpus adjugé, le gestionnaire peut décider d'une révision de celui-ci, afin de le rendre conforme à la dernière version du guide. Une dernière validation est réalisée sur une sélection aléatoire du corpus finalisé. Enfin, ce corpus est publié.

Les processus mis en œuvre lors de l'annotation produisent donc un corpus final, annoté, corrigé et évalué, des mesures d'évaluation et un guide d'annotation à jour.

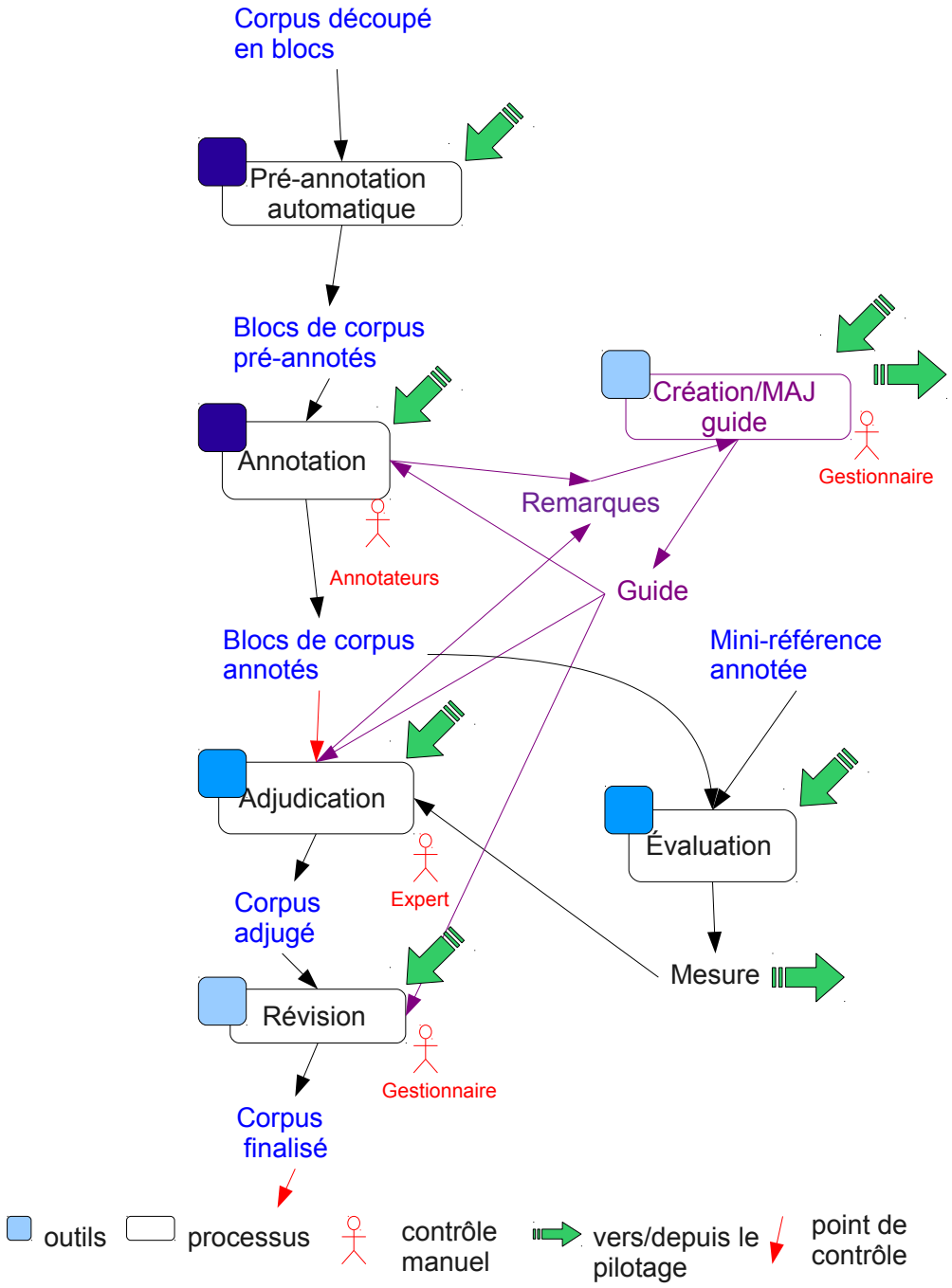


FIGURE 9.4: Processus d'annotation

9.4 Modules nécessaires à une campagne d'annotation

Les processus que nous venons de détailler ont besoin d'être outillés. Nous détaillons ici les modules nécessaires, qui peuvent être utilisés dans différentes phases de la campagne. Il s'agit d'une analyse abstraite et certains de ces modules n'existent pas encore.

Les modules que nous proposons ici permettent la ré-utilisation d'outils existants, en particulier d'outils d'aide à l'annotation et à la gestion de campagnes d'annotation. Il suffit pour cela que les entrées et sorties définies soient compatibles.

9.4.1 Module d'analyse du corpus

Description Ce module prend en entrée le corpus brut et le découpe en blocs équilibrés et représentatifs. En particulier, il crée un échantillon représentatif du corpus général qui va être annoté, évalué et corrigé pour former la mini-référence.

L'obtention de sous-corpus représentatifs et équilibrés est un sujet de recherche à part entière, que nous n'avons pas pu creuser au-delà de notre travail sur l'influence de l'hétérogénéité du corpus (Fort *et al.*, 2011b) sur l'annotation, qui met l'accent sur le fait que les caractéristiques du corpus doivent être prises en compte, notamment :
– la taille des fichiers (en octets) et leur densité (en nombre de tokens),
– la source des fichiers,
– la modalité source (si le corpus comprend différentes modalités source),
– la typologie éventuelle des fichiers.
Cette phase est donc largement manuelle pour le moment, le travail de sélection étant réalisé par le gestionnaire de la campagne. L'automatisation de ce processus devrait permettre, outre un découpage et un comptage (opérations relativement simples), de formaliser les critères de représentativité et d'équilibre utilisés.

Utilisateur Ce module est utilisé par le gestionnaire de la campagne.

Entrées Ce module prend en entrée un corpus non annoté.

Sorties Ce module produit en sortie des blocs équilibrés et représentatifs du corpus global, dont un échantillon qui deviendra la mini-référence. Chaque bloc est en outre accompagné de métadonnées (identifiant, taille, source, etc.) le concernant.

Logiciels existants Il n'existe à notre connaissance aucun logiciel permettant de faire ce type d'analyse et de découpage automatiquement.

9.4.2 Module de pré-annotation automatique

Description Ce module applique une pré-annotation automatique au corpus brut découpé ou à l'échantillon (future mini-référence). Cette pré-annotation peut être dans certains cas homogène à l'annotation et dans d'autres non. Ainsi, dans la campagne 3 d'annotation de matchs de football (voir section 5.3), nous avons appliqué une pré-annotation automatique limitée aux noms de joueurs et d'entraineurs, ce qui a permis de limiter les efforts d'annotation sur ces catégories très répétitives.

Du type de pré-annotation appliqué dépend donc la qualification de la tâche d'annotation qui va suivre. En particulier, celle-ci peut avoir à être découpée en différentes tâches d'annotation élémentaires.

Entrées Ce module prend en entrée :
- le corpus brut ou déjà annoté,
- un outil d'annotation automatique,
- les paramètres nécessaires à l'application de l'outil automatique, entrés par le gestionnaire (*via* le module de pilotage).

Sorties Ce module produit le sous-corpus annoté avec l'outil automatique.

Logiciels existants Certains outils d'aide à l'annotation permettent l'utilisation d'annotateurs automatiques externes. C'est le cas de `GATE`, par exemple.

9.4.3 Module d'annotation

Description Ce module permet de réaliser des annotations de type varié (de segments simples à des relations n-aires orientées ou non). Le module d'annotation intègre tous les paramétrages donnés par le module de pilotage.

Le corpus donné en entrée du module peut avoir été pré-annoté et cette pré-annotation doit être prise en compte et présentée aux annotateurs. Le modèle de données ainsi que le format des annotations doivent être cohérents avec le guide d'annotation, une vérification automatique de numéro de version doit donc être effectuée. Les annotateurs peuvent être amenés à modifier leurs annotations (de leur fait, ou à la demande de l'expert), un moteur de recherche permettant d'effectuer des requêtes à la fois dans le texte et sur les étiquettes est donc indispensable. Par ailleurs, une horloge doit fournir à l'annotateur des informations sur le temps qu'il passe sur chaque fichier, voire sur chaque couche d'annotation (selon le paramétrage du gestionnaire).

Utilisateurs Ce module est utilisé par les experts (pour la mini-référence) et les annotateurs. En fonction du statut de l'utilisateur, les fonctionnalités offertes par le module d'annotation ne seront pas nécessairement les mêmes.

Entrées Ce module prend en entrée :
- les fichiers à annoter, associés aux identifiants et statuts des annotateurs prévus pour ces fichiers ;
- des liens vers la version courante du guide d'annotation et les autres sources à consulter ;
- le modèle de données ;
- les messages éventuels en provenance du gestionnaire.

Sorties Ce module émet en sortie :
- les fichiers annotés, associés aux identifiants et statuts de leurs annotateurs ;
- les messages éventuels pour le gestionnaire.

Logiciels existants De nombreux outils d'annotation génériques existent, qui proposent plus ou moins de fonctionnalités et la prise en compte d'une plus ou moins grande richesse de types d'annotations (voir Annexe A.1). À notre connaissance, aucun de ces outils n'intègre directement un lien vers le guide d'annotation à jour (versionné).

9.4.4 Module d'adjudication

Description Ce module permet de réaliser des comparaisons entre annotations, soit en permettant de visualiser la totalité des annotations, soit en filtrant les accords pour ne montrer que les désaccords. L'utilisateur peut ensuite sélectionner l'annotation qu'il considère la plus juste parmi celles réalisées ou annoter des éléments non encore annotés (silence).

Utilisateurs Ce module est utilisé par les experts.

Entrées Ce module prend en entrée :
- les fichiers annotés en parallèle, avec les identifiants des annotateurs ;
- la mesure d'accord (ou de conformité) sur les fichiers concernés ;
- le modèle de données ;
- le guide d'annotation.

Sorties Ce module produit un fichier annoté unifié, soit uniquement par adjudication des désaccords, soit totalement revu.

Logiciels existants `GATE Teamware` propose ce type de fonctionnalité, mais apparemment limitée à la prise de décision sur les désaccords. Une fonctionnalité de comparaison puis de fusion existe également dans `Slate`, avec les mêmes limitations.

9.4.5 Module de révision

Description Le module de révision permet au gestionnaire d'appliquer au corpus annoté (mini-référence ou autres blocs de corpus) des traitements automatiques ne nécessitant pas d'expertise d'interprétation, mais uniquement la connaissance précise du guide et de ses mises à jour.

Ce module permet notamment de :
– supprimer une catégorie ;
– fusionner deux catégories ;
– changer le nom d'une catégorie ;
– redéfinir une catégorie lorsque cela est possible automatiquement (par exemple, tous les « qui » sont des relatifs dans une configuration donnée).

Utilisateurs Ce module est utilisé par le gestionnaire de la campagne.

Entrées Ce module prend en entrée un corpus adjugé, le guide d'annotation et des paramètres fournis par le module de pilotage.

Sorties Ce module produit un corpus révisé.

9.4.6 Module de calcul des dimensions de complexité (évaluateur de complexité)

Description Ce module permet, pour chaque tâche d'annotation élémentaire, de calculer les dimensions de complexité de la campagne. Il est décrit dans la figure 9.5.

La décomposition en tâches d'annotation élémentaires est réalisée par le gestionnaire de la campagne.

Pour chacune d'entre elles, le calcul des dimensions de complexité s'exécute dès que les informations nécessaires sont fournies par le gestionnaire *via* le module de pilotage.

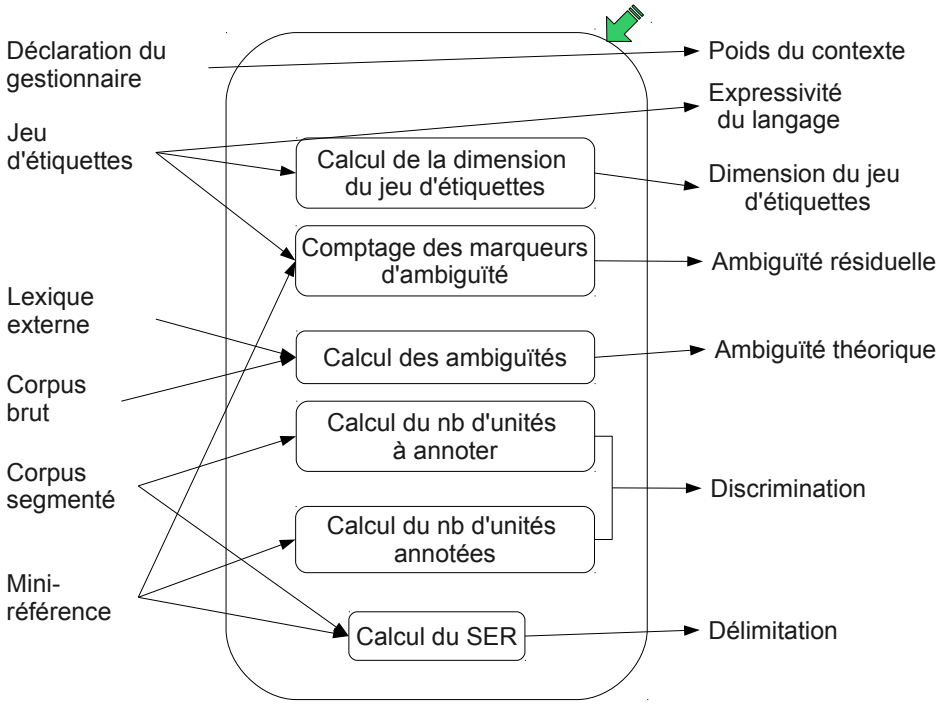


FIGURE 9.5: Module de calcul des dimensions de complexité d'une TAE (entrées et sorties)

Si le poids du contexte est déclaré par le gestionnaire et l'expressivité du langage d'annotation déduite directement du jeu d'étiquettes, les autres dimensions de complexité nécessitent des calculs.

La dimension du jeu d'étiquettes est ainsi calculée à partir du jeu d'étiquettes. Ce jeu d'étiquettes est également utilisé dans le comptage des marqueurs d'ambiguïté (ou d'incertitude), puisque ceux-ci font partie du jeu d'étiquettes. L'ambiguïté théorique nécessite un lexique externe et le corpus brut.

La dimension de discrimination implique deux types de calculs, l'un du nombre d'unités à annoter et l'autre du nombre d'unités effectivement annotées. Elle nécessite donc une mini-référence annotée et le corpus segmenté (même grossièrement).

Enfin, la délimitation est évaluée grâce au calcul du *Slot Error Rate* et nécessite pour cela une mini-référence annotée et le corpus segmenté. Bien entendu, plus la segmentation effectuée est grossière, plus la dimension de délimitation prendra une valeur élevée.

Entrées Ce module prend en entrée :

- le corpus brut ;
- le corpus segmenté (obtenu par un pré-traitement qui peut être différent de la pré-annotation) ;
- la mini-référence ;
- un lexique externe (optionnel) ;
- le jeu d'étiquettes ;
- le poids du contexte tel que défini par le gestionnaire (à partir de l'expérience des annotateurs sur la mini-référence).

Sorties Ce module produit les mesures des différentes dimensions de complexité :

- le score de discrimination ;
- le score de délimitation ;
- le score correspondant à l'expressivité du langage d'annotation ;
- le score correspondant à la dimension du jeu d'étiquettes ;
- le score d'ambiguïté (résiduelle et/ou théorique) ;
- le score correspondant au poids du contexte.

Il les présente sous forme de diagramme radar.

Logiciels existants Aucun logiciel ne propose encore ce type de traitement.

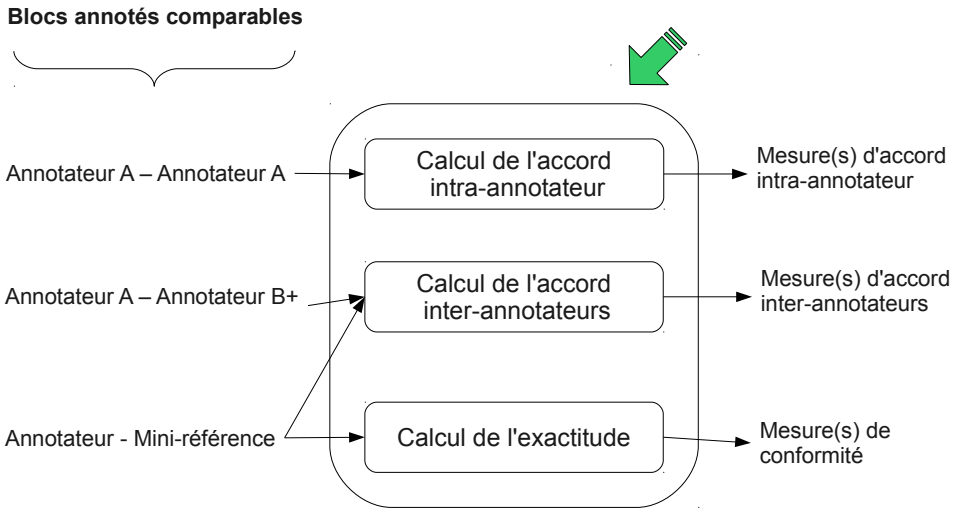


FIGURE 9.6: Module d'évaluation (entrées et sorties)

9.4.7 Module d'évaluation

Description Le module d'évaluation est décrit dans la figure 9.6. Il permet d'effectuer des mesures d'évaluation de différents types :

– accords inter-annotateurs ;
– accords intra-annotateurs ;
– mesure de conformité par rapport à la mini-référence.

L'accord inter-annotateurs peut bien entendu être calculé également entre un annotateur humain et un outil d'annotation automatique.

Le type de mesure à calculer est un paramètre fourni par le module de pilotage en fonction de la campagne et de la phase de la campagne. L'évaluation peut être réalisée par exemple sur certains blocs annotés, sur tout le corpus (dans les cas où il est annoté en totalité plus d'une fois), pour une certaine catégorie, pour un annotateur en particulier.

Les résultats de l'évaluation sont utilisés par le gestionnaire pour le contrôle des flux et par l'expert pour faciliter l'adjudication.

Entrées Ce module prend en entrée des ensembles de fichiers (blocs) annotés comparables :

– mini-référence et annotation de l'annotateur A ;
– annotation de l'annotateur A et annotation de l'annotateur A ;
– annotation de l'annotateur A et annotation de l'annotateur B.

Sorties Ce module produit des mesures d'accords inter- et intra-annotateur, ainsi que de conformité par rapport à la mini-référence.

Logiciels existants Certains outils d'aide à l'annotation, dont `Glozz`, proposent une fonctionnalité de calcul de l'accord inter-annotateurs. L'accord proposé est cependant d'un seul type (par exemple, pour `Glozz`, la mesure Glozz). Il n'existe à notre connaissance aucun outil d'évaluation paramétrable.

9.4.8 Module de création et mise à jour du guide d'annotation

Description La création et la mise à jour du guide d'annotation pourraient être intégrées au module de pilotage et transmises comme des paramètres, mais elles nous semble trop complexes pour être gérées comme tel. Nous avons donc choisi d'en faire un module de autonome.

Ce module permet de créer et modifier le guide d'annotation, qui définit :

– le cadre général de la campagne d'annotation : application finale, logique sous-jacente ;
– le modèle de données, y compris les exemples et les définitions associés à chaque étiquette, les tests spécifiques pour les distinguer, et les contraintes sur les catégories (qui doivent être vérifiées automatiquement par l'outil d'annotation).

Il doit également signaler aux annotateurs les éventuelles pré-annotations présentes dans le corpus à annoter et intégrer des avertissements contre les biais induits par ces pré-annotations.

Utilisateur Ce module est utilisé par le gestionnaire de la campagne.

Entrées Ce module prend en entrée les remarques des annotateurs (y compris des experts) que le gestionnaire choisit de prendre en compte, les pré-annotations éventuelles, ainsi que les paramètres fournis par le module de pilotage.

Sorties Ce module produit le guide d'annotation, c'est-à-dire :
– un texte expliquant le cadre de la campagne ;
– un modèle de données qui devrait idéalement pouvoir être utilisé directement par l'outil d'annotation ;
– les contraintes associées au jeu d'étiquettes.

Logiciels existants Il n'existe à ce jour aucun outil dédié permettant spécifiquement de produire un guide d'annotation.

9.4.9 Module de pilotage

Nous parlons ici de pilotage plutôt que de gestion, car celui-ci ne concerne qu'une sous-partie de la gestion de la campagne. Celle-ci englobe en effet également les interactions avec le client et les interventions manuelles, qui relèvent directement du gestionnaire de la campagne.

Description Ce module permet de piloter une campagne d'annotation, c'est-à-dire de spécifier les données sur lesquelles cette campagne doit s'appliquer, de définir les conditions dans lesquelles elle doit se dérouler et d'en fixer les objectifs. L'ensemble de ces informations définit le protocole d'annotation à mettre en place. Au fur et à mesure qu'elles sont disponibles, ces informations sont transmises aux différents processus qui composent une campagne où elles sont utilisées comme paramètres.

Ces paramètres sont les suivants :
– le type et l'outil de pré-annotation ;
– l'outil d'annotation ;
– le nombre d'annotateurs nécessaire ;
– le nombre d'adjudicateurs (experts) ;
– le corpus brut ;
– le guide d'annotation (uniquement son statut, complet ou non, et sa version) ;
– les mesures d'évaluation et leur base (taux de double annotation sur l'échantillon).
Ce module communique donc avec tous les autres modules. Les données du protocole sont pour certaines renseignées directement par le gestionnaire, alors que d'autres proviennent de l'analyse du corpus et des annotations produites. La communication est donc parfois à double sens.

Nous ne décrirons pas ici en détails le suivi et le contrôle du flux des blocs à annoter, qui se font automatiquement, sous contrôle du gestionnaire, qui peut intervenir aux points de contrôle suivants :
– pendant la création de la mini-référence : avant l'adjudication et une fois la mini-référence créée ;

– pendant la formation des annotateurs ;
– pendant la phase de rodage ;
– lors de la finalisation du corpus annoté.

Utilisateur Ce module est utilisé par le gestionnaire de la campagne.

Entrées Ce module prend en entrée :

– les identifiants des fichiers non annotés qui vont constituer le corpus ;
– les métadonnées concernant les blocs de fichiers, issues de l'analyse du corpus ;
– un lien vers la version courante du guide d'annotation (et d'éventuels liens vers les autres sources à consulter) ;
– les identifiants des fichiers annotés du module d'annotation ou pré-annotés du module de pré-annotation automatique ;
– des mesures d'évaluation (résultats), associées aux identifiants des fichiers évalués ;
– les mesures correspondant aux six dimensions de complexité ;
– des informations concernant les utilisateurs, entrées par le gestionnaire :
  – informations de connexion au système (identifiant, mot de passe, nom),
  – statut (expert ou annotateur), correspondant à des droits spécifiques.
– des informations concernant les utilisateurs, en provenance des modules d'évaluation et d'annotation :
  – informations sur le travail de l'utilisateur (accords inter-annotateurs et intra-annotateur, conformité) pour chaque campagne à laquelle il participe,
  – état d'avancement courant sur la campagne,
  – des messages en provenance des modules d'annotation et d'adjudication.

Sorties Ce module produit notamment :

– des informations concernant la répartition des blocs de fichiers (liées au nombre d'annotateurs, au volume à annoter et au protocole prévu) ;
– les paramètres nécessaires au module de pré-annotation, notamment le jeu d'étiquettes ;
– un lien vers la version courante du guide d'annotation (et d'éventuels liens vers les autres sources à consulter), vers le module d'annotation ou d'adjudication ;
– des fichiers non annotés vers le module d'annotation ou de pré-annotation automatique ;
– des fichiers annotés en parallèle pour le module d'évaluation ;
– la mini-référence pour le module de calcul des dimensions de complexité ;
– une ou des mesures d'évaluation sélectionnées par le gestionnaire ;
– un certain nombre d'informations renseignées par le gestionnaire pour le module de calcul des dimensions de complexité (voir sous-section 9.4.6) ;

– des informations sur le travail de l'utilisateur (accords inter-annotateurs et intra-annotateur, conformité) pour chaque campagne à laquelle il participe, vers le module d'annotation (pour visualisation par l'utilisateur),
– des messages vers les modules d'annotation et d'adjudication.

Logiciels existants Certains outils de gestion d'annotation existent déjà et sont présentés en détails en Annexe A.4. Ces outils sont les suivants : `Slate`, `Djangolody` (qui ne semble pas maintenu) et `GATE Teamware`. Si ces outils permettent de gérer les fichiers, les utilisateurs, les annotations et, pour `GATE Teamware`, les traitements automatiques à appliquer au corpus et un type d'évaluation, aucun ne propose de gérer l'analyse du corpus, la création du guide d'annotation, l'adjudication et le calcul des dimensions de complexité.

Un outil qui intègrerait tous ces modules serait difficile à maintenir, complexe à installer et à paramétrer. L'organisation modulaire que nous proposons permet de dépasser ces limites et de ré-utiliser des outils existants. Elle couvre toute la campagne d'annotation, y compris la pré-campagne, et impose des points de contrôle réguliers, par le biais d'évaluations diverses. Elle offre ainsi au gestionnaire de la campagne les moyens de mieux maîtriser celle-ci, afin de faire produire une qualité optimale d'annotation.

9.5 Retours sur les scénarios

9.5.1 Scénario 1 : création de la mini-référence lors du travail préparatoire avec le client

Le scénario 1, présenté en sous-section 9.1.1 correspond à une instanciation simple des processus décrits en section 9.2, sans pré-annotation. La seule différence est liée à la création de la mini-référence par le client, plutôt que par des annotateurs-experts.

Le fait de faire appel au client comme annotateur-expert permet de ne pas appliquer les points de contrôle prévus avant l'adjudication et à la fin de la phase, puisque le client est directement impliqué. Par ailleurs, l'annotation étant réalisée à deux sur une seule machine (*peer annotation*), l'évaluation par l'accord inter-annotateurs est impossible et l'adjudication devient inutile.

La création de la mini-référence est donc largement simplifiée. Une révision reste possible. Le calcul des dimensions de complexité doit être effectué. Le guide d'annotation est produit à l'issue de cette phase.

Le seul inconvénient de cette configuration est que la mini-référence est nécessairement de taille réduite (il est difficile de faire annoter un client pendant plusieurs jours). Par conséquent, il est probable que certains problèmes n'aient pas été vus. La phase de

rodage risque donc d'être plus longue. Il se peut également que certaines dimensions de complexité ne soient pas représentatives du corpus dans son ensemble (ça peut être le cas pour l'ambiguïté, par exemple).

9.5.2 Scénario 2 : création de la mini-référence lors de la formation des annotateurs

Dans ce scénario (voir sous-section 9.1.2), la formation des annotateurs et la création de la mini-référence sont réalisées en même temps. Les processus mis en œuvre ici sont ceux décrits en section 9.2.

Les mises à jour du guide sont dans ce cas plus nombreuses et le gestionnaire doit donc être très disponible. Les interactions avec le client sont elles aussi plus nombreuses.

Dans un cas comme celui-ci, la formation des annotateurs prend sans doute davantage de temps, puisque le guide est encore très instable. En revanche, les annotateurs sont plus impliqués dans les processus de décision.

9.5.3 Scénario 3 : création de la mini-référence pour un jeu (GWAP)

Dans ce scénario, présenté en sous-section 9.1.3, les processus mis en œuvre sont identiques à ceux que nous avons présentés en section 9.2, si ce n'est que le corpus est déjà pré-annoté et le guide est pré-existant (il est cependant probable qu'il sera modifié lors de la pré-campagne).

Une autre différence est que des annotateurs non-experts participent à la création de la mini-référence, afin d'évaluer la difficulté de la tâche.

Ce scénario montre que la pré-campagne est utile quel que soit le type de campagne prévu, traditionnel ou par *crowdsourcing*.

Conclusion

L'annotation manuelle de corpus est devenue fondamentale pour le TAL. Les corpus annotés sont en effet utilisés à la fois pour créer des outils d'annotation automatique et pour l'évaluation de ces outils. Pour être utile, voire ré-utilisable, cette annotation doit être de bonne qualité. Or, le processus d'annotation manuelle est encore mal connu et les outils proposés parfois mal utilisés.

Ce travail de thèse tente d'apporter une vision unifiée de l'annotation manuelle de corpus pour le TAL, sans pour autant la simplifier. Il est le fruit de diverses expériences de gestion et de participation à des campagnes d'annotation, mais également de collaborations avec différents chercheur(e)s.

Nous avons participé à six campagnes d'annotation dans le cadre du programme Quæro (dont trois que nous avons gérées) et à une campagne en dehors de ce programme. Nos différentes collaborations et l'état de l'art que nous avons réalisé nous ont permis d'élargir ce point de vue. À partir de ces expériences, nous avons effectué un travail de synthèse et d'organisation, puis nous avons tenté d'apporter des solutions à certains problèmes précis rencontrés par le gestionnaire de campagne d'annotation. Nous avons en chemin redéfini toute une terminologie autour de l'annotation manuelle de corpus.

Une première partie de notre travail a consisté à proposer une organisation des campagnes d'annotation, mettant l'évaluation au cœur de celle-ci par la création d'une mini-référence, l'intégration de la formation des annotateurs dans la pré-campagne et d'un contrôle qualité tout au long de la campagne. Ce contrôle est assuré par le gestionnaire de la campagne, qui est en contact avec le client et connaît ses besoins ainsi que l'application visée. Ce même gestionnaire définit le protocole appliqué lors de la campagne.

Nous avons également mis au point une grille d'analyse des dimensions de complexité d'une campagne d'annotation, qui donne une vision synthétique de la complexité d'une campagne et permet de prévoir les difficultés, voire de les éviter en fournissant les outils adaptés. Les dimensions de complexité que nous avons identifiées permettent de répondre aux deux questions fondamentales qui se posent lors d'une campagne d'annotation : quoi annoter ? et comment annoter ? Ces dimensions sont la discrimination des unités à annoter, la délimitation de celles-ci, l'expressivité du langage d'annotation utilisé, la dimension du jeu d'étiquettes, le degré d'ambiguïté de la tâche et le contexte à prendre en compte pour annoter.

Ces deux premiers axes de travail correspondent à une méthodologie globale pour la gestion de l'annotation manuelle de corpus. Un autre volet de ce travail de thèse a concerné les outils du gestionnaire de campagne.

Nous avons mené une série d'expériences sur la pré-annotation automatique et son influence sur la qualité et la rapidité de correction par les humains. Les résultats que nous obtenons sont surprenants, puisqu'ils montrent que pour l'annotation morpho-syntaxique de l'anglais, un outil de pré-annotation automatique entraîné sur 50 phrases annotées manuellement permet des gains très importants en temps et en qualité de l'annotation.

La qualité de l'annotation manuelle représente justement un autre axe de notre travail, sur lequel nous avons apporté des solutions pratiques. Nous avons ainsi proposé un tableau de synthèse de données permettant d'avoir une vue qualitative de l'annotation quel que soit le nombre d'annotateurs. Nous avons également élaboré un algorithme d'estimation du nombre d'annotables (indispensable au calcul des mesures de type Kappa) qui permet d'obtenir une approximation réaliste de celui-ci pour toutes les campagnes d'annotation. Enfin, nous avons participé à un travail de réflexion sur les accords inter-annotateurs qui a donné naissance à l'outil « Richter ». Cet outil applique des dégradations contrôlées sur des corpus annotés de référence, afin d'évaluer précisément le comportement des mesures d'évaluation en présence de tel ou tel type d'erreur d'annotation et de mieux caractériser les résultats obtenus.

Les campagnes auxquelles nous avons participé ont par ailleurs été décrites précisément, notamment en ce qui concerne leur évaluation, dans diverses publications. Nous espérons que cet effort de transparence permettra à d'autres chercheurs de comparer leurs campagnes aux nôtres en toute connaissance de cause et de profiter de nos expériences.

Enfin, nous avons mis au jour les processus en œuvre et les outils nécessaires dans une campagne d'annotation et instancié ainsi la méthodologie que nous avons décrite. La solution modulaire que nous proposons est suffisamment souple pour intégrer des outils existants et permettre de ne pas tout utiliser dans tous les cas, ce qui limite les efforts à fournir pour les petites campagnes d'annotation, par exemple.

Nous espérons surtout que le travail réalisé dans le cadre de cette thèse donne une vision plus cohérente de la tâche d'annotation manuelle de corpus pour le TAL et qu'il permettra une meilleure prise en compte de cette activité dans le domaine.

Nous avons posé le cadre de l'activité d'annotation manuelle de corpus, mais de très nombreuses pistes de recherche restent encore à creuser.

Ainsi, nous continuons à avancer sur le sujet des accords inter-annotateurs. Nous comptons compléter rapidement les expérimentations avec l'outil « Richter ». L'élar-

gissement de la réflexion sur l'évaluation de l'annotation de relations est également en cours.

Un autre champ de recherche très prometteur est l'annotation assistée par ordinateur. En effet, il existe un véritable continuum d'automatisation, depuis les fonctions `Rechercher-Remplacer` à l'annotation totalement automatique, et nous souhaitons l'explorer plus avant. Une typologie fine des méthodes d'annotation assistée permettrait sans doute de mieux en comprendre les possibilités. Des expériences semblables à celles que nous avons menées sur la pré-annotation pourraient également être menées sur l'apprentissage actif.

Nous souhaitons également mettre en œuvre au moins une partie des modules que nous décrivons dans le chapitre 9. Nous envisageons en particulier le développement d'un module de calcul des dimensions de complexité d'une campagne d'annotation. Nous pourrions l'intégrer dans `Slate`, par exemple, dont le développeur, Dain Kaplan, est intéressé par le projet.

Nous avons en outre commencé à travailler, au sein de l'équipe Sémagramme du Loria avec Bruno Guillaume et Guy Perrier, sur l'annotation manuelle d'un corpus arboré du français librement disponible. Nous prolongeons en cela les efforts effectués par Candito et Seddah (2012) sur le corpus Sequoia. Deux axes de recherche sont envisagés. Le premier concerne la correction d'annotations automatiques à l'aide d'un jeu (GWAP) et le second vise à étudier comment mieux intégrer le guide d'annotation dans l'outil d'aide à l'annotation. Notre projet de jeu doit permettre la correction d'annotations syntaxiques en dépendance, réalisées automatiquement, par des joueurs formés à la tâche. Nous espérons ainsi obtenir non seulement un corpus annoté en syntaxe pour le français qui soit totalement libre, mais également de nombreuses informations sur le développement d'un tel jeu, notamment en termes de coût et de difficulté gérable par les joueurs. Notre second projet consiste à proposer des méthodes et des outils pour aider le gestionnaire à développer un guide et un corpus en formalisant en partie la cohérence entre guide d'annotation et corpus annoté.

Plus généralement, nous pensons qu'une solution partielle à la difficulté de financement du développement de ressources langagières pourrait résider dans ce type de *crowdsourcing*. Une plate-forme générique de jeux, proposant des modules pour créer ses propres jeux et drainant ses propres joueurs serait sans doute utile au domaine. La plate-forme `GALOAP`[2], développée à l'Université de Pise dans le laboratoire de G. Attardi, est un premier pas en ce sens.

2. `GALOAP` est disponible sous licence BSD à l'adresse suivante : `http://galoap.codeplex.com/`

Bibliographie

Anne Abeillé, Lionel Clément et François Toussenel : Building a treebank for French. *In* Anne Abeillé, éditeur : *Treebanks*, pages 165 –187. Kluwer, Dordrecht, 2003.

Gilles Adda, Benoît Sagot, Karën Fort et Joseph Mariani : Crowdsourcing for language resource development : Critical analysis of Amazon Mechanical Turk overpowering use. *In Proceedings of the Language and Technology Conference (LTC)*, Poznań, Pologne, novembre 2011. URL http://hal.archives-ouvertes.fr/hal-00648187. 5 pages.

Madeleine Akrich et Dominique Boullier : *Savoir faire et pouvoir transmettre*, chapitre Le mode d'emploi, genèse, forme et usage, pages 113–131. éd. de la MSH (collection Ethnologie de la France, Cahier 6), 1991. URL http://www.uhb.fr/sc_humaines/las/IMG/pdf/emploi.pdf.

Beatrice Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, Michael Matthews, Stuart Roebuck, Richard Tobin et Xinglong Wang : Assisted curation : Does text mining really help ? *In Proceedings of the Pacific Symposium on Biocomputing*, 2008. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.137.4682.

Beatrice Alex, Claire Grover, Rongzhou Shen et Mijail Kabadjov : Agile corpus annotation in practice : An overview of manual and automatic annotation of CVs. *In Proceedings of the Fourth Linguistic Annotation Workshop (LAW)*, pages 29–37, Uppsala, Suède, juillet 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W10-1804.

Beatrice Alex, Malvina Nissim et Claire Grover : The impact of annotation on the performance of protein tagging in biomedical text. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 595–600, Gène, Italie, mai 2006. URL http://www.ltg.ed.ac.uk/np/publications/ltg/papers/Alex2006Impact.pdf.

Erick Alphonse, Sophie Aubin, Philippe Bessières, Gilles Bisson, Thierry Hamon, Sandrine Laguarigue, Adeline Nazarenko, Alain-Pierre Manine, Claire Nédellec, Mohamed Ould Abdel Vetah, Thierry Poibeau et Davy Weissenbacher : Event-based information extraction for the biomedical the CADERIGE project. *In Proceedings of the JNLPBA COLING 2004 Workshop*, Genève, Suisse, août

2004. URL http://www-lipn.univ-paris13.fr/~poibeau/articles/nlpba04.pdf.

Emilia Apostolova, Sean Neilan, Gary An, Noriko Tomuro et Steven Lytinen : Djangology : A light-weight web-based tool for distributed collaborative text annotation. *In* Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Mike Rosner Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, éditeur : *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, La Valette, Malte, mai 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/543_Paper.pdf.

Xabier Artola, A. Díaz de Ilarraza, Nerea Ezeiza, Koldo Gojenola, Aitor Sologaistoa et Aitor Soroa : EULIA : a graphical web interface for creating, browsing and editing linguistically annotated corpora. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC) Workshop on XML-based richly annotated corpora*, Lisbonne, Portugal, 2004. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.6.1082.

Ron Artstein et Massimo Poesio : Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008. ISSN 0891-2017. URL http://www.mitpressjournals.org/doi/abs/10.1162/coli.07-034-R2.

Collin F. Baker, Charles J. Fillmore et John B. Lowe : The Berkeley FrameNet project. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) and International Conference on Computational Linguistics (ICCL)*, ACL'98, pages 86–90, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/980845.980860.

Béatrice Bakhouche, Béatrice Beys, Daniel Delattre, Charles Guérin et Trung Tran : De l'annotation aux marginalia. http ://meticebeta.univ-montp3.fr/lelivre/partie2/de_lannotation_aux_marginalia.html, juin 2010. URL http://meticebeta.univ-montp3.fr/lelivre/partie2/de_lannotation_aux_marginalia.html.

Lucie Barque, Alexis Nasr et Alain Polguère : From the definitions of the trésor de la langue française to a semantic database of the French language. *In Proceedings of the 14th EURALEX International Congress*, Leeuwarden, Pays-Bas, 2010. URL http://lucie.barque.free.fr/mesdocs/euralex2010.pdf.

Claude Barras, Edouard Geoffrois, Zhibiao Wu et Mark Liberman : Transcriber : a free tool for segmenting, labeling and transcribing speech. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1373–1376, Grenade, Espagne, mai 1998.

Petra Saskia Bayerl et Karsten Ingmar Paul : What determines inter-coder agreement in manual annotations ? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725, décembre 2011. ISSN 0891-2017. URL http://dx.doi.org/10.1162/COLI_a_00074.

Kent Beck : Manifesto for agile software development. http ://agilemanifesto.org/, 02 2011. URL http://agilemanifesto.org/.

Edward M. Bennett, R. Alpert et A. C.Goldstein : Communications through limited questioning. *Public Opinion Quarterly*, 18(3):303–308, 1954.

Christophe Benzitoun, Karën Fort et Benoît Sagot : TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, pages 99–112, Grenoble, France, juin 2012. URL http://www.aclweb.org/anthology/F/F12/F12-2008.pdf.

Yves Bestgen : Quels indices pour mesurer l'efficacité en segmentation thématique ? *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, page p. 10, Senlis, France, juin 2009. URL lipn.univ-paris13.fr/taln09/pdf/TALN_94.pdf.

Vikas Bhardwaj, Rebecca Passonneau, Ansaf Salleb-Aouissi et Nancy Ide : Anveshan : A tool for analysis of multiple annotators' labeling behavior. *In Proceedings of the fourth linguistic annotation workshop (LAW IV)*, Uppsala, Suède, 2010. URL www.aclweb.org/anthology/W/W10/W10-1806.pdf.

Steven Bird, David Day, John S. Garofolo, John Henderson, Christophe Laprun et Mark Liberman : ATLAS : a flexible and extensible architecture for linguistic annotation. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1699–1706, Athènes, Grèce, 2000. ELRA. URL http://arxiv.org/abs/cs/0007022v1.

Steven Bird et Mark Liberman : A formal framework for linguistic annotation. *Speech Communication*, 33(1-2):23–60, 2001.

Alena Böhmová, Jan Hajič, Eva Hajičová et Barbora Hladká : The prague dependency treebank : Three-level annotation scenario. *In* Anne Abeillé, éditeur : *Treebanks : Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers, 2001. URL http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/References/Czech_PDT.pdf.

Hélène Bonneau-Maynard, Sophie Rosset, Christelle Ayache, Anne Kuhn et Djamel Mostefa : Semantic annotation of the French Media dialog corpus. *In Proceedings of the InterSpeech*, Lisbonne, Portugal, septembre 2005. URL ftp://tlp.limsi.fr/public/IS052010.PDF.

Kalina Bontcheva, Hamish Cunningham, Ian Roberts et Valentin Tablan : Web-based collaborative corpus annotation : Requirements and a framework implementation. *In* René Witte, Hamish Cunningham, Jon Patrick, Elena Beisswanger, Ekaterina Buyko, Udo Hahn, Karin Verspoor et Anni R. Coden, éditeurs : *Proceedings of the workshop on New Challenges for NLP Frameworks (NLPFrameworks 2010)*, La Valette, Malte, mai 2010. ELRA. URL http://www.lrec-conf.org/proceedings/lrec2010/workshops/W10.pdf.

Eric Brill : A simple rule-based part of speech tagger. *In* Association for Computational Linguistics, éditeur : *Proceedings of the workshop on Speech and Natural Language (HLT'91)*, pages 112–116, Morristown, NJ, USA, 1992. URL http://acl.ldc.upenn.edu/H/H92/H92-1022.pdf.

Manuel Burghardt : Usability recommendations for annotation tools. *In Proceedings of the Sixth Linguistic Annotation Workshop*, pages 104–112, Jeju, République de Corée, juillet 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W12-3613.

Chris Callison-Burch et Mark Dredze : Creating speech and language data with Amazon's Mechanical Turk. *In CSLDAMT '10 : Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Morristown, NJ, USA, 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W10-0701.

Marie Candito et Djamé Seddah : Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, Grenoble, France, juin 2012. URL http://hal.inria.fr/hal-00698938.

Emmanuelle Canut, Virginie André et Bertrand Gaiffe : Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement des Corpus Oraux en Français). *Pratiques : théorie, pratique, pédagogie*, Interactions et Corpus Oraux:147–148, 2010. URL http://hal.archives-ouvertes.fr/hal-00523397.

Jean Carletta : Assessing agreement on classification tasks : the kappa statistic. *Computational Linguistics*, 22:249–254, 1996. URL http://acl.ldc.upenn.edu/J/J96/J96-2004.pdf.

Marc Carmen, Paul Felt, Robbie Haertel, Deryle Lonsdale, Peter McClanahan, Owen Merkling, Eric Ringger et Kevin Seppi : Tag dictionaries accelerate manual annotation. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, La Valette, Malte, mai 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/451_Paper.pdf.

Jonathan Chamberlain, Karën Fort, Udo Kruschwitz, Mathieu Lafourcade et Massimo Poesio : *Theory and Applications of Natural Language Processing*, chapitre Using Games to Create Language Resources : Successes and Limitations of the Approach. Theory and Applications of Natural Language Processing. Springer, 2012. A paraître.

Jonathan Chamberlain, Udo Kruschwitz et Massimo Poesio : Constructing an anaphorically annotated corpus with non-experts : assessing the quality of collaborative annotations. *In Proceedings of the 2009 Workshop on The People's Web Meets NLP : Collaboratively Constructed Semantic Resources*, People's Web '09, pages 57–62, Stroudsburg, PA, USA, 2009a. Association for Computational Linguistics. ISBN 978-1-932432-55-8. URL http://portal.acm.org/citation.cfm?id=1699765.1699774.

Jonathan Chamberlain, Massimo Poesio et Udo Kruschwitz : Addressing the resource bottleneck to create large-scale annotated texts. *In STEP '08 : Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 375–380, Morristown, NJ, USA, 2008a. Association for Computational Linguistics. URL http://portal.acm.org/ft_gateway.cfm?id=1626511&type=pdf&coll=GUIDE&dl=ACM&CFID=66309750&CFTOKEN=80191814.

Jonathan Chamberlain, Massimo Poesio et Udo Kruschwitz : Phrase Detectives : a web-based collaborative annotation game. *In Proceedings of the International Conference on Semantic Systems (I-Semantics'08)*, Graz, Autriche, 2008b. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.162.362.

Jonathan Chamberlain, Massimo Poesio et Udo Kruschwitz : A new life for a dead parrot : Incentive structures in the phrase detectives game. *In Proceedings of WWW 2009*, Madrid, Espagne, avril 2009b. URL http://anawiki.essex.ac.uk/www2009.pdf.

Philipp Cimiano et Siegfried Handschuh : Ontology-based linguistic annotation. *In Proceedings of the ACL 2003 Workshop on Linguistic Annotation*, pages 14–21, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

Jacob Cohen : A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Jacob Cohen : Weighted kappa : Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968.

Kevin Bretonnel Cohen, Lynne Fox, Philip V. Ogren et Lawrence Hunter : Corpus design for biomedical natural language processing. *In Proceedings of the ACL-ISMB workshop on linking biological literature, ontologies and databases : mining biological semantics*, pages 38–45, 2005. URL http://acl.ldc.upenn.edu/W/W05/W05-1306.pdf.

David A. Cohn, Zoubin Ghahramani et Michael I. Jordan : Active learning with statistical models. *In* G. Tesauro, D. Touretzky et T. Leen, éditeurs : *Advances in Neural Information Processing Systems*, volume 7, pages 705–712. The MIT Press, 1995. URL ftp://publications.ai.mit.edu/ai-publications/pdf/AIM-1522.pdf.

Paul Cook et Suzanne Stevenson : Automatically identifying changes in the semantic orientation of words. *In* Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner et Daniel Tapias, éditeurs : *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, La Valette, Malte, mai 2010. ISBN 2-9517408-6-7. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/657_Paper.pdf.

Benoît Crabbé et Marie-Hélène Candito : Expériences d'analyses syntaxique statistique du français. *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, Avignon, France, 2008. URL http://www.linguist.univ-paris-diderot.fr/~mcandito/Publications/crabbecandi-taln2008-final.pdf.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva et Valentin Tablan : GATE : a framework and graphical development environment for robust NLP tools and applications. *In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, Philadelphia, USA, 2002. URL http://gate.ac.uk/sale/acl02/acl-main.pdf.

Sandipan Dandapat, Priyanka Biswas, Monojit Choudhury et Kalika Bali : Complex linguistic annotation - no easy way out ! a case from bangla and hindi POS labeling tasks. *In Proceedings of the third ACL Linguistic Annotation Workshop*, Singapour, 2009. URL http://www.aclweb.org/anthology/W/W09/W09-3002.pdf.

Mark Davies et Joseph L. Fleiss : Measuring agreement for multinomial data. *Biometrics*, 38(4):1047–1051, décembre 1982. URL http://www.jstor.org/stable/2529886.

David Day, Chad McHenry, Robyn Kozierok et Laurel Riek : Callisto : A configurable annotation workbench. *In Proceedings of the International Conference on Language Resources and Evaluation*, Lisbonne, Portugal, 2004. URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/612.pdf.

Eric Villemonte de la Clergerie : A collaborative infrastructure for handling syntactic annotations. *In Proceedings of the First International Workshop on Automated Syntactic Annotations for interoperable Language Resources*, Hong-Kong, Chine, janvier 2008. URL http://atoll.inria.fr/passage/docs/easyref.pdf.

Isin Demirsahin, Ihsan Yalcinkaya et Deniz Zeyrek : Pair annotation : Adaption of pair programming to corpus annotation. *In Proceedings of the Sixth Linguistic Annotation Workshop*, pages 31–39, Jeju, République de Corée, juillet 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W12-3605.

Deneme : How many turkers are there ? http ://groups.csail.mit.edu/uid/deneme/, décembre 2009. URL http://groups.csail.mit.edu/uid/deneme/?p=502.

Pascal Denis et Benoît Sagot : Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. *In Proceedings of the Pacific Asia Conference on Language Information and Computing (PACLIC)*, Hong-Kong, Chine, 2009. URL http://atoll.inria.fr/~sagot/pub/paclic09tagging.pdf.

Barbara Di Eugenio et Michael Glass : The kappa statistic : a second look. *Computational Linguistics*, 30(1):95–101, 2004. ISSN 0891-2017. URL http://dl.acm.org/ft_gateway.cfm?id=1005385&ftid=338723&dwn=1&CFID=114285513&CFTOKEN=84921344.

Stefanie Dipper, Michael Götze et Manfred Stede : Simple annotation tools for complex annotation tasks : an evaluation. *In* H. S. Thompson J. Carletta A. Witt, U. Heid et P. Wittenburg, éditeurs : *Proceedings of the International Conference on Language Resources and Evaluation (LREC) Workshop on XML-based richly annotated corpora (XBRAC)*, Lisbonne, Portugal, 2004. URL http://www.linguistics.ruhr-uni-bochum.de/~dipper/papers/xbrac04-sfb.pdf.

Sean P. Engelson et Ido Dagan : Minimizing manual annotation cost in supervised training from corpora. *In Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 319–326, Morristown, NJ, USA, 1996. Association for Computational Linguistics. URL http://acl.ldc.upenn.edu/P/P96/P96-1042.pdf.

Katrin Erk, Andrea Kowalski, Sebastian Padó et Manfred Pinkal : Towards a resource for lexical semantics : a large german corpus with extensive semantic annotation. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, ACL 03, pages 537–544, Morristown, NJ, USA, 2003. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/1075096.1075164.

Christiane Fellbaum : *WordNet : An Electronic Lexical Database*. MIT Press, 1998.

Paul Felt, Owen Merkling, Marc Carmen, Eric Ringger, Warren Lemmon, Kevin Seppi et Robbie Haertel : CCASH : A web application framework for efficient, distributed language resource development. *In* Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner et Daniel Tapias, éditeurs : *Proceedings of the*

International Conference on Language Resources and Evaluation (LREC), La Valette, Malte, mai 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/360_Paper.pdf.

David Ferrucci et Adam Lally : UIMA : an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10: 327–348, septembre 2004. ISSN 1351-3249. URL http://dl.acm.org/citation.cfm?id=1030318.1030325.

R. H. Finn : A note on estimating the reliability of categorical data. *Educational and Psychological Measurement*, 30:71–76, 1970.

Joseph L. Fleiss : Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 5:378–382, 1971.

Karën Fort, Gilles Adda et Kevin Bretonnel Cohen : Amazon Mechanical Turk : Gold mine or coal mine ? *Computational Linguistics (editorial)*, 37(2):413–420, juin 2011a. URL http://aclweb.org/anthology-new/J/J11/J11-2010.pdf.

Karën Fort et Vincent Claveau : Annotating football matches : Influence of the source medium on manual annotation. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turquie, mai 2012a. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/623_Paper.pdf. 6 pages.

Karën Fort et Vincent Claveau : Annotation manuelle de matchs de foot : Oh la la ! l'accord inter-annotateurs ! et c'est le but ! *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, pages 383–390, Grenoble, France, juin 2012b. URL http://www.aclweb.org/anthology/F/F12/F12-2031. Poster.

Karën Fort, Maud Ehrmann et Adeline Nazarenko : Vers une méthodologie d'annotation des entités nommées en corpus ? *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, Senlis, France, juin 2009. URL http://hal.archives-ouvertes.fr/hal-00402321/en/. 11 pages.

Karën Fort, Claire François, Olivier Galibert et Maha Ghribi : Analyzing the impact of prevalence on the evaluation of a manual annotation campaign. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turquie, mai 2012a. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/549_Paper.pdf. 7 pages.

Karën Fort, Claire François et Maha Ghribi : Evaluer des annotations manuelles dispersées : les coefficients sont-ils suffisants pour estimer l'accord inter-annotateurs ? *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, Montréal, Canada, juillet 2010. URL http://www.iro.umontreal.ca/~felipe/TALN2010/Xml/Papers/all/taln2010_submission_6.pdf. 10 pages.

Karën Fort et Bruno Guillaume : Sylva : plate-forme de validation multi-niveaux de lexiques. *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, Avignon, France, juin 2008. URL http://hal.archives-ouvertes.fr/hal-00336290/en/. 10 pages. Poster.

Karën Fort, Adeline Nazarenko et Claire Ris : Corpus linguistics for the annotation manager. *In Proceedings of the Corpus Linguistics Conference*, Birmingham, Angleterre, juillet 2011b. URL http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-195.pdf. 13 pages.

Karën Fort, Adeline Nazarenko et Sophie Rosset : Modeling the complexity of manual annotation tasks : a grid of analysis. *In Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 895–910, Mumbai, Inde, décembre 2012b.

Karën Fort et Benoît Sagot : Influence of pre-annotation on POS-tagged corpus development. *In Proceedings of the Fourth ACL Linguistic Annotation Workshop*, pages 56–63, Uppsala, Suède, juillet 2010. URL http://aclweb.org/anthology-new/W/W10/W10-1807.pdf.

Eduard Frunzeanu et Philippe Pons : Les "encyclopédies" médiévales et les digital humanities : l'évolution du programme sourcencyme. Journée d'étude sur l'annotation collaborative de corpus, mars 2012. URL http://www.msh-lorraine.fr/fileadmin/images/actualites/actus2012/Axe2_AnnotCorpus_9mars/AnnoColCorpus_EFrunzeanu_PPons.pdf.

Olivier Galibert, Ludovic Quintard, Sophie Rosset, Pierre Zweigenbaum, Claire Nédellec, Sophie Aubin, Laurent Gillard, Jean-Pierre Raysz, Delphine Pois, Xavier Tannier, Louise Deléger et Dominique Laurent : Named and specific entity detection in varied data : the Quaero named entity baseline evaluation. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2010. URL ftp://tlp.limsi.fr/public/191_Paper.pdf.

Olivier Galibert, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum et Ludovic Quintard : Extended named entities annotation in OCRed documents : From corpus constitution to evaluation campaign. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turquie, 2012. ELRA. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/343_Paper.pdf.

Sylvain Galliano, Guillaume Gravier et Laura Chaubard : The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. *In Interspeech 2009*, Brighton, Angleterre, 2009. URL http://www.afcp-parole.org/etape/docs/interspeech09-ester-final.pdf.

Roger Garside, Geoffrey Leech et Tony McEnery, éditeurs. *Corpus Annotation : Linguistic Information from Computer Text Corpora*. Longman, Londres, Angleterre, 1997.

Dan Gillick et Yang Liu : Non-expert evaluation of summarization systems is risky. *In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 148–151, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=1866696.1866718.

Frédéric Godefroy : *Complément du dictionnaire de l'ancienne langue française et de tous ses dialectes du IXème au XVème siècle*. 1895-1902. URL http://micmap.org/dicfro/introduction/complement-godefroy.

Daniela Goecke, Harald Lüngen, Dieter Metzing, Maik Stührenberg et Andreas Witt : Different views on markup. *In* Andreas Witt, Dieter Metzing et Nancy Ide, éditeurs : *Linguistic Modeling of Information and Markup Languages*, volume 40 de *Text, Speech and Language Technology*, pages 1–21. Springer Netherlands, 2010. ISBN 978-90-481-3331-4. URL http://dx.doi.org/10.1007/978-90-481-3331-4_1.

Annette M. Green : Kappa statistics for multiple raters using categorical classifications. *In Proceedings of the Twenty-Second Annual Conference of SAS Users Group*, San Diego, USA, 1997. URL http://www2.sas.com/proceedings/sugi22/POSTERS/PAPER241.PDF.

Ralph Grishman : TIPSTER architecture design document version 3.1. Rapport technique, DARPA, 1998. URL www-nlpir.nist.gov/related_projects/tipster/docs/arch31.doc.

Ralph Grishman et Beth Sundheim : Message Understanding Conference-6 : a brief history. *In Proceedings of the the 16th conference on Computational linguistics*, pages 466–471, Morristown, NJ, USA, 1996. Association for Computational Linguistics. URL http://acl.ldc.upenn.edu/C/C96/C96-1079.pdf.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert et Ludovic Quintard : Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. *In Proceedings of the 5th Linguistic Annotation Workshop*, pages 92–100, Portland, Oregon, USA, juin 2011. URL http://www.aclweb.org/anthology/W11-0411. Poster.

Ulrike Gut et Petra Saskia Bayerl : Measuring the reliability of manual annotations of speech corpora. *In Proceedings of the Speech Prosody*, pages 565–568, Nara, Japon, 2004. URL http://www.uni-giessen.de/germanistik/ascl/dfg-projekt/pdfs/SP2004_GutBayerl03.pdf.

Kilem Li Gwet : *Handbook of Inter-rater Reliability*. Advanced Analytics, LLC, third édition, 2012.

Benoît Habert : *Corpus. Méthodologie et applications linguistiques*, chapitre Détournements d'annotation : armer la main et le regard, pages 106–120. Champion and Presses Universitaires de Perpignan, 2000.

Benoît Habert : Portrait de linguiste(s) à l'instrument. *Texto !*, vol. X (4), 2005. URL http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html.

Hilda Hardy, Kurk Baker, Laurence Devillers, Lori Lamel, Sophie Rosset, Tomek Strzalkowski, Cristian Ursu et Nick Webb : Multi-layer dialogue annotation for automated multilingual customer service. *In ISLE workshop*, Edimbourg, Grande Bretagne, décembre 2002. URL http://www.research.att.com/~walker/isle-dtag-wrk/.

Jisup Hong et Collin F. Baker : How good is the crowd at "real" WSD ? *In Proceedings of the 5th Linguistic Annotation Workshop*, pages 30–37, Portland, Oregon, USA, juin 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W11-0404.

George Hripcsak et Daniel F. Heitjan : Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, 35(2):99–110, 2002. ISSN 1532-0464.

George Hripcsak et Adam S Rothschild : Agreement, the f measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association (JAMIA)*, 12(3):296–298, 2005. URL http://ukpmc.ac.uk/articles/PMC1090460.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore et Rebecca Passonneau : MASC : the manually annotated sub-corpus of american english. *In* Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, éditeur : *Proceedings of the International Language Resources and Evaluation (LREC)*, Marrakech, Maroc, mai 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/617_paper.pdf.

Nancy Ide, Laurent Romary et Eric de la Clergerie : International standard for a linguistic annotation framework. *Journal of Natural Language Engineering*, 10, 2003.

Nancy Ide et Keith Suderman : GrAF : A graph-based format for linguistic annotations. *In Proceedings of the Linguistic Annotation Workshop, held in conjunction with ACL 2007*, pages 1–8, Prague, République Tchèque, juin 2007. URL http://www.cs.vassar.edu/~ide/papers/LAW.pdf.

Nany Ide et Laurent Romary : Representing linguistic corpora and their annotations. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Gène, Italie, 2006. URL http://www.cs.vassar.edu/~ide/papers/LAF-LREC06.pdf.

Julien Jourde, Alain-Pierre Manine, Philippe Veber, Karën Fort, Robert Bossy, Erick Alphonse et Philippe Bessières : BioNLP shared task 2011 – bacteria gene interactions and renaming. *In Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 65–73, Portland, Oregon, USA, juin 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W11-1810.

Michael Kaisser et John B. Lowe : Creating a research collection of question answer sentence pairs with Amazon's Mechanical Turk. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2008. URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/565_paper.pdf.

Dain Kaplan, Ryu Iida, Kikuko Nishina et Takenobu Tokunaga : Slate - a tool for creating and maintaining annotated corpora. *Journal for Language Technology and Computational Linguistics*, 26(2):89 – 101, janvier 2012. URL http://www.cl.cs.titech.ac.jp/publication/archive/673.pdf.

Dain Kaplan, Ryu Iida et Takenobu Tokunaga : Annotation process management revisited. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 365 – 366, La Valette, Malte, mai 2010. URL http://www.cl.cs.titech.ac.jp/publication/archive/655.pdf.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi et Jun'ichi Tsujii : GENIA corpus–a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl. 1):180–182, 2003. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/suppl_1/i180.

Jin-Dong Kim, Tomoko Ohta et Jun'ichi Tsujii : Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1):10, 2008. ISSN 1471-2105. URL http://www.biomedcentral.com/1471-2105/9/10.

Jin-Dong Kim et Yue Wang : CSAF - a community-sourcing annotation framework. *In Proceedings of the Sixth Linguistic Annotation Workshop*, pages 153–156, Jeju, République de Corée, juillet 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W12-3622.

Beata Beigman Klebanov et Eyal Beigman : From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503, décembre 2009. URL http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2009.35.4.35402. From Maud Ehrmann.

Klaus Krippendorff : *Content Analysis : An Introduction to Its Methodology*, chapitre 12. Sage, Beverly Hills, CA., USA, 1980.

Klaus Krippendorff : *Content Analysis : An Introduction to Its Methodology, second edition*, chapitre 11. Sage, Thousand Oaks, CA., USA, 2004.

Udo Kruschwitz, Jon Chamberlain et Massimo Poesio : (linguistic) science through web collaboration in the ANAWIKI project. *In Proceedings of the WebSci'09 : Society On-Line*, Athènes, Grèce, mars 2009. URL http://journal.webscience.org/186/2/websci09_submission_90.pdf.

Henry Kucera et W. Nelson Francis : *Computational Analysis of Present-Day American English*. Brown University Press, Providence, Rhode Island, USA, 1967.

Solomon Kullback et Richard Leibler : On information and sufficiency. *Annals of Mathematical Statistics (en)*, pages 79–86, 1951.

Anna Kupsc : Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré. *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, Toulouse, France, juin 2007. URL http://llf.linguist.jussieu.fr/llf/Gens/Kupsc/kupsc-taln07.pdf.

Mathieu Lafourcade et Alain Joubert : JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. *In* Ecole normale supérieure Lettres et sciences humaines, éditeur : *Actes de Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Lyon, France, mars 2008. URL http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2008/pdf/lafourcade-joubert.pdf.

Mathieu Lafourcade, Alain Joubert, Didier Schwab et Michael Zock : Evaluation et consolidation d'un réseau lexical grâce à un assistant ludique pour le mot sur le bout de la langue. *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, pages 295–306, Montpellier, France, June 2011. URL http://www.lirmm.fr/~lopez/TALN2011/PDF_long/Lafourcade_taln11_submission_52.pdf.

Marion Laignelet et François Rioult : Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, Senlis, France, juin 2009. URL http://www-lipn.univ-paris13.fr/taln09/pdf/TALN_77.pdf.

J. Richard Landis et Gary G. Koch : The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006341X. URL http://www.jstor.org/stable/2529310.

Frederic Landragin, Thierry Poibeau et Bernard Victorri : ANALEC : a new tool for the dynamic annotation of textual data. *In* Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk et Stelios Piperidis, éditeurs : *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turquie, mai 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/638_Paper.pdf.

Thomas Lebarbé : CLELIA : Building a manucript archive throught interdisciplinary dialogue. *In The Marriage of Mercury and Philology : Problems and Outcomes in Digital Philology*, 2008. URL http://w3.u-grenoble3.fr/lebarbe/uploads/Main/pres_TL_edimbourg.pdf.

Geoffrey Leech : Corpus annotation schemes. *Literary and Linguistic Computing*, 8(4):275–281, 1993. URL http://llc.oxfordjournals.org/content/8/4/275.abstract.

Geoffrey Leech : *Corpus annotation : Linguistic information from computer text corpora*, chapitre Introducing corpus annotation, pages 1–18. Longman, Londres, Angleterre, 1997.

Geoffrey Leech : *Developing Linguistic Corpora : a Guide to Good Practice*, chapitre Adding Linguistic Annotation, pages 17–29. Oxford : Oxbow Books, 2005. URL http://ota.ahds.ac.uk/documents/creating/dlc/index.htm.

Gaelle Lortal, Myriam Lewkowicz et Amalia Todiracu-Courtier : *Annotations dans les documents pour l'action*, chapitre Des activités d'annotation : De la glose au document, pages 153–171. Hermes Publishing, Londres-Paris, 2006. URL http://lewkowicz.tech-cico.fr/publi/Hermes_annotation_Lortal_Lewkowicz_Todirascu.pdf.

Brian MacWhinney : *The CHILDES Project : Tools for Analyzing Talk*. Lawrence Erlbaum Associates, 3 édition, 2000. URL http://childes.psy.cmu.edu/manuals/chat.pdf.

Kazuaki Maeda, Haejoong Lee, Julie Medero et Stephanie Strassel : A new phase in annotation tool development at the linguistic data consortium : The evolution of the Annotation Graph Toolkit. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Gène, Italie, mai 2006. URL http://ssli.ee.washington.edu/people/jmedero/publications/2006_lrec_b.pdf.

Kazuaki Maeda et Stephanie Strassel : Annotation tools for large-scale corpus development : Using AGTK at the Linguistic Data Consortium. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Lisbonne, Portugal, 2004. URL http://papers.ldc.upenn.edu/LREC2004/AGTK.pdf.

John Makhoul, Francis Kubala, Richard Schwartz et Ralph Weischedel : Performance measures for information extraction. *In Proceedings of DARPA Broadcast News Workshop*, pages 249–252, 1999. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.27.4637.

Mitchell Marcus, Beatrice Santorini et Mary Ann Marcinkiewicz : Building a large annotated corpus of English : The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. URL http://acl.ldc.upenn.edu/J/J93/J93-2004.pdf.

Yann Mathet et Antoine Widlöcher : Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs. *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, Montpellier, France, juin 2011a. URL http://www.lirmm.fr/~lopez/TALN2011/PDF_long/Mathet_taln11_submission_78.pdf.

Yann Mathet et Antoine Widlöcher : Stratégie d'exploration de corpus multi-annotés avec GlozzQL. *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, Montpellier, France, juin 2011b. URL http://www.lirmm.fr/~lopez/TALN2011/PDF_court/Mathet_taln11_submission_186.pdf.

Yann Mathet, Antoine Widlöcher, Karën Fort, Claire François, Olivier Galibert, Cyril Grouin, Juliette Kahn, Sophie Rosset et Pierre Zweigenbaum : Manual corpus annotation : Evaluating the evaluation metrics. *In Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 809–818, Mumbai, Inde, décembre 2012. Poster.

Marie Mikulová et Jan Štěpánek : Annotation quality checking and its implications for Design of treebank (in building the prague czech-english Dependency treebank). *In Proceedings of the Eight International Workshop on Treebanks and Linguistic Theories*, volume 4-5, Milan, Italie, décembre 2009. URL http://tlt8.unicatt.it/FullPaper/B_2.pdf.

Thomas Morton et Jeremy LaCivita : WordFreak : an open tool for linguistic annotation. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology : Demonstrations - Volume 4*, NAACL-Demonstrations '03, pages 17–18, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/1073427.1073436.

Christoph Müller et Michael Strube : Multi-level annotation of linguistic data with MMAX2. *In* Sabine Braun, Kurt Kohn et Joybrato Mukherjee, éditeurs : *Corpus Technology and Language Pedagogy : New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfort, Allemagne, 2006. URL http://mmax2.sourceforge.net/mmaxpaper.pdf.

Denis Muzerelle : *Vocabulaire codicologique : répertoire méthodique des termes français relatifs aux manuscrits*. Editions CEMI, Paris, 1985. URL http://vocabulaire.irht.cnrs.fr/pages/vocab2.htm.

Adeline Nazarenko, Erick Alphonse, Julien Derivière, Thierry Hamon, Guillaume Vauvert et Davy Weissenbacher : The ALVIS format for linguistically annotated documents. *In Proceedings of the International conference on Language Resources and Evaluation (LREC)*, pages 1782–1786, Gène, Italie, 2006. ELDA. URL http://hal.archives-ouvertes.fr/hal-00080472/en/.

Claire Nédellec, Philippe Bessières, Robert Bossy, Alain Kotoujansky et Alain-Pierre Manine : Annotation guidelines for machine learning-based named entity recognition in microbiology. *In* M. Hilario et C. Nédellec, éditeur : *Proceedings of the Data and text mining in integrative biology workshop*, pages 40–54, Berlin, Allemagne, septembre 2006. URL http://genome.jouy.inra.fr/~cnedelle/Docs/NedellecECML06.pdf.

Scott Novotney et Chris Callison-Burch : Cheap, fast and good enough : automatic speech recognition with non-expert transcription. *In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, HLT'10, pages 207–215, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5. URL http://portal.acm.org/citation.cfm?id=1857999.1858023.

Mick O'Donnell : Demonstration of the UAM CorpusTool for text and image annotation. *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies : Demo Session*, HLT-Demonstrations '08, pages 13–16, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1564144.1564148.

Philip Ogren : Knowtator : A plug-in for creating training and evaluation data sets for biomedical natural language systems. *In Protégé Conference*, Stanford, USA, 2006. URL http://knowtator.sourceforge.net/docs/Ogren_ProtegeConference2006_KnowtatorAbstract.pdf.

Constantin Orasan : Palinka : A highly customisable tool for discourse annotation. *In Proceedings of SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japon, 2003. URL http://www.sigdial.org/workshops/workshop4/proceedings/29_SHORT_orasan_palinka-final.pdf.

Patrick Paroubek, Martine Hurault-Plantet, Catherine Gouttas et Alexander Pak : Un corpus de référence pour l'évaluation de la fouille d'opinion dans le contexte industriel du projet DOXA. *In Actes d'EvalECD 2010*, 2010. URL http://www.lirmm.fr/~bechet/EvalECD/actes-evalecd-2010.pdf.

Rebecca J. Passonneau : Computing reliability for coreference annotation. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1503–1506, Lisbonne, Portugal, 2004. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.132.7127.

Georgios Petasis : The SYNC3 collaborative annotation tool. *In* Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk et Stelios Piperidis, éditeurs : *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turquie, mai 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/700_Paper.pdf.

Lev Pevzner et Marti A. Hearst : A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, mars 2002. ISSN 0891-2017. URL http://dx.doi.org/10.1162/089120102317341756.

Massimo Poesio et Ron Artstein : The reliability of anaphoric annotation, reconsidered : Taking ambiguity into account. *In Proceedings of the Workshop on Frontiers in Corpus Annotations II : Pie in the Sky*, pages 76–83, Ann Arbor, Michigan, USA, juin 2005. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W05/W05-0311.pdf. Implicit ambiguity much more frequent than explicit ambiguity.

Adam Przepiórkowski et Grzegorz Murzynowski : Manual annotation of the national corpus of Polish with Anotatornia. *In* Stanisław Goźdź-Roszkowski, éditeur : *The proceedings of Practical Applications in Language and Computers PALC 2009*, Frankfort, Allemagne, 2009. Peter Lang. URL http://nlp.ipipan.waw.pl/~adamp/Papers/2009-palc-anotatornia/paper.pdf. Forthcoming.

Sylvie Ranwez et Michel Crampes : Méta-description en XML de documents vidéo. *In Actes de la conférence ISKO'99*, pages 12/1–12/8, Lyon, France, 1999. URL http://hal.archives-ouvertes.fr/hal-00371300/en/.

Adwait Ratnaparkhi : A maximum entropy model for part-of-speech tagging. *In Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pages 133–142, 1996. URL http://acl.ldc.upenn.edu/W/W96/W96-0213.pdf.

Christian Raymond et Julien Fayolle : Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, Montréal, Canada, juillet 2010. URL http://hal.archives-ouvertes.fr/inria-00561732/.

Ines Rehbein, Josef Ruppenhofer et Caroline Sporleder : Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation. *In Proceedings of*

the Third Linguistic Annotation Workshop, pages 19–26, Singapour, août 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W09/W09-3003.

Denis Reidsma et Jean Carletta : Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326, septembre 2008. URL http://doc.utwente.nl/64823/.

Dennis Reidsma, Natasa Jovanovic et Dennis Hofs : Designing annotation tools based on properties of annotation problems, 2004. URL http://doc.utwente.nl/49282/.

Sophie Rosset, Cyril Grouin, Karën Fort, Olivier Galibert, Juliette Kahn et Pierre Zweigenbaum : Structured named entities in two distinct press corpora : Contemporary broadcast news and old newspapers. *In Proceedings of the 6th Linguistic Annotation Workshop (LAW VI)*, pages 40–48, Jeju, République de Corée, juillet 2012. URL http://www.aclweb.org/anthology/W12-3606.

Sophie Rosset, Cyril Grouin et Pierre Zweigenbaum : *Entités Nommées Structurées : guide d'annotation Quæro*. LIMSI–CNRS, Orsay, France, 2011. URL http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf.

Benoît Sagot et Pierre Boullier : SxPipe 2 : architecture pour le traitement présyntaxique de corpus bruts. *Traitement Automatique des Langues*, 49(2):155–188, 2008. URL http://hal.inria.fr/inria-00515489.

Benoît Sagot, Karën Fort, Gilles Adda, Joseph Mariani et Bernard Lang : Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé. *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, Montpellier, France, juin 2011. URL http://hal.inria.fr/inria-00617067/. 12 pages.

Geoffrey Sampson : The role of taxonomy in language engineering. *Philosophical Transactions of the Royal Society of London. Series A :Mathematical, Physical and Engineering Sciences*, 358(1769):1339–1355, 2000. URL http://rsta.royalsocietypublishing.org/content/358/1769/1339.abstract.

Beatrice Santorini : Part-of-speech tagging guidelines for the Penn Treebank Project. Rapport technique MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania, 1990. URL ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz.

Helmut Schmid : *New Methods in Language Processing, Studies in Computational Linguistics*, chapitre Probabilistic part-of-speech tagging using decision trees, pages 154–164. UCL Press, London, 1997. URL http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

William A Scott : Reliability of content analysis : The case of nominal scale coding. *Public Opinion Quaterly*, 19(3):321–325, 1955.

Satoshi Sekine : Definition, dictionaries and tagger of extended named entity hierarchy. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Lisbonne, Portugal, 2004. ELRA. URL http://cs.nyu.edu/~sekine/papers/lrec04-65.pdf.

Sidney Siegel et N. John Castellan : *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, USA, 2nd édition, 1988.

John Sinclair : *Developing Linguistic Corpora : a Guide to Good Practice*, chapitre Corpus and Text - Basic Principles, pages 1–16. Oxford : Oxbow Books, 2005. URL http://ahds.ac.uk/linguistic-corpora/.

Rion Snow, Brendan O'Connor, Daniel Jurafsky et Andrew Y. Ng. : Cheap and fast - but is it good ? evaluating non-expert annotations for natural language tasks. *In Proceedings of EMNLP 2008*, pages 254–263, 2008. URL http://www.stanford.edu/~jurafsky/amt.pdf.

Drahomíra "Johanka" Spoustová, Jan Hajič, Jan Raab et Miroslav Spousta : Semi-supervised Training for the Averaged Perceptron POS Tagger. *In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 763–771, Morristown, NJ, USA, 2009.

Pontus Stenetorp, Goran Topić, Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim et Jun'ichi Tsujii : BioNLP shared task 2011 : Supporting resources. *In Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 112–120, Portland, Oregon, USA, juin 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W11-1816.

Maik Stührenberg, Daniela Goecke, Nils Diewald, Alexander Mehler et Irene Cramer : Web-based annotation of anaphoric relations and lexical chains. *In Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 140–147, Prague, République Tchèque, juin 2007. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1642059.1642082.

Katrin Tomanek, Udo Hahn, Steffen Lohmann et Jürgen Ziegler : A cognitive cost model of annotations based on eye-tracking data. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, ACL'10, pages 1158–1167, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1858681.1858799.

Stephen Tratz et Eduard Hovy : A taxonomy, dataset, and classifier for automatic noun compound interpretation. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Uppsala, Suède, juillet

2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P10-1070.

Jean Véronis : Annotation automatique de corpus : panorama et état de la technique. *Ingénierie des langues*, pages 111–129, 2000. URL http://sites.univ-provence.fr/~veronis/pdf/2000hermes4.pdf.

Anne Vilnat, Patrick Paroubek, Eric Villemonte de la Clergerie, Gil Francopoulo et Marie-Laure Guénot : PASSAGE syntactic representation : a minimal common ground for evaluation. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, La Valette, Malte, 2010. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/603_Paper.pdf.

Luis von Ahn : Games with a purpose. *IEEE Computer Magazine*, pages 96–98, juin 2006. URL http://www.cs.cmu.edu/~biglou/ieee-gwap.pdf.

Luis von Ahn et Laura Dabbish : Labeling images with a computer game. *In Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, pages 319–326, New York, NY, USA, 2004. ACM. ISBN 1-58113-702-8. URL http://doi.acm.org/10.1145/985692.985733.

Holger Voormann et Ulrike Gut : Agile corpus creation. *Corpus Linguistics and Linguistic Theory*, 4(2):235–251, 2008.

Davy Weissenbacher et Adeline Nazarenko : A bayesian classifier for the recognition of impersonal it pronoun occurrences : description of the system. *In Proceedings of the NIPS Workshop on Bayesian Methods for NLP*, Whistler, Canada, 2005. URL http://hal.archives-ouvertes.fr/hal-00162003/.

Antoine Widlöcher et Yann Mathet : La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, Senlis, France, 2009. URL http://www-lipn.univ-paris13.fr/taln09/pdf/TALN_120.pdf.

Peter Wittenburg : About annotation schemes and terminology. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 39–45, Athènes, Grèce, 2000. URL http://www.mpi.nl/ISLE/documents/papers/wittenburg.pdf.

Marcin Woliński, Katarzyna Głowińska et Marek Świdziński : A preliminary version of składnica - a treebank of polish. *In Proceedings of the 5th Language and Technology Conference*, Poznań, Pologne, 2011.

Martin Wynne, éditeur. *Developing Linguistic Corpora : a Guide to Good Practice*. Oxford : Oxbow Books, 2005. URL http://ota.ahds.ac.uk/documents/creating/dlc/index.htm.

Fang Xu et Dietrich Klakow : Paragraph acquisition and selection for list question using amazon's mechanical turk. *In Proceedings of the International Language Resources and Evaluation (LREC)*, pages 2340–2345, La Valette, Malte, mai 2010. URL http://www.lsv.uni-saarland.de/241_Paper.pdf.

Annexe A

Outils d'aide à l'annotation existants

La présentation traditionnelle d'un état de la technique sous forme de tableau comparatif nous semble ici inadaptée car elle ne permet pas de représenter les différentes logiques sous-jacentes à chaque outil. Nous préférons donc détailler chacun d'entre eux dans son contexte de création et d'utilisation.

L'organisation de l'inventaire proposée ici repose sur une typologie des outils que nous avons mise au jour lors de nos tests. La section A.1 présente ainsi les outils génériques, soit qu'ils aient été conçus comme tels, soit qu'ils aient évolué vers plus de généricité. Logiquement, la section suivante introduit certains outils spécifiques à une tâche d'annotation. Les plate-formes d'annotation, dont la finalité est plus directement l'application d'outils de TAL sur des corpus textuels, sont présentées en section A.3. Enfin, la section A.4 détaille les outils de gestion d'annotations.

A.1 Outils génériques

Les outils suivants, s'ils se veulent aujourd'hui génériques, n'ont souvent pas été créés comme tels. Ils ont été conçus pour une tâche spécifique, puis se sont diversifiés. Ils ne permettent pas la gestion de l'annotation en tant que telle, même si certains incluent des fonctionnalités comme le calcul de l'accord inter-annotateurs.

A.1.1 `PALinkA`

Perspicuous and Adjustable Links Annotator (`PALinkA`) [1] est un programme Java gratuit permettant d'annoter manuellement un corpus. Cet outil (Orasan, 2003) fait suite à `CLinkA`, qui avait été créé pour l'annotation de co-référence, mais n'était pas adaptable à d'autres campagnes d'annotation. `PALinkA` se veut donc adaptable à de nombreux types d'annotation, notamment à la co-référence, mais également à l'annotation de phénomènes discursifs variés. Les annotations sont ajoutées dans le document lui-même. Il gère le XML, les éventuelles balises pré-existantes à l'annotation ne sont donc pas visibles. Il permet également d'annoter des relations, mais seul un type de relation peut être ajouté à un élément. Il est intéressant de noter que `PALinkA` permet

1. http://clg.wlv.ac.uk/projects/PALinkA/

d'enregistrer le temps passé sur l'annotation. L'outil, après une période d'inactivité, est de nouveau maintenu et la dernière version disponible (2 1.0.2) date de février 2011. Il est possible d'en obtenir les sources auprès du développeur. `Palinka` nécessite une tokenisation préalable, ainsi que la création d'un fichier de préférence. L'outil est en outre peu documenté, même si une FAQ vient d'être ajoutée sur le site Web de l'outil.

A.1.2 `Cadixe`

`Cadixe` (Alphonse *et al.*, 2004) est un outil d'aide à l'annotation qui a été développé dans le cadre des projets Caderige [2] et ExtraPloDocs [3].

`Cadixe` est un logiciel qui permet d'annoter un texte à l'aide de balises XML à partir d'une DTD. Il gère le XML, les balises existantes et ajoutées ne sont donc pas visibles par défaut, mais les annotations sont intégrées au fichier source. Son interface permet de choisir et de visualiser les balises sous la forme d'attributs graphiques (couleur, style, police, taille de la police, inversion vidéo, etc.) définis dans une feuille de style modifiable (CSS). Écrit en Java, il fonctionne sur tous les systèmes d'exploitation, mais c'est un client lourd, qui ne permet pas l'annotation collaborative, ni le calcul automatique de l'accord inter-annotateurs. `Cadixe` ne permet en outre pas d'annoter des relations en tant que telles.

L'outil, gratuit mais pas open source, est raisonnablement documenté mais n'est aujourd'hui plus maintenu et la dernière version disponible (2.0a4) date de mars 2005 [4].

Nous avons utilisé `Cadixe`, à la demande de l'INRA, notre partenaire sur ces projets, pour les campagnes d'annotation noms de gènes, de protéines et d'espèces (voir section 5.1), puis de relations de renommage de gènes (voir section 5.2).

A.1.3 `Callisto`

`Callisto` (Day *et al.*, 2004) [5] est un programme Java permettant d'annoter manuellement un corpus. Il est gratuit, mais non open source (freeware). La dernière version (1.5.2.1) date de 2009 mais l'outil est maintenu par l'organisation américaine MITRE [6] et la nouvelle version est annoncée pour bientôt (au 6 mars 2012). Les annotations y sont déportées suivant le modèle de données ATLAS (Bird *et al.*, 2000). Il gère le XML.

2. http://caderige.imag.fr/
3. http://www-lipn.univ-paris13.fr/fr/rcln-projets
4. http://caderige.imag.fr/Cadixe/
5. http://www.softpedia.com/get/Office-tools/Other-Office-Tools/Mitre-Callisto.shtml
6. http://www.mitre.org/

`Callisto` repose sur une architecture modulaire et configurable. Des services d'annotation sont accessibles à tous les modules d'annotation, modules qui doivent être développés à part. S'il existe un module pour les annotations de relations de type temporel, TANGO [7], aucun module générique permettant d'annoter des relations quelles qu'elles soient n'existe encore.

`Callisto` est un outil très agréable à utiliser, s'il est déjà configuré pour le type d'annotation envisagé (comme la résolution d'anaphore, par exemple), mais il est difficile de l'adapter pour une campagne d'un autre type.

Dans sa dernière version (non encore disponible), `Callisto` offre des fonctionnalités intéressantes : enregistrement du temps passé sur chaque document d'un corpus et possibilité de poser des questions aux annotateurs après l'annotation. Les développements actuels s'orientent vers la gestion d'annotations (calcul de l'accord inter-annotateurs, validation croisée, etc).

A.1.4 `Amazon Mechanical Turk`

`Amazon Mechanical Turk` [8] est une plate-forme de myriadisation du travail parcellisé avec rémunération (Sagot *et al.*, 2011), ou *microworking crowdsourcing*, créée en 2005 par l'entreprise de vente en ligne *Amazon.com*. Si elle n'est pas unique en son genre, cette plate-forme est la plus connue et la plus utilisée. Elle permet à des demandeurs (*Requesters*) de proposer des groupes de tâches simples (*Human Intelligence Task*, ou HIT) à réaliser en ligne à un grande nombre de personnes (*Turkers*). Le travail est parcellisé car il est simplifié à l'extrême pour pouvoir être réalisé par une myriade ou une foule (*crowd*) de non-experts.

`Amazon Mechanical Turk` propose des composants Web que le *Requester* peut utiliser à sa guise dans des formulaires comportant une ou plusieurs questions auxquelles le *Turker* doit répondre (Snow *et al.*, 2008). Tout l'« art » du *Requester* consiste donc à décomposer des tâches complexes, notamment d'annotation (voir, entre autres, (Snow *et al.*, 2008; Hong et Baker, 2011)), en tâches simples (voire simplifiée, voir (Cook et Stevenson, 2010)) et de fixer une rémunération, généralement très faible (par exemple, cinq cents américains pour traduire une phrase).

Par ce biais, des ressources langagières peuvent donc être produites à très bas prix (1/10e du coût usuel, au moins (Callison-Burch et Dredze, 2010)). De ce point de vue, au moins, `Amazon Mechanical Turk` ne peut pas être considéré comme un jeu tel que `Phrase Detectives` (voir sous-section A.2.4), dans lequel la rémunération n'est pas au cœur du système (seuls les meilleurs joueurs ont droit à un prix).

7. http://timeml.org/site/tango/tool.html
8. https://www.mturk.com

Nous avons montré les limites de ce système dans différentes articles de prise de position (Fort *et al.*, 2011a; Sagot *et al.*, 2011; Adda *et al.*, 2011), nous ne nous étendons donc pas davantage ici.

A.1.5 `Dexter`

`Dexter` [9] est un ensemble de programmes Java permettant d'annoter manuellement un corpus. C'est un outil gratuit, dont le développement, après avoir été stoppé pendant plusieurs années, vient de reprendre en 2011. Les sources sont disponibles auprès du développeur.

Le premier programme, l'annotateur `Dexter` (`Dexter Coder`) permet de réaliser des annotations manuelles. Il gère le XML, les annotations produites sont déportées mais l'outil ne permet pas d'annoter des relations. Il donne en revanche accès à une interface de recherche dans les annotations et le texte source, qui permet l'utilisation d'expressions régulières.

Le convertisseur (`Converter`), est un autre outil `Dexter` qui propose une interface d'import, dans le format XML de `Dexter`, des fichiers de l'utilisateur (texte ou XML).

A.1.6 `Knowtator`

`Knowtator` (Ogren, 2006) [10] se présente sous la forme d'un plugin Java de l'éditeur d'ontologies `Protégé` [11]. Il s'agit d'un outil gratuit et open source (Mozilla Public License 1.1) dont la dernière version (1.9-beta2) date de 2010. Il a été développé à l'origine pour l'annotation dans le domaine biomédical et son originalité principale est de proposer un lien direct avec une ontologie.

Son interface est directement inspirée de celle de `WordFreak` (Morton et LaCivita, 2003) et il permet d'annoter des unités et des relations de manière simple et claire. Les annotations y sont déportées. Il possède un mode « consensus » permettant de comparer et de fusionner les annotations de plusieurs annotateurs.

En revanche, il n'accepte que du texte brut en entrée et ne prend donc pas en compte le balisage XML en tant que tel. Par conséquent, les balises apparaissent en clair et « bruitent » l'affichage, ce qui n'est pas adapté au travail des annotateurs. Une solution simple consiste à « nettoyer » le XML avec une XSLT, par exemple, avant de l'importer dans `Knowtator`. Une autre limitation de cet outil, par ailleurs très utilisé dans le domaine biomédical, est qu'il n'est pas compatible avec la dernière version de Protégé.

9. http://www.dextercoder.org/
10. http://knowtator.sourceforge.net/
11. http://protege.stanford.edu/

A.1.7 MMAX2

MMAX2 (Müller et Strube, 2006) [12] est un programme Java permettant d'annoter manuellement un corpus. Les annotations sont déportées. Il gère le XML et permet en outre d'annoter des relations.

L'outil est open-source (licence Apache V2.0) et maintenu, la dernière version (1.13) date de 2010 et le développeur répond aux mails et aux échanges sur le forum.

MMAX2 fournit un tokeniser intégré. Il distingue deux type de relations, les relations non orientées (*sets*) et les relations orientées (*pointers*). MMAX2 permet de réaliser des annotations selon différents niveaux (*levels*), ce qui permet de construire un projet d'annotation complexe avec des niveaux d'annotation différents, tels que *phrase*, *chunk* et *POS*. Il permet d'annoter des zones non contiguës et des zones partiellement recouvrantes ("Lyonnais étaient réduits" et "réduits à 10"). Il permet également la factorisation.

MMAX2 présente cependant différents lacunes. Il est difficile à utiliser car son interface est complexe et les actions souris peu claires. Il devient d'une lenteur extrême lorsque les fichiers font plus de quelques pages, et la visualisation est peu efficace (par exemple, les couleurs des attributs ne se retrouvent pas dans le panneau des attributs). Enfin, sa documentation est limitée.

Dans ses dernières versions, MMAX2 propose un plugin *AnnotationDiffWindow* permettant de visualiser des annotations concurrentes de chaînes de co-référence et donc les accords ou désaccords entre annotateurs.

A.1.8 UAM CorpusTool

UAM CorpusTool (O'Donnell, 2008) est un outil d'annotation destiné aux linguistes, qui est facile à installer et à configurer. Il a été développé en Python et est disponible [13], mais non open source et ne fonctionne que sur les systèmes d'exploitation MS Windows et Mac OS.

Il permet la mise en place rapide de projets d'annotation, grâce à une interface dédiée à l'organisation des fichiers du corpus et à la création du schéma d'annotation. Il est cependant à noter que l'outil ne distingue pas les annotateurs du gestionnaire de projet.

Une caractéristique intéressante de UAM CorpusTool est de permettre l'édition du schéma d'annotation en cours de projet. L'interface permet également d'associer une glose explicative à chaque étiquette, glose pouvant être visualisée par l'annotateur pendant l'annotation.

12. http://mmax2.sourceforge.net/
13. http://www.wagsoft.com/CorpusTool/index.html

`UAM CorpusTool` prend en compte différents « niveaux » linguistiques d'annotation (par exemple, des informations sur le texte lui-même, les portions du texte par fonction, les phrases/clauses, etc), mais ne permet pas encore d'annoter des relations. Les annotations créées sont stockées en mode déporté, un fichier par niveau d'annotation.

Une interface de recherche relativement performante est disponible, qui permet la recherche de patrons lexicaux sur plusieurs niveaux grâce à un lexique inclus. Des traitements automatiques sont également disponibles (découpage en phrases, en *chunks*), ainsi qu'une interface d'analyse statistique fournissant des informations d'analyse textuelle sur le texte et les annotations.

A.1.9 `Glozz`

`Glozz` (Widlöcher et Mathet, 2009) est un outil d'annotation créé dans le cadre du projet ANR Annodis, qui visait la création d'un corpus de référence en analyse du discours. `Glozz` se présente sous la forme d'un programme Java et est disponible gratuitement pour la recherche sous réserve de s'inscrire sur le site Web [14]. Une des originalités de l'outil par rapport à d'autres est cependant sa généricité. En effet, `Glozz` se veut utilisable pour toutes sortes d'annotations, sans limite d'échelle.

Le méta-modèle générique de `Glozz` propose trois types de composants : les unités (éléments textuels adjacents), les relations (rapport binaire entre deux unités) et les schémas (configuration textuelle complexe récurrente impliquant unités et relations). Pour `Glozz`, rien n'existe hors de ce méta-modèle. Cela signifie par exemple que les pré-annotations, s'il y en a, doivent être exprimées dans ce méta-modèle. Cela ne pose en général pas de problème, mais cela peut devenir très lourd à gérer si le fichier d'origine est très « chargé » en balises XML, par exemple (ce qui était le cas lors de la campagne 4 d'annotation de brevets en pharmacologie, voir section 5.4).

`Glozz` propose une annotation déportée et prend en entrée du texte. Visuellement, l'outil est très agréable et confortable, montrant toutes les imbrications et relations complexes de façon claire. Une feuille de style permet de paramétrer les couleurs des catégories à annoter. Par ailleurs, un mode *glue note* permet d'ajouter des commentaires en texte libre.

`Glozz` présente la particularité d'envisager l'annotation sous l'angle de l'exploration de corpus, en permettant de masquer certaines informations pour une tâche donnée et en proposant différentes vues sur le texte (vue « ruban », vue principale). Il est également possible dans `Glozz` de spécifier le niveau d'imbrication minimal et maximal des unités à afficher pour alléger la visualisation (ainsi, en mettant les curseurs respectivement sur 2 et 4, ne s'affichent que les unités de niveau 2, 3 ou 4). Il propose enfin un outil

14. http://www.glozz.org/

de recherche, portant à la fois sur le texte lui-même et sur les annotations. Cependant, cet outil de recherche est peu pratique : son interface minimaliste est peu claire, la visualisation des résultats est difficile au milieu des annotations, il ne gère pas les expressions régulières, la recherche sur les annotations ne semble pas fonctionner.

Dans ses dernières versions (depuis la version 1.1.0-beta), `Glozz` propose une fonctionnalité de calcul de l'accord inter-annotateurs (Mathet et Widlöcher, 2011a), qui reste limitée aux unités et ne s'applique pas aux relations. Cette fonctionnalité s'accompagne d'un langage de requête sur les annotations et le texte de type SQL, GlozzQL. Si ce langage, très puissant, permet d'effectuer des recherches complexes, de sélectionner certaines « configurations » du corpus annoté et de vérifier certaines contraintes (Mathet et Widlöcher, 2011b), il est dans les faits lourd à manipuler, en particulier dès que le corpus atteint une taille de plus de quelques centaines de kilo-octets.

Malgré ces limitations, nous avons utilisé `Glozz` pour la campagne d'annotation de matchs de football, pour laquelle il était particulièrement adapté. Nous l'avons également utilisé pour le calcul de l'accord intra- et inter-annotateurs de la campagne (Fort et Claveau, 2012a). Cela étant, pour les raisons sus-citées, la correction du corpus devra être réalisée en dehors de `Glozz`.

A.1.10 `CCASH`

`CCASH`, ou *Cost-Conscious Annotation Supervised by Humans* (Felt *et al.*, 2010), est, comme son nom l'indique, « orienté coût », autrement dit, conçu pour intégrer des traitements automatiques permettant de limiter le travail effectif des annotateurs, tels que la propagation d'étiquettes (Carmen *et al.*, 2010), et pour mesurer le coût de l'annotation. Une originalité de l'outil est d'évaluer ce coût non seulement par le temps passé sur la tâche, mais également en nombre d'interactions nécessaires pour effectuer une action.

L'outil est développé sous forme d'une application Web, qui permet de gérer l'annotation de plusieurs annotateurs, éventuellement à distance. Il comprend une interface d'administration limitée, permettant à l'administrateur de créer des tâches d'annotation et d'assigner ces tâches à des utilisateurs. Une extension vers davantage de fonctionnalités de gestion est prévue pour 2013. Cet outil a donc vocation à devenir un outil plus large, d'annotation et de gestion d'annotation (voir annexe A.4). Les annotations sont déportées. CCASH permet l'annotation de relations.

La flexibilité et les possibilités de personnalisation sont au cœur du système.Ainsi, le client Web est développé à l'aide du `Google Web Toolkit` (JavaScript), qui propose des widgets variés. Le serveur est en Java et la persistance est assurée dans une base de données (au choix), *via* l'API Hibernate. Un fournisseur d'instances séparé permet d'intégrer d'éventuels traitements automatiques.

CCASH est disponible en open source sous licence AGPL [15] sur sourceforge [16]. Une documentation minimaliste est disponible sur un autre site Web [17]. L'outil est maintenu et toujours en cours d'amélioration.

A.1.11 brat

brat (Stenetorp *et al.*, 2011) a été créé à l'origine comme une extension d'un outil de visualisation d'annotations de texte et est aujourd'hui un outil d'annotation à part entière en version 1.2, proposant un environnement d'annotation collaborative en ligne. C'est un outil open source sous licence MIT [18] dont le serveur est écrit en Python avec une interface CGI. Les annotations sont déportées. Il est possible de le tester en ligne sur le site qui lui est dédié [19].

La spécificité de brat est de permettre aux annotateurs de travailler simultanément sur le même corpus, voire sur le même document, et de visualiser les modifications des uns et des autres au fur et à mesure de leurs ajouts (ce qui correspond au mode « collaboratif » de SYNC3, voir sous-section A.3.4). De ce point de vue, brat est véritablement un outil d'annotation *collaborative*, au sens de Wikipedia.

brat propose une interface simple, qui permet d'annoter des unités comme des relations. Une fonctionnalité majeure de brat est que chaque annotation est associée à une adresse URL unique, ce qui permet d'y faire référence dans la documentation.

En outre, il est possible de configurer brat pour enregistrer le temps passé par un annotateur sur un document et sur chaque action d'édition et de typage. En revanche, l'outil ne prévoit pas de calculer l'accord inter-annotateurs et ne distingue pas de rôle particulier pour le gestionnaire de la campagne.

Une autre limitation importante de l'outil est qu'il n'est pleinement supporté que par un nombre très limité de navigateurs (Chrome et Safari).

15. http://www.gnu.org/licenses/agpl-3.0.html
16. http://sourceforge.net/projects/ccash/
17. https://facwiki.cs.byu.edu/nlp/index.php/CCASH
18. http://www.opensource.org/licenses/mit-license.php
19. http://brat.nlplab.org/index.html

A.2 Outils spécifiques à une tâche TAL

A.2.1 Outils du LDC

Le *Linguistic Data Consortium* (LDC [20]) est un consortium d'entreprises, de laboratoires de recherches gouvernementaux et d'universités américains fondé en 1992 qui produit et distribue des ressources langagières. A ce titre, les développeurs du LDC ont créé un certain nombre d'outils d'aide à l'annotation pour la parole et le texte, autour de l'Annotation Graph Toolkit, ou `AGTK`, qui instancie un formalisme dérivé du modèle des graphes d'annotation (Bird et Liberman, 2001). `AGTK` est un ensemble de logiciels utiles au développement d'outils d'annotation, qui inclut des API permettant de manipuler des graphes d'annotation ainsi que des composants graphiques spécialisés.

Ces outils sont détaillés dans (Maeda et Strassel, 2004) et comprennent, outre un outil de transcription (`XTrans`) et un outil pour l'annotation d'entités nommées pour la campagne ACE (Automatic Content Extraction), des outils de comparaison et d'adjudication d'annotations concurrentes. Une évolution plus récente est l'Annotation Collection Toolkit ou `ACK` (Maeda *et al.*, 2006), une application Web écrite en PHP permettant de mettre en place rapidement des campagnes d'annotation simples.

Le LDC a donc fait le choix de la spécialisation des outils, tout en se fondant sur un « cœur » commun à ceux-ci. Ce choix, s'il permet une très grande adaptabilité aux besoins de chaque campagne, implique une bonne maîtrise d'`AGTK` et du développement supplémentaire pour pratiquement chaque campagne.

A.2.2 `Serengeti`

`Serengeti` (Stührenberg *et al.*, 2007) est un outil développé spécifiquement pour l'annotation de relations et de chaînes anaphoriques *via* le Web. Une version de démonstration est accessible sur le Web [21]. Pour l'instant, l'outil n'est utilisable qu'avec le navigateur Firefox et n'est pas librement disponible.

Une spécificité intéressante de `Serengeti` est un compte utilisateur particulier, *Consensus User*, qui permet de comparer des annotations réalisées par plusieurs annotateurs, et de les valider ou non. Cependant, cette fonctionnalité reste trop limitée pour être considérée comme une véritable interface de gestion (pas de calcul d'accord inter-annotateurs, pas de possibilité d'assigner des travaux d'annotation à tel ou tel annotateur, etc).

20. http://www.ldc.upenn.edu/
21. http://coli.lili.uni-bielefeld.de/serengeti/

A.2.3 EasyRef

EasyRef (De la Clergerie, 2008) est un outil pour l'annotation collaborative en syntaxe, mis au point dans le cadre du projet ANR PASSAGE (*Produire des annotations syntaxiques à grande échelle*), qui prévoit l'annotation de 100 millions de mots du français par 10 équipes, dont 500 000 annotés manuellement pour servir de référence.

A l'origine, EasyRef a été développé pour corriger les annotations manuelles produites pour servir de référence dans la campagne d'évaluation d'analyse syntaxique EASy [22]. L'outil a pris la forme d'une application Web, car plusieurs équipes dispersées devaient pouvoir l'utiliser. Il ne nécessite donc aucune installation côté client. Cette correction d'annotation (4 000 phrases) n'a cependant pas donné des résultats satisfaisants. Elle a été réalisée par différentes équipes, le guide d'annotation étant mis à jour au fur et à mesure. En outre, les annotateurs devaient remplir les champs par eux-mêmes, ce qui a été source de nombreuses erreurs. Le résultat n'est par conséquent pas homogène. Les participants à la campagne ont depuis soulevé un certain nombre d'erreurs, mais celles-ci n'ont pas été associées à des exemples, rendant leur correction difficile.

Profitant de cette expérience, EasyRef a été ensuite amélioré, le développement étant orienté selon quatre directions :

1. conserver une structure collaborative, afin de permettre à de nombreuses personnes de travailler et d'utiliser l'annotation, ainsi que pour faciliter la gestion du projet,
2. présenter une information riche et complexe de manière claire, en utilisant des zooms par exemple,
3. réduire le risque d'erreur grâce à des interfaces intelligentes et contraignantes,
4. conserver des traces des activités sur le système, grâce à des logs, un système de gestion de version, etc.

EasyRef [23] permet aujourd'hui de :

– visualiser des annotations, avec notamment un outil de recherche par expressions régulières et par critères administratifs (présence/absence d'un rapport de bug, par exemple) et la possibilité de cacher certaines informations,
– éditer les annotations de manière contrainte tout en conservant la trace précise de ces modifications : chaque action d'édition crée une nouvelle version de la phrase, avec un numéro de révision incrémenté,
– gérer des rapports de bugs reliés à une phrase, y compris des discussions sur un bug particulier.

Par ailleurs, EasyRef n'est pour l'instant pas open-source.

22. http://atoll.inria.fr/passage/eval2.fr.html
23. http://atoll.inria.fr/easyrefpub/login

A.2.4 `Phrase Detectives`

`Phrase Detectives` (Chamberlain *et al.*, 2008b) est une application Web créée pour l'annotation manuelle d'anaphores. L'originalité de cette application réside dans sa forme, puisqu'il s'agit d'un jeu de type *Game With A Purpose* (GWAP, jeu ayant un but).

Les textes proposés aux joueurs sont automatiquement pré-annotés suivant un processus détaillé dans (Kruschwitz *et al.*, 2009) : normalisation, tokenisation, analyse syntaxique (avec le Berkeley Parser) et identification des annotables[24]. Une phase de formation sur un corpus déjà annoté par des experts (*gold standard*) permet aux joueurs de se familiariser avec la tâche, de voir leurs erreurs et de gagner en compétence. Une fois atteint un niveau suffisant, ils peuvent commencer à véritablement « jouer », c'est-à-dire à annoter l'antécédent le plus proche d'un élément surligné dans le texte ou à valider des annotations réalisées par d'autres.

Les joueurs marquent des points si d'autres joueurs sont d'accord avec eux. Ils sont par ailleurs régulièrement testés par rapport au *gold standard*.

Différents moyens sont mis en place pour motiver les joueurs (Chamberlain *et al.*, 2009b), en particulier des classements et des prix accordés aux meilleurs. `Phrase Detectives` a ainsi permis d'annoter plus de 400 documents, soit plus de 162 000 tokens, avec une qualité d'annotation évaluée comme étant satisfaisante (Chamberlain *et al.*, 2012).

Ce type d'outil reste cependant limité à une tâche particulière (l'annotation d'anaphore) et n'est pas facilement adaptable à d'autres types d'annotations, car la trame du jeu (ou *gameplay*) est étroitement liée à l'annotation d'anaphore et tout autre type d'annotation nécessiterait une adaptation conséquente du jeu. En outre, `Phrase Detectives` n'est pas une application librement disponible.

A.3 Plate-formes d'annotation automatique

L'avantage de ces plate-formes d'annotation automatique est qu'elles proposent de nombreux outils de TAL, tels que des tokenizers, des analyseurs morphosyntaxiques (taggers), des analyseurs syntaxiques de surface (shallow parsers) ou encore des extracteurs d'entités nommées. Ces outils sont extrêmement utiles pour la pré-annotation automatique du corpus. Cependant, si ces plate-formes offrent la possibilité d'annoter manuellement un corpus, il ne s'agit que d'une fonctionnalité annexe, ajoutée parfois tardivement (`GATE`) ou en cours de développement (`UIMA`).

24. Voir section 3.3 pour une définition des annotables.

A.3.1 GATE

GATE (Cunningham *et al.*, 2002) (*General Architecture for Text Engineering*) est une plate-forme open source librement téléchargeable[25], développée par l'université de Sheffield, qui permet à la fois d'accéder à des outils et à des ressources de TAL existants et d'en développer de nouveaux. GATE propose pour cela une interface d'annotation manuelle (GATE Developer). L'outil est maintenu (la dernière version, 7.0, date de février 2012), documenté, il existe même des formations et des certifications GATE.

Jusqu'à la version 5 de l'outil, l'interface d'annotation manuelle était mal documentée, l'outil étant utilisé à l'origine pour appliquer des traitements automatiques et non pour faire de l'annotation manuelle. Les dernières versions sont mieux documentées de ce point de vue, mais l'utilisation de l'outil pour l'annotation manuelle reste marginale et limitée à de la correction d'annotations automatiques. Ainsi, un simple test montre qu'il est difficile de revenir sur une action (pas de fonctionnalité *undo* pour cela), par exemple.

En revanche, GATE propose toute une série d'options intéressantes, notamment la possibilité de rendre le document source éditable ou non (*Read Only*). Une autre option proposée par GATE est la possibilité de créer ses propres catégories pour l'annotation manuelle (*Restricted/Unrestricted annotation set*).

Si leur manipulation n'est pas toujours aisée, les fonctionnalités de l'outil sont variées. Il est par exemple possible d'annoter automatiquement toutes les occurrences d'une unité dans le texte (*Annotate all*). La visualisation des annotations réalisées, ainsi que la recherche dans le texte sont en outre faciles d'accès.

Enfin, GATE comprend un éditeur de chaînes de co-références, mais ne permet pas l'annotation de relations en tant que telles.

GATE accepte en entrée tous les formats balisés (XMLL, HTML, SGML), ainsi que le texte brut. Les annotations sont déportées en format TIPSTER (Grishman, 1998) et sont stockées soit sous forme de base de données relationnelle, soit sous forme de fichiers (XML ou sérialisation Java). L'outil offre un accès aux applications qu'il propose *via* une API. Il est également intégré dans UIMA.

GATE propose également un outil permettant de comparer deux ensembles d'annotations et génère des mesures de précision, rappel et F-mesure.

Récemment, l'équipe de GATE a proposé une extension de l'outil, Teamware (Bontcheva *et al.*, 2010), qui présente des fonctionnalités de gestion d'annotation et qui est par conséquent présentée dans la section A.4.

25. Sur le site http://gate.ac.uk/

A.3.2 EULIA

EULIA (Artola *et al.*, 2004) est un outil d'annotation collaboratif et plus largement de traitement linguistique sur le Web. Il prend en entrée et en sortie des documents XML TEI et manipule des structures de traits. Les annotations sont produites de manière déportée. EULIA propose également un moteur de recherche.

L'outil a la particularité de proposer non seulement un environnement d'annotation, mais également de traitement linguistique (tokenisation, analyse morpho-syntaxique, analyse syntaxique de surface).

Il ne semble en revanche pas proposer l'annotation de relations. Il ne nécessite aucune installation côté client. Par ailleurs, EULIA n'est pas open-source et ne semble pas maintenu.

A.3.3 UIMA

UIMA [26] (*Unstructured Information Management Applications*) est un projet de la fondation Apache, hérité d'IBM (Ferrucci et Lally, 2004) visant à fournir une architecture et des applications pour le traitement de l'information non structurée (et notamment, du texte).

Inclus dans l'environnement de développement Eclipse, Apache UIMA est long à installer, mais en contrepartie il propose de nombreux « annotateurs » (outils d'annotation automatique) pré-installés, dont un tokeniser simple (selon les espaces), un stemmer (Snowball), un analyseur morpho-syntaxique (HMM), un annotateur selon des expressions régulières, un extracteur d'entités nommées (OpenCalais), ainsi qu'un compilateur de dictionnaire permettant de reconnaître des éléments textuels prédéfinis. D'autres sont également disponibles mais ne sont pas installés par défaut, notamment pour le français [27].

Le CAS Editor permet quant à lui l'annotation manuelle de textes. Il reste pour l'instant limité et ne permet en particulier pas d'annoter des relations par le biais de l'interface. Il est en outre complexe à installer, même si depuis la version 2.3.0 d'UIMA, il se présente sous la forme d'un plugin Eclipse. Cela le rend peu utilisable par des utilisateurs en local et nécessite donc une mise en place sous forme client/serveur.

A notre connaissance, le CAS Editor n'est aujourd'hui pas encore utilisé pour de vrais projets d'annotation manuelle de corpus.

26. http://uima.apache.org/
27. Voir le blog de Nicolas Hernandez : http://enicolashernandez.blogspot.fr/p/uima.html

A.3.4 SYNC3

SYNC3 (Petasis, 2012) est un outil d'aide à l'annotation collaborative, librement disponible sous licence open source LGPL dans le cadre de la plate-forme de TAL Ellogon [28].

Cet outil a pour spécificité d'être un outil collaboratif tout en n'étant pas une application Web. Il s'agit d'une application installée en local, qui communique avec un serveur centralisé stockant les documents et les méta données dans une base de données. Ce parti pris original est justifié par son créateur par la plus grande souplesse offerte par une application « lourde » en termes d'interface utilisateur (possibilité de définir des raccourcis clavier, de personnaliser davantage l'outil, de lancer des traitements automatiques localement), ainsi que la possibilité, pour les annotateurs, de ne pas être connectés à Internet pour travailler.

SYNC3 propose les fonctionnalités disponibles dans Ellogon, dont le calcul de l'accord inter-annotateurs, mais il ne propose pas d'interface spécifique pour la gestion de l'annotation, ni d'utilisateur possédant des droits particuliers. Il enregistre également les annotations réalisées et en déduit des expressions régulières, révisables par l'utilisateur, permettant d'annoter automatiquement le reste du texte.

Enfin, l'outil distingue deux modes d'annotation : distribué et collaboratif. Dans le mode distribué, chaque annotateur annote séparément, alors que dans le mode collaboratif, les annotateurs partagent un même document. Dans ce dernier mode, les annotations sont toutes conservées. Ce mode « collaboratif » cumulatif est semblable à ce qui est proposé dans brat (voir sous-section A.1.11).

A.4 Outils de gestion d'annotations

A.4.1 Slate

Slate (Kaplan *et al.*, 2010, 2012), *Segment and Link-based Annotation Tool* Enhanced, est également un outil de gestion de projets d'annotation et d'aide à l'annotation, mais il est beaucoup plus ambitieux dans sa couverture que le précédent. Il s'agit, là encore, d'une application Web, dont le serveur est en Java et l'interface en Flash. Même si la licence qui lui est associée nous est inconnue, Slate semble maintenant être téléchargeable [29].

Il reconnaît deux types d'utilisateurs, les administrateurs et les annotateurs. Du point de vue administrateur, l'outil permet de créer des projets d'annotation, de gérer les

28. http://www.ellogon.org/
29. https://bitbucket.org/dainkaplan/slate/overview

annotateurs, de leur assigner des tâches d'annotation (et éventuellement de leur en enlever) et de suivre la progression du travail d'annotation, d'importer des données, de les exporter. L'administrateur peut également extraire les différences sur les parties du corpus annotées en parallèle et les fusionner pour créer une référence. L'outil offre enfin une fonctionnalité fondamentale de versionnage du projet d'annotation, au même titre qu'un projet de développement standard. Dans `Slate`, toutes les instances d'annotations sont identifiées par le numéro de version du projet, et donc du jeu d'étiquettes, au moment de l'annotation.

Il est intéressant de noter que dans `Slate`, l'administrateur ne peut pas annoter. Les annotateurs, quant à eux, ne peuvent pas voir le travail des autres annotateurs. Les deux rôles sont donc totalement indépendants l'un de l'autre.

Dans `Slate`, une annotation est une étiquette posée soit sur un empan de texte, soit sur une relation entre empans de texte (*links*). Ces relations peuvent être orientées ou non. `Slate` permet en outre de définir des contraintes sur l'annotation des segments et des relations. L'outil permet l'ajout de nouvelles couches d'annotation, avec de nouveaux jeux d'étiquettes, référençant les couches inférieures. Les annotations créées sont bien entendu déportées. Il est largement personnalisable par l'annotateur, notamment en ce qui concerne les couleurs utilisées. Une visualisation intégrée, dans un panneau réservé, de la documentation externe est également prévue. Enfin, `Slate` est multilingue et permet l'import et l'export des données annotées.

A.4.2 `Djangology`

`Djangology` (Apostolova *et al.*, 2010) est un outil de gestion de projets d'annotation et d'aide à l'annotation. C'est une application Web écrite en Python, utilisant le framework Django [30], Ajax et reposant sur une base de données quelconque. L'outil est conçu pour être rapide et facile à installer et permettre l'annotation distribuée et collaborative de documents textuels. Il est disponible librement, en open source [31], mais semble peu maintenu (la dernière modification date de 2010) et est très peu documenté.

Du point de vue de la gestion du projet d'annotation, cet outil offre une interface d'administration permettant de gérer les documents et les utilisateurs et de définir le modèle de données pour l'annotation. Il permet également d'effectuer des calculs d'accords inter-annotateurs et propose une interface de comparaison des annotations parallèles. Il permet enfin de calculer la précision d'une annotation automatique par rapport à une référence. Il n'inclut cependant pas de moteur de pré-annotation automatique.

30. https://www.djangoproject.com/
31. http://sourceforge.net/projects/djangology/

En tant qu'outil d'annotation, `Djangology` offre des fonctionnalités d'annotation de base, notamment la propagation automatique d'annotations déjà réalisées (équivalent de l'*Annotate all* de `GATE`). Grâce au couplage entre les fonctionnalités de gestion et d'annotation, l'annotateur peut prévenir le gestionnaire qu'il a fini d'annoter son document. Enfin, l'outil ne semble pas permettre l'annotation de relations.

A.4.3 `GATE Teamware`

`GATE Teamware` (Bontcheva *et al.*, 2010) est une extension de `GATE` (voir sous-section A.3.1) prévue pour organiser et réaliser de l'annotation dite « collaborative » en ligne. Il s'agit donc d'un outil d'annotation et de gestion de campagnes. Comme `GATE`, `Teamware` est open-source (sous licence GPL Affero [32]) et disponible sous SourceForge [33]. L'architecture de l'outil est décrite dans (Bontcheva *et al.*, 2010) et distingue trois couches de traitement : des services Web SOAP pour le stockage des données, des interfaces utilisateurs Web et une couche d'exécution permettant de gérer les flux des projets d'annotation.

A l'instar de `Slate`, `Teamware` a la particularité de rendre explicite la division des rôles d'une campagne d'annotation. Cependant, là où `Slate` prévoit deux types acteurs, le gestionnaire de la campagne et les annotateurs, `Teamware` en prévoit trois : les gestionnaires de la campagne, les éditeurs ou curateurs (annotateurs experts) et les annotateurs. Les curateurs mesurent les accords inter-annotateurs, font l'adjudication, sont responsables de la formation des annotateurs, et répondent à leurs questions. Les gestionnaires d'annotation, quant à eux, définissent les nouveaux projets d'annotation, en gèrent l'avancement, participent aux choix méthodologiques (organisation du calcul de l'accord, choix d'une pré-annotation éventuelle, etc.). Enfin, les annotateurs sont, dans ce cadre, des « non-spécialistes », qu'il convient de former en début de campagne. Un moyen de communication entre annotateurs et éditeurs par message instantané est également prévu.

`Teamware` propose les services d'annotation automatique présents dans `GATE` et en reprend largement l'interface d'annotation. L'outil permet également d'annoter en utilisant une ontologie, comme `Knowtator` (voir sous-section A.1.6). Une interface d'adjudication est fournie aux éditeurs, qui leur permet de sélectionner les annotations qu'ils jugent bonnes parmi celles proposées par les annotateurs. Enfin, l'interface du gestionnaire de la campagne lui permet de créer des corpus, de définir le schéma d'annotation, de configurer les éventuelles pré-annotations et de les exécuter sur les corpus. Cette interface lui propose également un suivi de la campagne : avancement (nombre de documents annotés, restant à annoter, en cours d'annotation) et statistiques sur les annotateurs (temps par document, temps de travail effectif, accord inter-annotateurs,

32. http://www.gnu.org/licenses/agpl-3.0.html
33. https://gate.svn.sourceforge.net/svnroot/gate/teamware/trunk/

etc). Un rôle d'administrateur et de super-administrateur sont également prévus pour gérer l'application et les droits des différents acteurs.

Cet outil est très complet, mais présente apparemment les mêmes problèmes d'interface que `GATE`. Certaines difficultés spécifiques sont également apparues lors d'expériences d'annotation, notamment de vitesse et donc de trafic réseau. Enfin, `GATE Teamware` ne peut pas être utilisé pour annoter sur Firefox.

A.5 Autres outils

Il existe d'autres outils de gestion et d'aide à l'annotation, mais qui ne sont pas utilisables, pour différentes raisons.

Ainsi, `WordFreak` (Morton et LaCivita, 2003) est un outil d'aide à l'annotation open source, disponible sous Mozilla Public License[34] et écrit en Java, qui a inspiré certains outils d'aide à l'annotation actuels. Il n'est cependant plus maintenu depuis longtemps.

De même, `Annotatornia` (Przepiórkowski et Murzynowski, 2009), s'il est open source (licence GPL) et disponible[35], n'est pas du tout maintenu et son interface n'existe qu'en polonais. C'est à notre connaissance le seul outil d'aide à l'annotation écrit en Ruby.

Enfin, `CSAF` (Kim et Wang, 2012) apporte d'intéressantes fonctionnalités de gestion d'une « communauté » d'annotateurs, mais n'en est encore qu'au stade du prototype. De même, `ANALEC` (Landragin *et al.*, 2012), s'il est très souple d'utilisation (le modèle de données est modifiable à tout moment), reste encore trop peu documenté. Par ailleurs, ses fonctionnalités spécifiques de visualisation (par exemple, une chaine de co-référence peut être directement représentée par sa longueur) et d'analyse géométrique (interface sur laquelle les unités partageant des contextes similaires apparaissent regroupés) sont à usage encore trop restreint pour que nous le présentions en détails[36].

34. http://wordfreak.sourceforge.net
35. http://nlp.ipipan.waw.pl/Anotatornia/
36. http://www.lattice.cnrs.fr/Telecharger-Analec

Table des matières

Table des figures

Liste des tableaux

Liste des définitions

Résumé

L'annotation manuelle de corpus est devenue un enjeu fondamental pour le Traitement Automatique des Langues (TAL). En effet, les corpus annotés sont utilisés aussi bien pour créer que pour évaluer des outils de TAL. Or, le processus d'annotation manuelle est encore mal connu et les outils proposés pour supporter ce processus souvent mal utilisés, ce qui ne permet pas de garantir le niveau de qualité de ces annotations. Nous proposons dans cette thèse une vision unifiée de l'annotation manuelle de corpus pour le TAL. Ce travail est le fruit de diverses expériences de gestion et de participation à des campagnes d'annotation, mais également de collaborations avec différents chercheur(e)s. Nous proposons dans un premier temps une méthodologie globale pour la gestion de campagnes d'annotation manuelle de corpus qui repose sur deux piliers majeurs : une organisation des campagnes d'annotation qui met l'évaluation au cœur du processus et une grille d'analyse des dimensions de complexité d'une campagne d'annotation. Un second volet de notre travail a concerné les outils du gestionnaire de campagne. Nous avons pu évaluer l'influence exacte de la pré-annotation automatique sur la qualité et la rapidité de correction humaine, grâce à une série d'expériences menée sur l'annotation morpho-syntaxique de l'anglais. Nous avons également apporté des solutions pratiques concernant l'évaluation de l'annotation manuelle, en donnant au gestionnaire les moyens de sélectionner les mesures les plus appropriées. Enfin, nous avons mis au jour les processus en œuvre et les outils nécessaires pour une campagne d'annotation et instancié ainsi la méthodologie que nous avons décrite.

Mots clés : annotation manuelle, méthodologie, évaluation, accords inter-annotateurs

Abstract

Manual corpus annotation has become a key issue for Natural Langage Processing (NLP), as manually annotated corpora are used both to create and to evaluate NLP tools. However, the process of manual annotation remains underdescribed and the tools used to support it are often misused. This situation prevents the campaign manager from evaluating and guarantying the quality of the annotation. We propose in this work a unified vision of manual corpus annotation for NLP. It results from our experience of annotation campaigns, either as a manager or as a participant, as well as from collaborations with other researchers. We first propose a global methodology for managing manual corpus annotation campaigns, that relies on two pillars: an organization for annotation campaigns that puts evaluation at the heart of the process and an innovative grid for the analysis of the complexity dimensions of an annotation campaign. A second part of our work concerns the tools of the campaign manager. We evaluated the precise influence of automatic pre-annotation on the quality and speed of the correction by humans, through a series of experiments on part-of-speech tagging for English. Furthermore, we propose practical solutions for the evaluation of manual annotations, that provide the campaign manager with the means to select the most appropriate measures. Finally, we brought to light the processes and tools involved in an annotation campaign and we instantiated the methodology that we described.

Keywords: manual annotation, methodology, evaluation, inter-annotator agreements