Classifications flexionnelles. Étude quantitative des structures de paradigmes

Sacha Beniamine

Thèse de doctorat
de l'Université Sorbonne Paris Cité
Préparée à l'Université Paris Diderot
Ecole doctorale Sciences du langage (ED 132)
Laboratoire de Linguistique Formelle / Labex EFL

Classifications flexionnelles

Étude quantitative des structures de paradigmes.

par Sacha Beniamine
Thèse de doctorat de Linguistique Théorique et Descriptive
Dirigée par Olivier Bonami

Présentée et soutenue publiquement à Paris le 6 juillet 2018

Rapporteur :	Dunstan Brown,	Professeur,	U. of York
Rapporteur et président du jury :	Nabil Hathout,	Directeur de recherche,	CNRS - CLLE ERSS (Toulouse)
	Delphine Tribout	Maître de Conférences,	U. de Lille
	Ana R. Luís	Maître de Conférences,	U. de Coimbra
	Berthold Crysmann	Chargé de recherche,	CNRS - U. Paris Diderot
Directeur :	Olivier Bonami,	Professeur,	U. Paris Diderot

Titre : Classifications flexionnelles. Étude quantitative des structures de paradigmes.

Résumé : Dans les systèmes flexionnels, il arrive que des propriétés morphosyntaxiques soient exprimées différemment d'un lexème à l'autre. Pour décrire ces systèmes, il est usuel d'énumérer un petit nombre de classes flexionnelles formant une partition des lexèmes. Les systèmes flexionnels suivent cependant une organisation beaucoup plus complexe, et en raison d'un flou méthodologique, les travaux sont souvent en désaccord sur l'inventaire exact des classes.

Cette thèse se place dans la perspective Mot et Paradigme et élabore des outils computationnels permettant d'observer précisément la structure de similarité des systèmes de classes flexionnelles en se fondant sur des lexiques de formes fléchies. Nous étudions les verbes de l'arabe, de l'anglais, du chatino de Zenzontepec, du chatino de Yaitepec, du français, du navajo, du portugais européen, ainsi que les noms du Russe.

Dans une première partie, nous proposons de décrire le comportement flexionnel des lexèmes au moyen des alternances entre leur formes. Nous présentons un algorithme pour inférer automatiquement des patrons d'alternances reliant deux formes de surface d'un même lexème. Nous employons ces patrons pour quantifier le problème de remplissage des cases de paradigme. Dans une seconde partie, nous nous interrogeons sur la structure de similarité des systèmes flexionnels. Nous commençons par classer les lexèmes en microclasses, fondées sur l'identité du comportement flexionnel. Celles-ci sont nombreuses, et parfois très similaires entre elles. Nous proposons ensuite un algorithme fondé sur la longueur de description permettant de regrouper les microclasses en un petit nombre de macroclasses correspondant à la notion traditionnelle de classe flexionnelle. Enfin, nous montrons que le modèle le plus fidèle pour décrire les similarités au sein de ces systèmes est un treillis dont chaque nœud constitue une classe flexionnelle. Pour déduire cette hiérarchie à héritage multiple des patrons d'alternances, nous employons l'analyse de concepts formels.

Mots clés : Linguistique computationnelle, morphologie flexionnelle, Traitement Automatique des Langues Naturelles (TALN), approche mot et paradigme, classes flexionnelles, patrons d'alternances, PCFP, longueur de description minimale, analyse de concepts formels, hiérarchie flexionnelle.

Title : Inflectional classifications. A quantitative study of paradigm structures.

Abstract : In some inflectional systems, the same morphosyntactic properties can be expressed differently across lexemes. These systems are usually described through the enumeration of a small number of inflection classes partitioning the inventory of lexemes. However, the actual structure of inflection class systems is much more complex, and methodological vagueness leads to contradictory accounts regarding inventories of inflection classes.

This dissertation adopts the Word and Paradigm approach and elaborates computational tools to investigate precisely the similarity structure of inflection class systems based on inflectional lexicon. We study Arabic, Yaitepec Chatino, Zenzontepec Chatino, English, French, Navajo and European Portuguese verbs as well as Russian nouns.

The first part defines the inflectional behavior of lexemes through the set of all surface alternations between their forms. We describe an algorithm to infer automatically alternation patterns between any two forms of a lexeme. We use alternation patterns to quantify the Paradigm Cell Filling Problem (PCFP). The second part investigates the similarity structure of inflectional systems. We start by classifying lexemes into microclasses, based on identity of inflectional behavior. These classes are numerous, and sometimes very similar. We then describe an algorithm based on minimal description length to gather microclasses into macroclasses which conform to the traditional notion of inflection class. Finally, we show that the most faithful model to describe similarities in inflectional systems is a lattice in which each node is an inflection class. To deduce this multiple inheritance hierarchy from alternation patterns, we use Formal Concept Analysis.

Keywords : Computational linguistics, inflectional morphology, Natural Language Processing, Word and Paradigm, inflection classes, alternation patterns, PCFP, minimal description length, gormal concept analysis, inflectional hierarchy.

Remerciements

Je souhaite remercier Olivier Bonami pour sa direction précieuse. Cette thèse a été mûrie au fil de discussions quasi hebdomadaires qui m'ont captivé dès la première et m'ont guidé tout au long du travail. Je suis reconnaissant pour sa disponibilité exceptionnelle, sa générosité intellectuelle, son exigence bienveillante et son soutien en toutes choses. Je n'aurais pas pu espérer meilleure direction.

Je remercie mes rapporteurs, Dunstan Brown et Nabil Hathout, ainsi que Delphine Tribout, Ana R. Luís et Berthold Crysmann, qui ont accepté de participer au jury.

La préparation de cette thèse a été financée par un contrat doctoral et diverses autres aides de l'opération Morph 1 de l'axe 2 du Laboratoire d'excellence « Fondements Empiriques de la Linguistique » (Labex EFL).

Cette thèse doit énormément à tous·tes les collaboratrices·eurs avec qui nous avons élaboré, adapté ou phonémisé les lexiques que j'ai étudiés dans cette thèse, et qui m'ont fait bénéficier de leur expertise. Je remercie Joyce McDonough pour les nombreuses réunions de travail autour d'un exemplaire du dictionnaire Young & Morgan, qui ont conduit à l'élaboration du lexique du navajo. Je remercie également son étudiant Benjamin Goehring, qui a saisi manuellement l'ensemble des tables de conjugaison des bases$_1$ du dictionnaire Young & Morgan. Je remercie Dunstan Brown et Enrique Palancar d'avoir mis à ma disposition respectivement les lexiques du russe et du chatino, ainsi que pour leurs explications, leur patience et leur minutie lors de la constitution des données phonétisées. Enfin, je remercie Kenza Ould Hamouda qui a validé manuellement les transcriptions phonémiques du lexique des verbes de l'arabe.

Je remercie Farrell Ackerman pour m'avoir accueilli chaleureusement à San Diego au printemps 2016. Je suis reconnaissant à Farrell Ackerman, ses étudiants et à Rob Malouf pour les

séminaires de lecture du mercredi et pour avoir organisé à cette occasion des séminaires de morphologie passionnants le dimanche après midi.

Je souhaite remercier Gilles Boyé pour les discussions éclairantes que nous avons eu lors de conférences, ainsi que Pascal Amsili et Benoît Crabbé pour leurs enseignements captivants et leur grande disponibilité. Le texte de cette thèse est écrit dans la police d'écriture Linux Libertine B, une adaptation de Linux Libertine par Berthold Crysmann, auquel je suis également reconnaissant pour ses précieux conseils en matière de LaTeX.

Merci à tous·tes mes collègues doctorant·e·s et post-doctorant·e·s des bureaux du 6e et du 5e, qui ont été mes compagnons de route au cours de ces quatre années. En particulier, je remercie pour leur présence inestimable Maximin et Olga qui m'ont conseillé et relu un nombre de fois incalculable, Marianne, Céline et Ingrid pour leur amitié et leur solidarité sportives ; ainsi qu'Aixiu, Charlotte, Chloé, Corentin, Gabriel, Hector, Jiaying, Kristina, Marianne, Marion, Patty, Saida, Sandro, Shrita, Timothée, Valérie, Vincent, Wei, Yiqin, et tous·tes les autres.

Enfin, je suis profondément reconnaissant à Max, Laure, Marc, et tous les Davids pour m'avoir hébergé, précédé, épaulé, et encouragé durant la rédaction de cette thèse.

Table des matières

Liste des tableaux

Table des figures

Introduction

La description d'un système verbal ou nominal commence souvent par l'énumération d'un petit nombre de classes flexionnelles, dites conjugaisons ou déclinaisons. Ces classes offrent une image succincte des comportements flexionnels parfois très variés des lexèmes du système. Par exemple, Brown et Hippisley (2012) distinguent quatre classes flexionnelles (déclinaisons) des noms du russe, que nous illustrons pour une partie du paradigme dans le tableau 1.

	I	II	III	IV
NOM.SG	zakon	karta	rukop'is'	boloto
DAT.SG	zakonu	karte	rukop'is'i	bolotu
NOM.PL	zakoni	karti	rukop'is'i	bolota
DAT.PL	zakonam	kartam	rukop'is'am	bolotam

TABLEAU 1 – Principales déclinaisons du russe selon Brown et Hippisley (2012, p. 48).

Les systèmes flexionnels suivent cependant une organisation beaucoup plus complexe que l'image qu'en donnent les grammaires pédagogiques. Dans le tableau 1, les cases de paradigme sélectionnées montrent des points de similarités complexes entre les classes. Dans les classes I et III, la forme de nominatif singulier ne présente pas d'affixe. Dans les quatre classes, le datif pluriel se termine en /-am/. Dans les classes I et III, le nominatif peut se former sur le datif pluriel en retirant /-am/, mais dans la classe II, le /-a-/ demeure au nominatif, et dans la classe IV, le nominatif se termine en /-o/. Au datif singulier, les classes I et IV se terminent en /-u/, tandis que les classes II et III se terminent respectivement en /-e/ et /-i/. Au nominatif pluriel, ce sont les classes I, II et III qui se terminent identiquement en /-i/, tandis que la classe IV est marquée par

un /-a/ final. Aucune des quatre cases du paradigme présentées ne suffit à distinguer les classes, et chaque point de similarité entre elles pourrait nous inciter à réunir les classes concernées. De fait, on trouve souvent de nombreuses descriptions concurrentes d'un même système. Corbett (1982) commente ainsi la situation en russe :

> Le lecteur qui n'est pas familier avec cette littérature s'attendra raisonnablement à une description claire des paradigmes du russe. La tradition en distingue trois, certains auteurs en reconnaissent quatre, et plus récemment il a été suggéré que seuls deux paradigmes sont nécessaires. [...] Il n'existe pas de procédure établie pour déterminer le nombre de paradigmes nécessaires à partir d'un ensemble de données ; il y a quelque vagues intuitions et la tradition joue un rôle important [1].

Cette thèse propose de répondre à ce problème, qui est loin de se limiter aux noms du russe, en élaborant des outils computationnels [2] et une méthodologie systématique pour observer les systèmes de classes flexionnelles en se fondant strictement sur les données. Au cours de cette thèse, nous nous demanderons comment observer la structure de similarité entre les comportements flexionnels des lexèmes, comment classer les lexèmes en fonction de ces comportements, et quelles structures émergent empiriquement de ces systèmes.

Au centre de la question des classes flexionnelles se situe celle, plus épineuse encore, de l'EXPONENCE : comment identifier la partie d'une forme d'un mot qui manifeste des valeurs flexionnelles ? Nous nous plaçons donc dans le cadre du courant MOT ET PARADIGME (Hockett 1954, 1967, 1987 ; Matthews 1965, 1972 ; Blevins 2006, 2016) qui propose que les mots entiers manifestent de telles valeurs et propose d'observer le comportement flexionnel des mots à travers les relations qui les lient dans les paradigmes. Cette approche n'accorde pas de place centrale à la notion de morphème, et abandonne la tentation d'une segmentation catégorique en exposants pour s'appuyer plutôt sur l'observation des contrastes flexionnels au sein des

1. [En anglais dans le texte] « *The reader not familiar with the literature will quite reasonably expect a straightforward account of the paradigms in Russian. Tradition answers three, some writers claim four, and more recently it has been suggested that only two paradigms are required. [...] There is no set procedure for determining the number of paradigms required given a set of data; there are certain vague intuitions and tradition plays a large part* ».

2. Les outils développés dans le cadre de cette thèse sont publiés sous le nom Qumin (QUantitative Modelling of INflection) et distribués librement à l'adresse : https://github.com/XachaB/Qumin

Langue	Taille en lexèmes	Taille en cases de paradigme	Source
Français	5249	51	Bonami, Caron et Plancq (2014)
Anglais	6064	8	Baayen, Piepenbrock et Gulikers (1995)
Chatino de Zenzontepec	392	4	Feist et Palancar (2015)
Chatino de Yaitepec	324	12	Feist et Palancar (2015)
Portuguais européen	1996	69	Veiga, Candeias et Perdigão (2013)
Arabe	1018	109	Kirov et al. (2016)
Russe	1539	13	Brown (1998)
Navajo	2157	70	Young et Morgan (1987)

Tableau 2 – Lexiques flexionnels étudiés.

paradigmes. Nous appellerons ces contrastes PATRONS D'ALTERNANCE (Bonami et Luís 2014). Cette perspective est nécessaire afin d'avoir les moyens de travailler sur des données qui ne sont pas pré-analysées (segmentées en radicaux et affixes), et donc de nous permettre d'élaborer des outils fournissant des comparaisons qui ne soient pas typologiquement biaisées.

Notre approche est inductive, computationnelle, quantitative et sensible à la diversité des langues. Elle vise à constituer des outils qui puissent se prêter à des études typologiques.

Nous nous fondons sur huit lexiques flexionnels dont les sources sont récapitulées dans le tableau 2. Le lexique du russe concerne la flexion nominale, tous les autres concernent la flexion verbale. Chaque lexique documente un ensemble de 324 (chatino de Yaitepec) à 6064 (anglais) lexèmes, chacun fléchi dans 4 (chatino de Zenzontepec) à 109 (arabe) formes fournies en notation phonémique. Chaque lexique est accompagné d'une *clé* phonémique qui décompose chaque caractère utilisé en traits phonologiques distinctifs ; l'inventaire de traits utilisé est dans l'esprit de Chomsky et Halle (1968). La recherche, l'adaptation et la phonémisation des lexiques ont occupé une partie importante du temps de préparation de cette thèse. Nous documentons le détail de ce travail dans l'annexe A.

Les lexiques choisis ne constituent pas un échantillon typologiquement représentatif, mais

ils manifestent des types d'exponence variés. En particulier, la flexion verbale de l'arabe présente de la transfixation, les systèmes du chatino de la flexion tonale, en russe et en portugais on rencontre des alternances accentuelles qui ont valeur d'exposant, et les verbes du navajo présentent de la flexion polysynthétique.

Les méthodes de classification élaborées dans cette thèse sont entièrement implémentées. Nous nous appuyons sur la théorie de l'information et en particulier des mesures d'entropie et de longueur de description, sur des algorithme de classification (*clustering*), ainsi que sur le formalisme de mathématiques appliquées nommé analyse formelle de concepts.

Le chapitre 1 introduit et problématise les notions essentielles à la compréhension de cette thèse, en particulier celles de paradigme, d'exposant, de patrons d'alternance, de classe flexionnelle et de canonicité. Nous argumentons en faveur d'une approche Mot et Paradigme, et proposons d'identifier le comportement flexionnel d'un lexème par l'ensemble des alternances manifestées entre ses formes. Nous montrons que le terme de « classe flexionnelle » désigne divers types de classification des lexèmes. D'une part, les auteurs cherchant à fournir une définition précise s'appuient généralement sur l'identité des comportements flexionnels. D'autre part, la tradition grammaticale fonde les classes flexionnelles sur la similarité entre lexèmes. En pratique, la plupart des descriptions de linguistes emploient cette dernière définition, car la première mène à un nombre de classes trop important pour donner une image générale des systèmes. Suivant Dressler, Mayerthaler et al. (1987), nous nommons respectivement MICROCLASSES et MACROCLASSES les classes fondées sur l'identité ou la similarité des comportements flexionnels. L'essentiel du problème des classes flexionnelles se situe dans cette distinction. D'une part les microclasses sont aisées à identifier si l'on dispose d'une caractérisation précise du comportement flexionnel des lexèmes, mais elles sont trop petites et nombreuses pour fournir des généralisations intuitives. D'autre part, les macroclasses sont conçues pour offrir de telles généralisation, mais leur identification pose problème, car la similarité est une mesure continue. Comment décider quel degré de similarité fonde l'appartenance à une macroclasse ? Enfin, nous discutons des propriétés d'un système de classes flexionnelles canoniques (Corbett 2009) dans lequel les deux types de classes coïncident.

La suite de cette thèse s'organise en deux parties. La première partie comporte deux cha-

pitres et propose une étude des alternances flexionnelles à travers l'inférence automatique de patrons d'alternances et l'étude de leur prédictibilité. La seconde partie comporte trois chapitres et étudie différents modèles de classification flexionnelles.

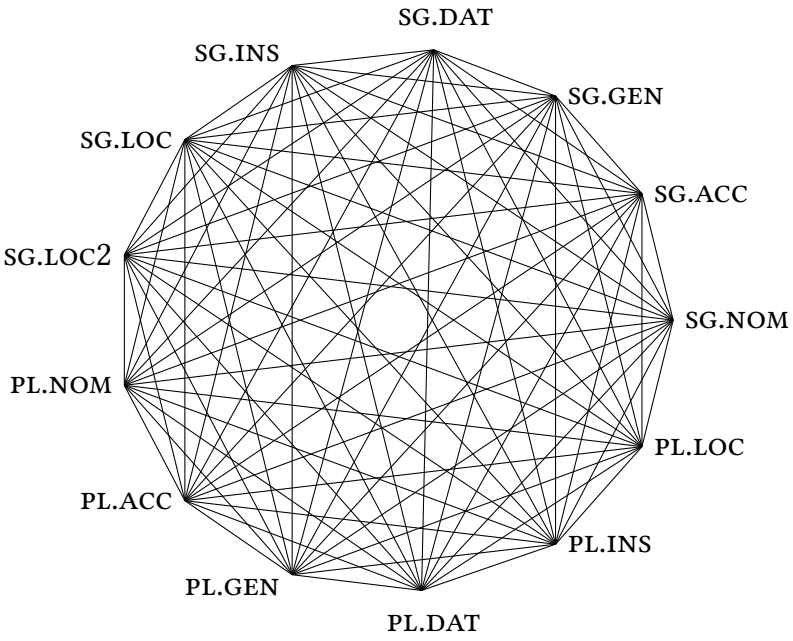


FIGURE 1 – Contrastes morphologiques examinés pour les noms du russe

Les données des lexiques sur lesquels nous travaillons se présentent sous la forme de tables où figurent des formes de surface non analysées. Le second chapitre de cette thèse propose un algorithme permettant d'identifier et de caractériser automatiquement les contrastes morphologiques au sein du paradigme de chaque lexème. La figure 1 illustre par des arcs l'ensemble des contrastes morpholgiques examinés pour les noms du russe. Ces contrastes sont formulés sous la forme de patrons d'alternance bidirectionnels, qui classent l'ensemble des lexèmes pour chaque paire de cases de paradigme. L'algorithme permettant d'inférer les patrons à partir des formes brutes procède sans connaissance préalable du type d'exponence à l'œuvre dans les données qu'il analyse. Il est conçu pour une application à des langues typologiquement variées, et sensible aux propriétés phonologiques des formes. Il s'inspire fortement du *minimum generalization learner* d'Albright et Hayes (2002), dont il étend le principe d'une part pour pouvoir

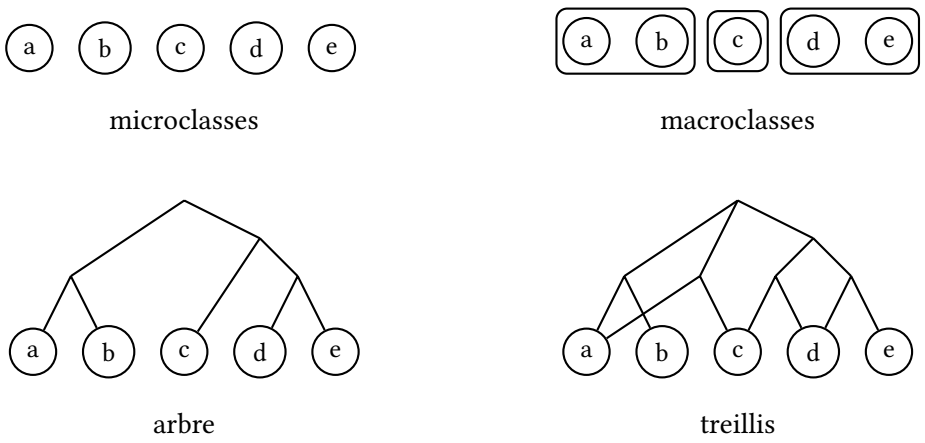


FIGURE 2 – Modèles de classifications.

calculer efficacement des patrons sur un grand nombre de paires de cases, et d'autre part pour pouvoir calculer automatiquement un alignement optimal entre les formes à contraster. Ce chapitre est une version étendue d'un article publié dans les actes de la conférence TALN (Beniamine 2017).

Le second chapitre se penche sur l'évaluation de la complexité des systèmes flexionnels, et en particulier sur le problème de remplissage des cases de paradigmes, ou PCFP (Ackerman, Blevins et Malouf 2009) : pour un locuteur qui connaît une forme de paradigme, quelle est la difficulté à produire une forme inconnue du même lexème ? Ackerman, Blevins et Malouf (2009) évaluent l'incertitude d'un locuteur face à ce problème au moyen d'une mesure d'entropie conditionnelle calculée sur des distributions d'affixes. Nous montrons, suivant Bonami et Boyé (2014), qu'une mesure similaire fondée sur les patrons d'alternances constitue une meilleure mesure du PCFP, et que les locuteurs ont souvent à leur disposition plus d'une forme connue sur laquelle fonder leur prédiction. Nous proposons donc également une mesure permettant d'évaluer la difficulté d'une telle entropie jointe. Nous présentons les résultats de l'évaluation du PCFP pour l'ensemble des huit systèmes considérés dans cette thèse. Nos conclusions confirment et étendent celles de Ackerman, Blevins et Malouf (2009), Ackerman et Malouf (2013), Bonami et Boyé (2014) et Bonami et Luís (2014). Ce chapitre se fonde sur un article publié dans la revue *Word Structure* (Bonami et Beniamine 2016), dans lequel nous appliquions cette méthodologie exclusivement aux verbes du français et du portugais.

Dans la seconde partie de cette thèse, nous employons les patrons d'alternance comme des traits permettant de caractériser le comportement flexionnel de chaque lexème et de fonder des classifications flexionnelles. La question qui sous-tend l'ensemble de cette partie est celle du modèle de classification : quel type de classification est le mieux à même de rendre compte de la structure des systèmes flexionnels ? La figure 2 présente les quatre modèles envisagés dans cette thèse : les microclasses (énumérées a, b, c, d et e dans la figure) sont des groupes de lexèmes qui présentent exactement les mêmes patrons d'alternance. Les macroclasses sont des groupes de microclasses qui forment une partition des lexèmes. Elles sont peu nombreuses et correspondent aux classes flexionnelles généralement décrites par les linguistes. Les systèmes flexionnels sont parfois décrits comme des arbres dont les microclasses constituent les feuilles. Nous proposons que les treillis, dans lesquels une classe peut avoir plusieurs parents, sont des structures hiérarchiques plus fidèles aux réseaux de similarités au sein des classes flexionnelles.

Le chapitre 4 s'attache à décrire les systèmes de microclasses. Il s'intéresse aux réseaux de similarité qui les connectent. Les microclasses se déduisent immédiatement de l'observation des inventaires de patrons d'alternance instanciés par chaque lexème. Les classes obtenues sont, à travers tous les systèmes étudiés, nombreuses et, sauf pour quelques classes majoritaires, plutôt petites. Elles ne correspondent pas du tout à la notion traditionnelle et intuitive dont les descriptions linguistiques font généralement usage. Nous discutons également de la possibilité d'organiser automatiquement les microclasses en hiérarchies arborescentes. Celles-ci permettent de mettre en relief certaines similarités saillantes au sein des réseaux de microclasses, mais ne fournissent qu'une image très partielle de leur structure. Par ailleurs, ils ne se prêtent pas aisément à la comparaison ou à l'évaluation des systèmes de classes proposées dans la littérature.

Le chapitre 5 propose un algorithme fondé sur la longueur de description minimale (Rissanen 1978) et permettant de produire une classification des microclasses en macroclasses. Cet algorithme se fonde sur une description formelle des systèmes de macroclasses. Les macroclasses ont l'avantage de fournir une image des systèmes plus intuitives que les microclasses, car elles sont plus grandes et moins nombreuses. L'algorithme que nous avons élaboré recherche un équilibre entre la perte d'information liée à cette simplification et le gain résultant de la réunion

de microclasses similaires. Les macroclasses étant de la même nature que les classes décrites dans la plupart des grammaires descriptives, nous pouvons comparer les résultats de nos analyses aux inventaires de classes connus pour les systèmes flexionnels étudiés. Nous montrons que l'algorithme proposé fournit des macroclasses généralement plus informatives que celles qui sont proposées dans la littérature. Nous concluons cependant que les macroclasses sont en général un modèle trop pauvre pour refléter correctement la structure des systèmes flexionnels. Ce chapitre est une version étendue d'un article publié dans le *Journal of Linguistic Modelling* (Beniamine, Bonami et Sagot 2017).

Le chapitre 6 propose d'appeler classe flexionnelle tout ensemble maximal de lexèmes pouvant être associé à un ensemble de patrons d'alternances. Ces classes peuvent être ordonnées entre elles par des relations d'inclusion, et elles forment ainsi de très grandes et très précises hiérarchies. Une caractéristique importante de ces hiérarchies est que ce ne sont pas des arbres, mais des treillis au sens mathématique : une même classe peut avoir plusieurs superclasses. Cette caractéristique est cruciale pour pouvoir rendre compte des phénomènes d'hétéroclise, où des lexèmes exhibent un comportement intermédiaire entre deux classes préalablement identifiées. Un des résultats importants de ce chapitre est de montrer, à travers une exploration systématique, que les phénomènes d'hétéroclise sont omniprésents dans les systèmes flexionnels. Nous les construisons automatiquement à partir des patrons d'alternance, en employant un formalisme de mathématiques appliquées : l'analyse formelle de concepts. Nous montrons que les hiérarchies résultantes sont trop complexes pour être construites à la main, et s'éloignent beaucoup d'un idéal canonique de classes flexionnelles au sens de Corbett (2009). Cependant, elles sont beaucoup plus proches de cet idéal que du maximum théorique de complexité possible. En outre, ces hiérarchies se prêtent à des évaluations quantitatives, mais sont difficilement exploitables pour des descriptions qualitatives fines. Nous fournissons donc une façon de sélectionner des classes dans la hiérarchie pour en fournir une vue simplifiée explorable manuellement et donnons quelques exemples de l'utilité d'une telle exploration manuelle pour étudier les propriétés d'un système et évaluer les classifications traditionnelles.

Chapitre 1

Les notions de classe flexionnelle

Qu'est-ce qu'une classe flexionnelle ? Les descriptions de systèmes flexionnels prennent souvent pour acquis un inventaire spécifique de classes. Pourtant, comme le montre Corbett (1982) pour les noms du russe, il existe souvent dans la littérature différentes analyses en classes concurrentes pour un même système. Plus encore, d'un auteur à l'autre, les analyses se fondent implicitement ou explicitement sur des définitions variées de la notion de classe flexionnelle. Ce chapitre propose de clarifier les définitions et les usages de cette notion, et présente les principaux problèmes qui se présentent lorsque l'on souhaite produire automatiquement des analyses en classes flexionnelles à partir des formes de surface.

Après avoir défini quelques notions essentielles (section 1.1), nous présentons le rôle des classes flexionnelles dans les analyses constructives au sens de Blevins (2006), c'est à dire les analyses qui proposent une grammaire capable de générer les formes de surface (section 1.2). Dans la section 1.3, nous argumentons en faveur d'une analyse abstractive (Blevins 2006) de type Mot et Paradigme, qui selon nous permet mieux de saisir l'ensemble des similarités entre les paradigmes de surface des lexèmes. Enfin dans la section 1.4, nous discutons de la notion de canonicité, et de la façon dont elle peut être employée pour comparer des systèmes de classes flexionnelles différents.

1.1 Paradigme de flexion : définitions

Afin de décrire les systèmes flexionnels, nous utilisons le vocabulaire descriptif proposé par Matthews (1974). Un MOT est un signe dont le signifié est un contenu syntaxique et sémantique

et le signifiant est un MOT FORME. Le mot forme est la réalisation d'un ensemble de propriétés morphosyntaxiques, c'est à dire qu'il est fléchi. Il constitue une unité phonologique de surface. Un LEXÈME est une unité lexicale abstraite. Il représente le lien entre un ensemble maximal de mots formes qui ne diffèrent entre eux que par la flexion. On le représente généralement par une FORME DE CITATION conventionnelle (le nominatif pour les noms latin, l'infinitif pour les verbes français) notée en petite capitales. Les PROPRIÉTÉS MORPHOSYNTAXIQUES sont des paires d'attributs (catégories) et de valeurs. Par exemple, le genre, le nombre ou le temps sont des attributs, et le féminin, le pluriel ou le passé sont respectivement des valeurs possibles de ces attributs. Nous notons souvent ces propriétés par leur seules valeurs, par exemple : F ou PL. Il existe deux principaux sens du mot PARADIGME dans le contexte de la morphologie flexionnelle, présentés comme suit par Carstairs-McCarthy (1991, p. 639) :

> Nommons la notion abstraite « paradigme$_1$ » et la notion plus concrète « paradigme$_2$ » et définissons-les comme suit :
>
> PARADIGME$_1$: l'ensemble de combinaisons de propriétés ou traits morphosyntaxiques (ou l'ensemble des "cases") réalisées par les formes fléchies des mots (ou lexèmes) pour une classe de mot donnée (ou catégorie, ou classe de lexème) dans une langue donnée.
>
> PARADIGME$_2$: l'ensemble de réalisations flexionnelles exprimant un PARADIGME$_1$ pour un mot (ou lexème) donné dans une langue donnée. [1]

Suivant Carstairs-McCarthy, nous nommons CASES les ensembles maximaux de propriétés morphosyntaxiques réalisés par des mot formes. Carstairs-McCarthy s'appuie ici sur la notion de RÉALISATIONS, dont nous montrerons qu'elle n'est pas une donnée de l'analyse mais constitue elle-même une forme d'analyse. Nous distinguerons le PARADIGME ABSTRAIT, ensemble de

1. [En anglais dans le texte] « *Let us call the abstract notion "paradigm$_1$" and the more concrete one "paradigm$_2$", and define them as follows:*

PARADIGM$_1$: the set of combinations of morphosyntactic properties or features (or the set of 'cells') realized by inflected forms of words (or lexemes) in a given word-class (or major category or lexeme-class) in a given language.

PARADIGM$_2$: the set of inflectional realizations expressing a PARADIGM$_1$ for a given word (or lexeme) in a given language ».

cases pour lesquelles un lexème se fléchit, et le PARADIGME CONCRET, l'ensemble de mots-formes fléchis d'un lexème. Lorsque deux lexèmes partageant le même paradigme abstrait utilisent des contrastes formels exactement parallèles pour distinguer les cases, on dit qu'ils partagent le même PARADIGME TYPE. La notion de Carstairs-McCarthy de PARADIGME$_2$ fondée sur les réalisations affixales est donc une façon de rendre compte d'un paradigme type. Dans les perspectives Mot et Paradigme, chaque paradigme type est souvent illustré par le paradigme concret entier d'un lexème par type, nommé PARADIGME EXEMPLAIRE. Lorsque le contexte rend l'interprétation non ambiguë, nous parlons parfois simplement de PARADIGME pour désigner l'une ou l'autre de ces notions.

	NOM.SG	VOC.SG	ACC.SG	GEN.SG	DAT.SG	ABL.SG	NOM.PL	VOC.PL	ACC.PL	GEN.PL	DAT.PL	ABL.PL
ROSA	rosa	rosa	rosam	rosae	rosae	rosā	rosae	rosae	rosās	rosārum	rosīs	rosīs
DOMINUS	dominus	domine	dominum	dominī	dominō	dominō	dominī	dominī	dominōs	dominōrum	dominīs	dominīs

TABLEAU 1.1 – Deux paradigmes de noms latins

Le tableau 1.1 exemplifie ces notions. Les CASES DE PARADIGME y figurent en en-tête des colonnes, et les LEXÈMES en indice des lignes. Le tableau est rempli par les mots formes qui instancient un lexème pour une case donnée. La structure d'une ligne constitue le paradigme abstrait, chaque ligne elle-même représente un PARADIGME CONCRET. Les paradigmes abstraits des lexèmes d'un même système flexionnel sont les mêmes dans les cas simples.

Ce format est à peu de choses près le format numérique des données d'entrées que nous utilisons lors d'investigations automatisées. Stump et Finkel (2013) nomment « *plat* » un tableau organisé similairement qui présente l'ensemble des paradigmes types d'un système.

Lorsque cela est pertinent, nous nommons RADICAL le matériel phonologique qui, dans un mot forme manifeste l'identité du lexème, et EXPOSANT le matériel phonologique qui exprime des propriétés morphosyntaxiques. Souvent, les exposants considérés par ces analyses sont des AFFIXES. Dans le tableau 1.1, on analyse généralement ros- comme le radical du lexème ROSA et -a comme l'exposant suffixal du nominatif singulier.

De manière cruciale cependant, les notions de paradigme et de paradigme type ne sont pas dépendantes de la possibilité (ou de l'impossibilité) d'une segmentation des mots-formes en

radicaux et exposants.

1.2 Les classes flexionnelles dans les analyses constructives

Les CLASSES FLEXIONNELLES, souvent nommées déclinaisons pour les noms et les adjectifs et conjugaisons pour les verbes, sont un outil récurrent dans la description des paradigmes de flexion. Aronoff (1994, p. 64) en propose la définition suivante : « Une classe flexionnelle est un ensemble de lexèmes dont les membres sélectionnent tous le même ensemble de réalisations flexionnelles [2] ». Carstairs-McCarthy (1994, p. 639) propose une définition très similaire : « un ensemble de mots (lexèmes) présentant le même paradigme$_2$ dans une langue donnée [3] ».

Nous illustrons ces définitions par les douze paradigmes types de noms latins masculins présentés dans le tableau 1.2, tels qu'ils sont présentés par Stump et Finkel (2013). Dans ce tableau, les paradigmes types sont représentés sous la forme de réalisations suffixales.

Dans cette table, chaque ligne représente une classe flexionnelle distincte. Ces classes sont habituellement présentées à travers le paradigme concret entièrement fléchi d'un lexème exemplaire par classe. Nous indiquons la forme de citation de ces lexèmes sur chaque ligne. Suivant ces définitions, un système flexionnel s'organise en classes dès lors qu'il existe au moins deux lexèmes qui n'utilisent pas la même réalisation pour au moins une case.

Blevins (2006) nomme CONSTRUCTIVES les analyses morphologiques fondées sur un lexique et une grammaire formelle, et qui visent à construire les formes observées à partir d'éléments abstraits ou plus petits. Dans ce cadre, le problème posé par les classes flexionnelles est celui de l'affectation des règles (ou des morphèmes) aux lexèmes (ou radicaux) correspondants : s'il existe plusieurs types de paradigmes concrets distincts exprimant le même paradigme abstrait, comment associer correctement chaque lexème avec les réalisations appropriées ? La solution consiste généralement à nommer chaque classe (par exemple « Première déclinaison » ou « 1 »), et à étiqueter les entrées lexicales pour leur affecter l'une de ces classes (Chomsky 1965 ; Matthews 1965). Aronoff (1994) décrit ainsi cette solution : « L'entrée lexicale du nom doit donc

2. [En anglais dans le texte] « *An inflectional class is a set of lexemes whose members each select the same set of inflectional realizations* ».

3. [En anglais dans le texte] « *a set of words (lexemes) displaying the same paradigm$_2$ in a given language* ».

			Singulier						Pluriel					
Déclinaison		Lexème exemplaire	NOM	VOC	ACC	GEN	DAT	ABL	NOM	VOC	ACC	GEN	DAT	ABL
Première	(1)	ROSA	a	a	am	ae	ae	ā	ae	ae	ās	ārum	īs	īs
Seconde	(2a)	DOMINUS	us	e	um	ī	ō	ō	ī	ī	ōs	ōrum	īs	īs
	(2b)	AGER	—	—	um	ī	ō	ō	ī	ī	ōs	ōrum	īs	īs
	(2c)	TEMPLUM	um	um	um	ī	ō	ō	a	a	a	ōrum	īs	īs
Troisième	(3a)	REX	s	s	em	is	ī	e	ēs	ēs	ēs	um	ibus	ibus
	(3b)	CONSUL	—	—	em	is	ī	e	ēs	ēs	ēs	um	ibus	ibus
	(3c)	CORPUS	—	—	—	is	ī	e	a	a	a	um	ibus	ibus
(i-stems)	(3d)	CIVIS	is	is	em	is	ī	e	ēs	ēs	ēs	ium	ibus	ibus
	(3e)	URBS	s	s	em	is	ī	e	ēs	ēs	ēs	ium	ibus	ibus
	(3f)	MARE	e	e	e	is	ī	ī	ia	ia	ia	ium	ibus	ibus
Quatrième	(4a)	MANUS	us	us	um	ūs	uī	ū	ūs	ūs	ūs	uum	ibus	ibus
	(4b)	CORNU	ū	ū	ū	ūs	ū	ū	ua	ua	ua	uum	ibus	ibus
Cinquième	(5)	RES	ēs	ēs	em	ēī	ēī	ē	ēs	ēs	ēs	ērum	ēbus	ēbus

TABLEAU 1.2 – Terminaisons des noms latins organisés en déclinaisons. Cette table étend Stump et Finkel (2013, p. 183, Table 7.1). Nous ajoutons les neutres et féminins, excluons le locatif, réordonnons les cases de paradigme, et ajoutons une numérotation afin de faciliter la référence à des déclinaisons spécifiques.

porter quelque étiquette pour garantir qu'il manifestera le bon ensemble de désinences. Cette étiquette est la classe flexionnelle du nom [4] ». Par exemple, l'entrée lexicale du lexème ROSA pourrait comprendre la propriété {CLASSE : 1 }. Par la suite, les règles, morphèmes, processus et autres dispositifs de flexion qui génèrent les formes de la déclinaison 1 seront également étiquetées de façon à sélectionner exclusivement les lexèmes ayant la propriété correspondante.

Ces étiquettes de classe flexionnelles constituent des propriétés d'une nature particulière : contrairement aux propriétés morphosyntaxiques, elles ne sont pas accessibles par la syntaxe. Cette différence de statut se reflète chez Stump (2001) par un traitement distinct des propriétés de classes et des autres propriétés lors de la sélection des règles. Aronoff (1994) écrit : « Cependant les classes flexionnelles ne se situent pas à l'interface entre la morphologie et un autre niveau linguistique, ni n'ont aucun propriété substantielle caractéristique d'un autre niveau. Elles sont purement morphologiques [5] ». Dans ses termes, les classes flexionnelles sont MORPHOMIQUES. Stump et Finkel (2013, p. 11) font de cette propriété une partie de la définition des classes :

> Une classe flexionnelle est une classe J de lexèmes tels que (i) les membres de J sont distingués par un patron de flexion commun et (ii) l'appartenance à J n'a aucune pertinence syntaxique [6].

L'existence des classes flexionnelles soulève donc quatre questions principales.

Premièrement, les étiquettes de classe ne sont justifiables que lorsqu'aucun autre prédicteur ne permet de motiver des réalisations concurrentes, ce qui conduit les linguistes à chercher des solutions optimales minimisant les distinctions de classes.

Deuxièmement, et en conséquence, les classes flexionnelles étant le lieu d'une complexité qui semble « gratuite » ou « inutile », leur présence est au premier abord surprenante. Depuis Carstairs (1987), de nombreux travaux se sont interrogés sur les bornes de cette complexité

4. [En anglais dans le texte] « *The lexical entry for the noun must therefore bear some sort of flag to assure that it will manifest the appropriate set of inflections. This flag is the inflectional class of the noun* ».

5. [En anglais dans le texte] « *The inflectional classes, however, neither mediate between morphology and another linguistic level nor have any substantial properties characteristic of another. They are purely morphological* ».

6. [En anglais dans le texte] « *An inflection class is a class J of lexemes such that (i) J's members are distinguished by a common pattern of inflection and (ii) membership in J has no syntactic significance* ».

morphologique, et les conditions de son existence.

Troisièmement, si le problème des classes flexionnelles se rapporte à une stratégie d'écriture de grammaire constructive, il est tout à fait possible qu'il existe plus d'une bonne solution, comme semble en témoigner les inventaires de classes contradictoires trouvés dans la littérature pour de mêmes systèmes. Pour autant, faut-il penser que les classes flexionnelles ainsi définies ne sont qu'un dispositif descriptif ? Comment évaluer leur réalité ?

Quatrièmement, l'un des problèmes cruciaux posés par cette définition est qu'elle présuppose une segmentation des formes en radicaux et affixes. Or cette segmentation n'est pas une donnée de l'analyse, et elle est extrêmement difficile à systématiser.

Dans les sections qui suivent, nous examinons tour à tour chacune de ces questions, en nous demandant comment une étude automatisée des systèmes flexionnels peut contribuer à y répondre.

1.2.1 Prédictibilité et étiquettes de classe

Le seul indice de l'existence des étiquettes de classes est l'impossibilité de sélectionner correctement les réalisations sans y référer. Le principe du rasoir d'Occam veut donc qu'une analyse ne présente de telles propriétés que s'il est impossible de faire autrement.

Dans le tableau 1.2, les lignes (4a) et (4b) représentent des classes flexionnelles distinctes au sens où elles ne partagent pas les mêmes terminaisons. Cependant, dans une perspective constructive, il n'est pas nécessaire de définir deux étiquettes de classe distinctes pour ces deux classes, car la classe (4b) ne comporte que des noms neutres, et la classe (4a) aucun neutre. Plutôt que de sélectionner d'une part {CLASSE : 4a} et d'autre part {CLASSE : 4b }, il est possible de sélectionner respectivement {CLASSE : 4, GENRE : m $\vee$ f} et d'autre part {CLASSE : 4, GENRE : n}.

Il est possible d'utiliser le genre comme prédicteur de façon à réduire également l'ensemble des autres classes neutres : (2c), (3c) et (3f). Carstairs-McCarthy (1994) va jusqu'à fusionner, pour les masculins, la première et la seconde classe, suivant l'argument que les rares noms masculins appartenant à la première classe sont en fait grammaticalement marqués par un genre féminin, mais sont masculins sémantiquement (ils désignent des hommes par leur pro-

fession). Il est donc possible de les distinguer des masculins de la seconde déclinaison qui, eux, sont marqués grammaticalement par le masculin.

La forme phonologique des radicaux des lexèmes peut également servir de prédicteur afin de réduire le nombre d'étiquettes de classe. Il est alors possible de réinterpréter les réalisations divergentes de ces classes comme des allomorphes phonologiquement conditionnés. Il existe donc une relation de vases communicants entre allomorphie et classes flexionnelles, lorsqu'il s'agit de rendre compte de la variation entre paradigmes types. Carstairs-McCarthy (1994, p. 759) argumente sur cette base que l'ensemble des classes (3a) à (3f) peuvent être considérées comme une unique classe : « La troisième déclinaison constituait une unique macroclasse au sein de laquelle des affixes apparemment rivaux étaient distribués en fonction d'un mélange complexe de traits phonologiques (comme la forme du radical), de traits syntaxiques lexicalement déterminés (comme le genre) et de traits sémantiques [7] ». L'intuition centrale ici est celle de la compétition : pour la dérivation formelle d'une forme par la grammaire, il est inutile de postuler l'existence d'étiquettes de classe distinctes si les stratégies envisageables ne sont pas en compétition.

C'est pour cette raison que des systèmes flexionnels par ailleurs bien connus ont pu être décrits comme présentant un nombre variable de classes flexionnelles, suivant l'attribution de certaines différences entre paradigmes types à la phonologie, à l'allomorphie, ou à des différences de propriétés sémantiques ou grammaticales. Par exemple, Plénat (1987) propose d'analyser le système verbal du français, traditionnellement décrit comme comportant trois conjugaisons, en seulement deux conjugaisons distinctes. Corbett (1982, p. 202), déjà cité dans l'introduction, remarque qu'« il n'y a aucun accord concernant le nombre de paradigmes nominaux en russe. La tradition répond qu'il y en a trois, certains auteurs en revendiquent quatre, et plus récemment il a été suggéré que seulement deux paradigmes étaient nécessaires [8] ». Blevins (2004) mentionne que les descriptions du système nominal de l'Estonien comptent entre 26 et 400

7. [En anglais dans le texte] « *the third declension constituted a single macroclass within which apparently synonymous rival affixes were distributed on the basis of a complex mixture of phonological features (such as stem shape), lexically determined syntactic features (such as gender) and semantic features* ».

8. [En anglais dans le texte] « *there is no agreement as to the number of nominal paradigms in Russian. Tradition answers three, some writers claim four, and more recently it has been suggested that only two paradigms are required* ».

paradigmes types, réunis en 6 à 12 classes.

1.2.2 Bornes à la complexité des systèmes de classes

Carstairs (1987) remarque qu'étant donné un inventaire d'affixes disponibles pour chaque case d'un système flexionnel, le nombre maximum de classes flexionnelles distinctes correspond au produit du nombre d'affixes disponible par cases. Le tableau 1.3 présente cet inventaire pour le système nominal du latin, ainsi que le compte des affixes par case. Le plus grand nombre de classes possible pour le latin est donc $9\times9\times6\times5\times6\times6\times7\times7\times7\times6\times3\times3 = 1\,620\,304\,560$, c'est à dire supérieur au milliard [9]. Pour le système du latin, ces valeurs maximales sont de toute façon beaucoup plus hautes que le nombre de noms attestés.

S'il existait autant de classes que de lexèmes, la notion même de classe flexionnelle serait inutile. L'ensemble du système devrait être appris, comme l'écrit Carstairs (1987), « Chaque nom du langage L devrait être appris avec non moins de douze "règles" spécifiant comment chaque combinaison de nombre et de cas devait être réalisé [10] ».

Il existe un lien entre la complexité d'un système de classes et l'interprédictibilité entre ses cases (Carstairs 1987 ; Matthews 1991). En effet, dans un système maximal tel que nous venons de le décrire, toutes les combinaisons d'affixes possible existent. En conséquence, connaître l'affixe que prend un lexème dans une ou plusieurs cases ne restreint pas, pour un locuteur qui n'en connaîtrait pas la classe flexionnelle, le choix des affixes de ce lexème pour les autres cases. En somme, la connaissance de la réalisation d'une case ne serait pas informative. En ce cas, les réalisations des cases de paradigme pourraient être dites entièrement indépendantes.

Mais on observe au contraire seulement treize classes en latin : son système flexionnel présente donc des relations de prédictibilité entre formes ou réalisations. Carstairs (1987) montre que l'on peut généraliser cette observation. Les classes flexionnelles présentent donc en gé-

9. Le tableau 1.3 révèle que trois paires de cas présentent exactement les mêmes inventaires d'exposants : le nominatif et le vocatif singulier ; le nominatif et vocatif pluriels ainsi que le datif et l'ablatif pluriel. On pourrait argumenter que ces cas ne doivent figurer qu'une fois dans le calcul. On obtient alors : $9\times6\times5\times6\times6\times7\times7\times6\times3 = 8\,573\,040$, soit plus de huit millions de classes potentielles.

10. [En anglais dans le texte] « *each noun in L would have to be learnt along with no less than twelve "rules" specifying how each Number and Case combination was to be spelt/* ».

Nombre	Cas	Affixes	Compte
SG	NOM	—, -a, -e, -ēs, -is, -s, -ū, -um, -us	9
	VOC	—, -a, -e, -ēs, -is, -s, -ū, -um, -us	9
	ACC	—, -am, -e, -em, -ū, -um	6
	GEN	-ī, -ae, -ēī, -is, -ūs	5
	DAT	-ae, -ēī, -ī, -ō, -ū, -uī	6
	ABL	-ā, -ō, -e, -ī, -ū, -ē	6
PL	NOM	-a, -ae, -ēs, -ī, -ia, -ua, -ūs	7
	VOC	-a, -ae, -ēs, -ī, -ia, -ua, -ūs	7
	ACC	-a, -ās, -ēs, -ia, -ōs, -ua, -ūs	7
	GEN	-ārum, -ērum, -ium, -ōrum, -um, -uum	6
	DAT	-ēbus, -ibus, -īs	3
	ABL	-ēbus, -ibus, -īs	3

TABLEAU 1.3 – Inventaires d'affixes latins par cas.

néral une structure implicative (Wurzel 1989) : connaître un affixe réduit considérablement le nombre de possibilités pour les autres cases.

Le nombre minimum théorique de classes dans un tel système correspond au nombre maximal d'affixes concurrents pour une case, ici 9 (nominatif ou vocatif singulier). Sur la base des données du latin, du hongrois, du zulu et du dyirbal, Carstairs (1987) définit le principe d'économie du paradigme (PEP), cité ci-dessous, selon lequel ce minimum correspond à la limite effective du nombre de classes distinctes dans un système flexionnel. Cependant, Carstairs ne compte pas comme classes chaque paradigme type distinct mais plutôt ce qu'il nomme MACRO-PARADIGME, ensemble de paradigmes types dont les réalisations ne sont pas en concurrences pour le locuteur, soit en raison d'une prédictibilité phonologique, soit en raison d'autres traits distincts, comme par exemple le genre.

> PRINCIPE D'ÉCONOMIE DU PARADIGME Quand dans une langue L donnée plus d'une réalisation flexionnelle est disponible pour un ou des ensembles de propriétés morphosyntaxiques non lexicalement déterminées associées avec une catégorie N, le nombre de macroparadigmes pour N n'est pas supérieur au nombre de macroinflexions rivales distinctes disponibles pour l'ensemble qui est le plus fourni en réalisations rivales [11].

Formulé autrement, le PEP est satisfait si, dans un système flexionnel, il existe au moins une case pour laquelle toutes les réalisations affixales sont des identifiants de classes. Il est alors possible, à partir de n'importe quelle forme de cette case, de prédire l'ensemble du reste de son paradigme. Par la suite, Carstairs-McCarthy (1994) raffine son analyse, en formulant le No Blur Principle (NBP), dont le nom fait référence au principe de contraste. Selon le NBP, dans un système flexionnel, tout affixe est soit un identifiant de sa classe, soit une réalisation par défaut. Il ne peut donc y avoir qu'un défaut par case. Le NBP est donc violé dès lors que pour au moins une case, il existe plusieurs affixes qui sont communs à plusieurs classes.

11. [En anglais dans le texte] « *PARADIGM ECONOMY PRINCIPLE When in a given language L more than one inflexional realisation is available for some bundle or bundles of non lexically determined morphosyntactic properties associated with some part of speech N, the number of macroparadigms for N is no greater than the number of distinct "rival" macroinflexions available for that bundle which is most generously endowed with such rival realisations.* ».

Ce principe ne semble pas toujours respecté par les systèmes. Par exemple, dans notre système du latin (tableau 1.2), les affixes -um et -em sont tous deux répétés entre plusieurs classes à l'accusatif singulier, et ne peuvent donc être ni des identifiants ni des défauts. Suivant le raisonnement de Carstairs-McCarthy, si un système de classes peut être décrit d'une façon qui satisfasse le NBP, alors cette classification est la bonne, même s'il est possible de formuler d'autres classifications qui ne satisfont pas ce principe. Il dispose de deux moyens principaux afin de trouver une classification qui respecte le NBP. Le premier découle du fait qu'il ne considère comme réalisations flexionnelles que les affixes, et que les affixes dépendent d'une segmentation des formes qui est sujette à arbitrages. Le choix d'une segmentation plutôt qu'une autre peut donc altérer les paradigmes types observés en rejetant ou non une partie de la variation dans les radicaux. Le second se situe dans la notion de MACROPARADIGME citée plus haut, qui permet de fusionner des paradigmes types.

Carstairs-McCarthy explore deux segmentations alternatives des affixes des noms latins masculins (les autres genres n'étant par définition par en concurrence avec ceux-ci, ils sont ignorés de l'analyse). Nous reproduisons ces deux analyses dans le tableau 1.4. Nous montrons les colonnes « floues » en grisant les cases concernées.

L'analyse traditionnelle, présentée en haut du tableau, ne respecte pas entièrement le NBP : par exemple au datif pluriel, -īs n'est ni un identifiant de classe (deux autres classes présentent le même), ni un défaut (-ibus est également répété entre plusieurs classes). L'analyse alternative, présentée en bas du tableau, abstrait les voyelles thématiques, considérées comme une partie du radical. Cette analyse présente deux fois plus de colonnes qui enfreignent le NBP. En conséquence, Carstairs-McCarthy choisit la première segmentation.

Afin de supprimer le flou restant, Carstairs-McCarthy fait appel à la notion de macroparadigme. il propose de fusionner les noms masculins des deux premières déclinaisons, comme nous l'avions mentionné plus haut, et de fusionner ensemble les variantes de la troisième déclinaison dont les alternances sont prédictibles phonétiquement. Il obtient ainsi une analyse des noms latins masculins en trois classes qui respectent le NBP.

	Singulier						Pluriel			
Déclinaison	NOM	VOC	ACC	GEN	DAT	ABL	N/V	ACC	GEN	D/A
Première	a	a	am	ae	ae	ā	ae	ās	ārum	īs
Seconde	us / —	e	um	iī	ō	ō	ī	ōs	ōrum	īs
Troisième	s / — / ēs	s / — / ēs	em	is	ī	e	ēs	ēs	um	ibus
	s	s	em	is	ī	e	ēs	ēs	ium	ibus
	is	is	im >em	is	ī	ī >e	ēs	īs >ēs	ium	ibus
Quatrième	us	us	um	ūs	uī	ū	ūs	ūs	uum	ibus

	Singulier						Pluriel			
Déclinaison	NOM	VOC	ACC	GEN	DAT	ABL	N/V	ACC	GEN	D/A
Première	—	—	m	ī	ī	V:	ī	:s	:rum	īs
Seconde	s / —	e	m	ī	V:	V:	ī	:s	:rum	īs
Troisième	s / —	s / —	m	s	ī	e	ēs	ēs	um	bus
	s	s	m	s	ī	e	ēs	ēs	um	bus
	s	s	m	s	ī	V: >e	ēs	:s >ēs	um	bus
Quatrième	s	s	m	s	ī	V:	:s	:s	um	bus

TABLEAU 1.4 – Affixes nominaux latins avec et sans voyelles thématiques. Cette table est adaptée de Carstairs-McCarthy (1994, pp. 749-750). Le symbole > signifie « tend à être remplacé par ». Une barre oblique sépare les affixes qui sont distribués sur une base partiellement phonologique. Dans chaque colonne, les cases grisées indiquent les affixes répétés qui violent le No Blur Principle.

1.2.3 La réalité du concept de classes flexionnelles

On peut cependant se demander si cette solution est unique : il se peut qu'il existe plusieurs façon concurrentes de regrouper les paradigmes afin de satisfaire le NBP. Carstairs (1987, p. 71), conscient de cette possibilité, nomme l'existence de telles analyses concurrentes « *macroparadigm ambivalence* ». Il illustre ce problème sur un langage imaginaire, et formule l'hypothèse qu'il ne se pose en fait pas réellement en pratique, justifiant (p.72) : « Pour le matériel du russe, du Dyirbal, du Zulu et du latin que nous avons considéré, il est façile de montrer qu'une seule solution en macroparadigmes est plausible [12] ».

Cependant, ses analyses de ces systèmes ne semblent pas si arrêtées. Tout d'abord, comme nous l'avons déjà remarqué, il existe pour le russe des analyses argumentées proposant un nombre de classes très variable. L'analyse du Latin par Carstairs ne fait pas non plus consensus. Par ailleurs, il existe un problème de circularité à définir le PEP ou le NBP comme des propriétés observables des systèmes, tout en construisant les systèmes de classes de façon à ce qu'ils s'y conforment. Blevins (2004) compare la pratique d'ajuster les frontières de classes de cette façon au procédé qui consiste à manipuler les circonscriptions électorales afin d'influencer l'issue d'un vote, ou « charcutage électoral » (en anglais : « *gerrymandering* »). Il montre que si Carstairs-McCarthy (1994) prend toujours soin de consolider ensemble des macroparadigmes qui semblent intuitivement associés, il serait tout aussi possible de fusionner des paradigmes qui présentent des contrastes discriminants mais ne sont pas habituellement groupés ensembles. Dès lors, il s'interroge : « Quel principe évite la consolidation de paradigmes non apparentés qui ne sont jamais réunis ensembles [13] » ?

Il nous semble au contraire qu'il se peut qu'un unique système accepte plusieurs analyses en classes distinctes mais équivalentes, comme semblent l'indiquer l'existence d'analyses concurrentes des mêmes systèmes. En général, les analyses constructives se donnent non pour une possible grammaire générant les données parmi d'autres, mais pour une image plausible du

12. [En anglais dans le texte] « *For the Russian, Dyirbal, Zulu and Latin material we have considered, it is easy to show that only one macroparadigm solution is plausible* ».

13. [En anglais dans le texte] « *what principle prevents the consolidation of unrelated paradigms that are never grouped together ?* ».

système mental des locuteurs. S'il existe plusieurs solutions équivalentes, il semble impossible de deviner laquelle fait partie du modèle des locuteurs. Il faudrait alors admettre que divers locuteurs peuvent en avoir des modèles différents. S'attend-on alors à ce qu'il existe des traces de ces différences de modèles dans le comportement linguistique de ces locuteurs ?

L'existence de systèmes de classe alternatifs équivalents suggère que les classes flexionnelles ne sont pas des objets naturels, mais des classifications construites à partir des données. Dès lors, il est compréhensible que différentes questions de recherche mènent à différentes classifications, sans que l'une soit nécessairement meilleure que l'autre.

1.2.4 Identité des unités de flexion

Nous avons argumenté que, étant donné un ensemble d'affixes déterminés pour un système flexionnel, l'organisation du lexique en classes flexionnelles constitue un construit de l'analyse et non une propriété brute des données. Le même argument peut être mené à propos de la segmentation des formes, qui nécessite des arbitrages parfois difficiles entre radical et affixes puis entre affixes concurrents et allomorphie. Revenant à l'argumentation de Carstairs concernant les voyelles thématiques du latin, nous remarquons que différentes hypothèses concernant la segmentation des formes peuvent mener à des systèmes de classes apparemment très différents.

Il n'existe pas de consensus sur la façon d'inférer systématiquement une segmentation de formes en radicaux et affixes (Spencer 2012). Les contraintes que l'on peut considérer pour cela peuvent même constituer des pressions contradictoires. Par exemple, si l'on veut minimiser le nombre d'affixes par case, on peut employer la stratégie appliquée par Boyé (2000, p. 125) qui consiste à définir les affixes comme la partie commune à l'ensemble des mots formes d'une case. Quel que soit le lexème, cette stratégie fait disparaître toute classe flexionnelle si l'on regarde exclusivement les affixes, mais les restitue potentiellement dans les radicaux, qui comprennent alors en partie du matériel parfois classé comme flexionnel. À l'opposé, on peut tenter de maximiser la régularité des radicaux (Embick et Halle 2005). Trouver un équilibre entre ces extrêmes dépend généralement de l'intuition du linguiste, et il n'existe pas de formalisation détaillée de ce processus, permettant de l'implémenter et de garantir sa reproductibilité.

Walther et Sagot (2011) fournissent un moyen de comparer de façon computationnelle

des analyses concurrentes de systèmes flexionnels. Walther (2013) conçoit un cadre formel très général permettant d'encoder des analyses concurrentes. Ce formalisme génératif, nommé Alexina, est implémenté et permet de générer des lexiques de grande tailles. Walther et Sagot (2011) comparent des analyses distinctes implémentées dans ce formalisme en évaluant leur compacité en s'appuyant sur la notion de longueur de description minimale. Ils peuvent ainsi mesurer pour chaque analyse les coûts respectifs du lexique et de la grammaire, et choisir l'analyse présentant la plus petite longueur de description. La méthodologie de Sagot et Walther (2011), Walther et Sagot (2011) et Walther (2013) permet d'identifier l'analyse la plus élégante lorsqu'on dispose de plusieurs analyses distinctes d'un même système flexionnel. Elle ne permet cependant pas de trouver une analyse affixale à partir des formes, de découvrir une analyse meilleure que celle dont on dispose, ni même d'évaluer une analyse unique dans l'absolu sans la comparer à d'autres. Par ailleurs, les comparaisons de Sagot et Walther (2011), Walther et Sagot (2011) et Walther (2013) aboutissent souvent à des mesures de complexités très proches pour des analyses qualitativement très différentes. Plus encore, il n'est pas certain que les grammaires mentales des locuteurs reflètent toujours l'organisation la plus compacte, ni que cette analyse soit la seule utile ou saillante pour les locuteurs. En somme, si les travaux de Sagot et Walther constituent la tentative la plus convaincante d'arbitrer entre des analyses constructives, ils soulignent en creux le caractère largement arbitraire du choix d'une telle analyse.

Par ailleurs, la focalisation sur les affixes fait l'hypothèse implicite d'une primauté des systèmes agglutinatifs, qui pose problème dans une perspective typologique. Ainsi Hockett (1987) écrivait dans un chapitre intitulé « La grande fraude agglutinative [14] » :

> Avec nos techniques ingénieuses, nous rendions compte des alternances en concevant un « analogue agglutinatif » de la langue et formulions des règles qui convertissaient les expressions de cet analogue vers les formes réellement exprimées. Bien sûr, même un tel analogue agglutinatif, accompagné de ses règles de conversion, pourrait être interprété seulement comme un outil descriptif. Mais il n'était généralement pas conçu de cette façon ; à l'opposé, il était considéré comme un reflet

14. [En anglais dans le texte] « *The great agglutinative Fraud* ».

> direct de la réalité. Nous semblions convaincus que, quelles que soient les apparences, toutes les langues étaient « vraiment » agglutinatives [15].

Ce sentiment est partagé par Matthews (1991), qui écrit des morphèmes : « Un défenseur des grammaires anciennes [NDA : les approches Mot et Paradigme] répondrait que ces éléments sont des fictions. Ils sont créés par la méthode moderne [NDA : les approches morphémiques] ; et, si nous les imposons sur un système flexionnel, il est inévitable que nous le décrivions comme un système agglutinatif qui a, d'un façon ou d'une autre, mal tourné [16] ».

La segmentation des formes, qu'elle prenne spécifiquement la forme d'une division radical-affixe ou non, constitue donc en elle-même une classification flexionnelle des lexèmes, et n'est pas un donné de l'analyse. Comment envisager une analyse qui prenne au sérieux ce constat ?

1.3 Vers une perspective abstractive

Les grammaires « anciennes » dont parle Matthews sont les grammaires de la tradition latine, qui présentent les systèmes flexionnels au moyen d'une série de paradigmes exemplaires illustrant l'ensemble des paradigmes types d'un système. Par la suite, il suffit d'un petit nombre de formes pour savoir de quel paradigme type relève un lexème, et pour le fléchir par analogie aux lexèmes exemplaires. L'ensemble de cases servant ainsi de prédicteur se nomme PARTIES PRINCIPALES. Ces analyses sont ABSTRACTIVES, dans les termes de Blevins (2006), plutôt que CONSTRUCTIVES : elles s'attachent à décrire les relations entre les formes de surface plutôt qu'à fournir une grammaire qui les génère. Les unités fondamentales considérées par ces analyses sont les mots et les liens qu'ils entretiennent au sein des paradigmes : ce sont des analyses de

15. [En anglais dans le texte] « *with our clever morphophonemic techniques[, we] were providing for alternations by devising an "agglutinative analog" of the language and formulating rules that would convert expressions in that analog into the shapes in which they are actually uttered. Of course, even such an agglutinative analog, with its accompanying conversion rules, could be interpreted merely as a descriptive device. But it was not in general taken that way; instead, it was taken as a direct reflection of reality. We seemed to be convinced that, whatever might superficially appear to be the case, every language is "really" agglutinative* ».

16. [En anglais dans le texte] « *An apologist for ancient grammar would answer that these elements are fictions. They are created by the modern method; and, if we fist them on a flectional system, we are bound to describe it as an agglutinating system that has somehow gone wrong* ».

type Mot et Paradigme (Hockett 1954).

1.3.1 Le mot comme unité de prédictibilité

Pour les noms du latin, la tradition considère que le nominatif et le génitif fournissent conjointement assez d'information pour sélectionner le reste du paradigme. Dans le tableau 1.2, cette paire de cases suffit à distinguer la plupart des autres cases : seules les formes qui se déclinent comme ROSA prennent /-a/ au nominatif singulier et /-ae/ au génitif singulier, seules celles qui se déclinent comme DOMINUS prennent /-us/ au nominatif singulier et /-i/ au génitif singulier. Mais ces seules deux cases ne suffisent pas à distinguer les paradigmes types de REX et URBS, qui se distinguent uniquement par le génitif pluriel, ou ceux de CONSUL et CORPUS, qui se distinguent par l'accusatif singulier, et les nominatif, vocatif et accusatif pluriels (ainsi que par leur genre, puisque la déclinaison de CORPUS est propre aux noms neutres).

Stump et Finkel (2013) se sont appuyés sur la notion de système de parties principales pour développer une typologie des systèmes flexionnels. Ils utilisent un outillage informatisé afin de mesurer précisément les variables typologiques étudiées. Tandis que les grammaires traditionnelles et pédagogiques avancent l'existence d'un unique ensemble de parties principales permettant de prédire l'ensemble des autres cases du paradigme pour l'ensemble des lexèmes, ils s'intéressent à toutes les formes que peut prendre un tel ensemble. Le Principal Part Analyzer est un programme qu'ils ont développé et qui permet de calculer les propriétés des systèmes de parties principales. Si on lui soumet le paradigme d'affixes du tableau 1.2, celui-ci propose en sortie les quatre ensembles de parties principales possibles suivants [17] :

1. {nom.sg, acc.sg, gen.pl }
2. {nom.sg, nom.pl, gen.pl }
3. {voc.sg, acc.sg, gen.pl }
4. {voc.sg, nom.pl, gen.pl }

17. Stump et Finkel (2013) étudient les implications entre cases de paradigme à travers les relations entre règles d'exponence, et non entre mots formes. Leurs analyses ne sont pas constructives, et s'apparentent à des analyses abstractives sans toutefois s'intéresser directement aux formes de surface. Elles constituent un compromis entre approches fondées sur le mot et approches fondées sur l'exponence.

Comme nous l'avons vu, Carstairs-McCarthy a conclu de l'observation du système nominal du latin qu'une analyse qui conserve les voyelles thématiques dans les affixes respecte mieux le NBP. Or le NBP, comme le PEP, est une contrainte sur l'indépendance des cases de paradigmes. Ces principes imposent une prédictibilité des réalisations entre elles : si une réalisation est diagnostique d'une classe, elle suffit à prédire l'ensemble des autres réalisations des lexèmes de cette classe. Il n'est pas surprenant que la segmentation qui attribue le plus de matériel aux affixes mène à une meilleure prédictibilité au sein du paradigme : l'ajout d'information ne peut en effet qu'améliorer cette prédictibilité. Nous pourrions donc employer le même moyen pour justifier une analyse qui prend en compte les radicaux au même titre que les affixes : un mot entier est nécessairement un prédicteur de classe égal ou meilleur qu'un affixe. C'est pour cette raison que Blevins (2004, p. 42) écrit : « les généralisation sur les exposants affixaux dérivent des patrons d'interdépendance qui impliquent les mots entiers [18] ». Blevins (2016) montre que les segmentations en affixes et radicaux opacifient artificiellement les paradigmes, et que la pratique d'utiliser des étiquettes de classes constitue un palliatif à ce problème.

Il apparaît donc que le mot constitue une unité plus utile que l'affixe – ou tout autre forme d'exposant – pour l'étude de la complexité, et donc de la prédictibilité des paradigmes, dans les systèmes flexionnels (Robins 1959). Blevins (2004) remarque justement que dans une perspective Mot et Paradigme, il n'est pas besoin d'un principe d'économie pour expliquer la dépendance entre cases du paradigme. Celle-ci est au contraire un des fondements théoriques de l'approche : ce type d'analyse n'est possible que parce qu'il existe des dépendances fortes entre cases de paradigme. Par ailleurs, une analyse Mot et Paradigme a l'avantage de ne pas nécessiter de référence explicite à une segmentation particulière des formes en affixes, qui comme nous l'avons vu ne va pas de soi.

L'exemple du burmeso illustre de manière spectaculaire le contraste entre approches constructives et approches abstractives. Le tableau 1.5 présente deux classes flexionnelles des verbes du burmeso telles qu'elles sont décrites par Corbett (2009), sous la forme d'exposants affixaux. D'un point de vue constructif, ce système flexionnel semble extrêmement complexe : chaque

18. [En anglais dans le texte] « *generalizations over affixal exponents are derivative of patterns of interdependence involving whole words* ».

	Classe 1		Classe 2	
	SG	PL	SG	PL
I	j-	s-	b-	t-
II	g-	s-	n-	t-
III	g-	j-	n-	b-
IV	j-	j-	b-	b-
V	j-	g-	b-	n-
VI	g-	g-	n-	n-

TABLEAU 1.5 – Classes flexionnelles verbales du burmeso (Corbett 2009, d'après Donohue 2001).

classe constitue un système d'affixes entièrement indépendant. Ces moyens concurrents pour exprimer les mêmes valeurs semblent, dans la perspective constructive, excédentaires, et se traduiraient par une grammaire « double ». Pourtant, dans une perspective abstractive, cette complexité disparaît. En effet, il est toujours trivial dans ce système de savoir à quelle classe appartient un mot, et de prédire l'ensemble de ses autres formes. N'importe quelle case constitue une partie principale parfaite.

Afin de rendre compte de ce contraste, Ackerman et Malouf (2013) distinguent la COMPLEXITÉ ÉNUMÉRATIVE d'un système, ou E-COMPLEXITÉ, qui concerne le nombre de distinctions faites par un système, et la taille de l'inventaire de moyens servant à les marquer ; et la COMPLEXITÉ INTÉGRATIVE, ou I-COMPLEXITÉ définie comme la difficulté que pose le système pour les locuteurs. En raison de la distribution zipfienne des mots formes en corpus, les locuteurs sont exposés à certaines formes, mais jamais au paradigme entier d'un lexème. En conséquence, ils doivent faire des inférences afin de produire les formes auxquelles ils n'ont pas été exposés. Ackerman et Malouf (2013) nomment ce problème le PARADIGM CELL FILLING PROBLEM (PCFP). Ils mesurent la complexité intégrative comme la difficulté moyenne que pose le PCFP au sein d'un système flexionnel.

Tandis que les systèmes flexionnels varient beaucoup en E-complexité, Ackerman et Malouf (2013) observent que ce n'est guère le cas en termes d'I-complexité, qui est uniformément

basse dans les systèmes qu'ils mesurent. Puisque celle-ci reflète une organisation en patrons implicatifs des paradigmes, ils concluent que « les approches grammaticales qui considèrent les mots, les paradigmes et les constructions clausales comme des unités primaires de l'analyse semblent bien fondées [19] ».

1.3.2 Des classes d'identité aux classes de similarité

Dans une perspective abstractive, l'élaboration d'une classification flexionnelle des lexèmes ne répond plus au besoin de minimiser une grammaire. S'il n'est plus nécessaire de fournir des étiquettes de classes servant au ré-assemblage de mots segmentés, il n'en reste pas moins que les systèmes flexionnels peuvent présenter un ensemble de paradigmes types, liés entre eux par des points de similarité. On peut donc se demander comment classer les lexèmes en fonction de leur comportement flexionnel. Nous nous demanderons tout d'abord quels types de classification sont le plus appropriés à décrire les analogies dans les paradigmes de flexion.

Revenons aux deux définitions des classes flexionnelles selon Carstairs-McCarthy (1994, p. 639) et Aronoff (1994, p. 64). Suivant celles-ci, un système de classes flexionnelles est une partition exhaustive de l'ensemble des lexèmes en plusieurs classes sans chevauchement. Les membres d'une classe partagent un comportement flexionnel identique. Par exemple, les classes (2a) et (2b) du tableau 1.2, quoiqu'elles partagent l'ensemble de leurs autres réalisations, sont distinctes en ce que (2b) ne présente aucune réalisation affixale au nominatif et vocatif singuliers.

Les 13 paradigmes types qui constituent les lignes du tableau 1.2 correspondent aux définitions d'Aronoff et de Carstairs-McCarthy, mais ne correspondent en fait pas à la classification traditionnelle du système flexionnel des noms latins. La tradition ne distingue que cinq classes, qui réunissent ensemble certaines des lignes et qui sont numérotées dans la première colonne du tableau. Au sein de ces classes, comme Dressler, Kilani-Schoch et al. (2008, p. 52) nous le rappellent, « tous les noms d'une classe ne se fléchissent pas exactement de la même façon [20] ».

19. [En anglais dans le texte] « *grammatical approaches that posit words and paradigms and clausal constructions as primary units of analysis appear to be well grounded* ».

20. [En anglais dans le texte] « *not all nouns of one class inflect in exactly the same way* ».

Par exemple, tandis que certains lexèmes de la troisième déclinaison se terminent en -ium au génitif pluriel (3d, 3e, 3f), d'autres se terminent en -um (3a, 3b, 3c). Plutôt qu'une identité, les membres de classes flexionnelles au sens traditionnel présentent un degré de similarité fort.

Cet exemple est représentatif d'une observation générale selon laquelle les descriptions traditionnelles de systèmes flexionnels distinguent un petit nombre de grandes classes, comprenant à la fois des patrons communs vus comme réguliers et des patrons plus rares vus comme des déviations à la situation régulière. Dans les faits, les analyses de linguistes se fondent souvent également sur ce type de classes.

Cette seconde définition rappelle les « macroparadigmes » de Carstairs-McCarthy, qui réunit certaines classes à condition qu'elles ne soient pas en concurrence. Le principe directeur pour leur élaboration est le respect des contraintes de complexité du système (comme le NBP) ; cependant leur définition mentionne explicitement un critère de similarité (Carstairs 1987, p. 69) :

> Un macroparadigme consiste en :
> (a) Deux ou plusieurs paradigmes similaires dont toutes les différences flexionnelles peuvent être expliquées phonologiquement, ou qui corrèlent de façon conséquente avec des différences de propriétés déterminées sémantiquement ou lexicalement ;
> ou
> (b) tout paradigme qui ne peut être combiné de cette façon avec un autre paradigme. [21]

D'autres auteurs s'appuient également sur ce type de définitions fondées sur la similarité. Brown et Hippisley (2012, p. 4) définissent les classes flexionnelles comme des « classes de lexèmes qui partagent des contrastes morphologiques similaires [22] ». Matthews (1991, p. 129) propose quant à lui : « les classes flexionnelles [...] sont des classes de lexèmes qui vont en-

21. [En anglais dans le texte] « *A macroparadigm consists of:*
(a) any two or more similar paradigms all of whose inflexional differences either can be accounted for phonologically, or else correlate consistently with differences in semantic or lexically determined syntactic properties;
or
(b) any paradigm which cannot be thus combined with other paradigm(s) ».

22. [En anglais dans le texte] « *classes of lexemes that share similar morphological contrasts* ».

semble relativement à une certaine flexion.[23] », enfin, Blevins (2004, p. 42) décrit ainsi les classes flexionnelles dans le modèle WP : « Les lexèmes qui se fléchissent similairement sont assignés à une conjugaison ou déclinaison commune ; c'est-à-dire une classe flexionnelle[24] ».

Dans une série de publications influentes, Dressler et ses collègues proposent de ne pas choisir entre les deux définitions, et fournissent des noms distincts pour décrire les deux types de classes : MICROCLASSES et MACROCLASSES (Dressler, Mayerthaler et al. 1987 ; Dressler et Thornton 1996 ; Kilani-Schoch et Dressler 2005) : Les MICROCLASSES sont de petites classes uniformes dont les membres se comportent de façon identiques. Il existe autant de microclasses qu'il y a de paradigmes types. Les MACROCLASSES sont de grandes classes fondées sur la similarité qui présentent un certain degré de variation interne.

Dressler et Thornton (1996) appellent également « paradigme isolé » une microclasse ne comportant qu'un lexème. Nous prendrons plutôt le parti de les voir comme un type (extrême) de microclasse. Ainsi, les microclasses correspondent aux définitions d'Aronoff et de Carstairs-McCarthy. Comme différents travaux utilisent des termes variables pour désigner microclasses et macroclasses, le tableau 1.6 résume ces usages.

nous	microclasses	macroclasses
Corbett (1982)	*types*	*paradigms*
Carstairs (1987)	*paradigm*	*macroparadigm*
Dressler et Thornton (1996)	*microclasses* et paradigmes isolés	*macroclasses*
Blevins (2006)	*declensions*	*macrodeclensions*
Brown et Hippisley (2012)	*class*	*superclass*
Bonami et Boyé (2014)	classes flexionnelles élémentaires	classes flexionelles irréductibles

TABLEAU 1.6 – Terminologie utilisée pour décrire les micro et macroclasses.

23. [En anglais dans le texte] « *inflectional classes [...] are classes of lexemes that go together in respect of some inflection* ».

24. [En anglais dans le texte] « *Lexemes that inflect alike are assigned to a common conjugation or declension; i.e., inflection class* ».

1.3.3 Les microclasses

L'identification d'un système de microclasses se fait de façon transparente si l'on dispose d'un inventaire des propriétés flexionnelles de chaque lexème : il suffit alors de réunir les lexèmes dont toutes les propriétés sont identiques. Cependant l'extraction de cet inventaire est un problème difficile en soi.

Comment caractériser le comportement flexionnel d'un lexème dans une approche Mot et Paradigme ? Ces modèles se fondent traditionnellement sur des ANALOGIES PROPORTIONNELLES qui réfèrent implicitement au comportement flexionnel des lexèmes, comme le décrit Matthews (1991, p. 192) :

> Tel est le génitif singulier *domini* au nominatif singulier *dominus*, tel doit être x (l'inconnue) au nominatif singulier *servus*. Quel est donc x ? Réponse : ce doit être *servi*. En notation, *dominus* : *domini* = *servus* : *servi* . Le patron tient pour de nombreux autres noms de ce qui est traditionnellement appelé la seconde déclinaison [25].

La question des microclasses est donc la suivante : comment évaluer le domaine dans lequel ces égalités sont valides ? Pour cela, il nous faut évaluer la relation entre *dominus* : domini et rosa : rosae. Il nous faut donc exprimer explicitement la fonction qui les relie, c'est à dire qu'il nous faut trouver f tel que $f(\text{dominus}) = \text{domini}$, g tel que $f(\text{rosa}) = \text{rosae}$, puis nous demander si $f = g$. Matthews (1991, p. 192) suggère d'écrire les fonctions telles que f et g, qu'il nomme « transformations morphologiques », comme indiqué ci-dessous (1) :

$$(1) \quad \textit{dominus} : \textit{domini} = \begin{bmatrix} \text{Nominative}, \\ \text{Singular}, \\ \text{X+us} \end{bmatrix}_N \rightarrow \begin{bmatrix} \text{Genitive}, \\ \text{X+i} \end{bmatrix}$$

Ces fonctions sont utilisées dans des IMPLICATIONS PARADIGMATIQUES. Dans les termes de Bonami (2014, p. 88) :

25. [En anglais dans le texte] « *As Genitive Singular* domini *is to Nominative Singular* dominus, *so* x *(unknown) must be to Nominative Singular* servus. *What then is* x *? Answer : it must be* servi. *In notation,* dominus : domini = servus : servi. *The pattern holds for many other Nouns of what is traditionally called the 2nd Declension* ».

> « a. Pour tout lexème L qui vérifie la condition C, si la case φ du paradigme de L est réalisée par la forme X , alors la case ψ du paradigme est réalisée par la forme $f(X)$.
> b. $C : [\varphi : X \Rightarrow \psi : f(X)]$ »

Nous proposons d'expliciter les fonctions f par des PATRONS D'ALTERNANCE, comme exemplifié en (2). Ce patron peut se lire : *Dans une forme du nominatif singulier, /-us/ final précédé d'une consonne alterne au génitif singulier avec /i/*. Les patrons décrivent des alternances qui sont *bidirectionnelles*, car les analogies représentent des relations qui peuvent se lire dans les deux sens. Ils sont constitués d'une ALTERNANCE formelle entre deux éléments (_us, _i) et d'un CONTEXTE D'APPLICATION qui formule des contraintes phonologiques à l'application du patron (XC_). On peut dire d'un patron qu'il est *applicable* à une forme si la forme remplit la contrainte formée par le contexte combiné par le membre approprié de l'alternance. Cette notation s'inspire des règles phonologiques de Chomsky et Halle (1968), ainsi que des patrons morphologiques de Albright et Hayes (2003) et Bonami et Boyé (2014).

(2) NOM.SG⇌GEN.SG : _us ⇌ _i / XC_

Le patron (2) exprime deux implications : d'une part, *tout lexème nominal dont la forme de nominatif singulier se termine par /-us/ précédé d'une consonne se termine en /-i/ au génitif singulier*, et inversement, *Tout lexème nominal dont la forme de génitif singulier se termine par /-i/ précédé d'une consonne se termine en /-us/ au génitif singulier*. Il n'est pas nécessaire que ces implications soient toujours satisfaites : on s'intéressera (chapitre 3) à leur probabilité d'être satisfaites. Elles constituent donc une forme de règle analogique plutôt qu'une implication logique. Guzman Naranjo (2017) remarque que dans cette perspective, « les analogies proportionnelles peuvent (et doivent) être étendues à des ensembles. Par exemple, ce n'est pas seulement la relation škola-školu qui détermine la relation muščina-muščinu, mais c'est plutôt l'ensemble de toutes les paires nominatif-accusatif connues par les locuteurs [26] ».

26. [En anglais dans le texte] « *proportional analogy can (and should) be extended to sets. For example, it is not just the relation škola-školu which determines the relation muščina-muščinu, it is rather the whole set of nominative-accusative pairs speakers know* ».

Nous proposons de qualifier le comportement flexionnel d'un lexème non pas par une série d'exposants, mais par l'ensemble des patrons d'alternances qui peuvent être définis pour l'ensemble des cases de son paradigme abstrait.

Wurzel (1989), Ackerman et Malouf (2013) et Stump et Finkel (2013) examinent les relations implicatives entre cases de paradigmes en s'appuyant sur des exposants préalablement segmentés. Mais pour les besoins de l'évaluation ultérieure des points de similarités entre paradigmes, il est crucial que ces patrons soient élaborés sur la base des mots entiers, plutôt que d'une segmentation préalable en affixes. Ainsi, des patrons fondés sur les affixes présentés dans le tableau 1.2 des noms latins présumeraient entre ACC.PL et GEN.PL le patron (3) pour l'alternance *dominōs* : *dominōrum* et le patron (4) pour l'alternance *rosās* : *rosārum* :

(3) ACC.PL⇌GEN.PL : _ōs ⇌ _ōrum / X_

(4) ACC.PL⇌GEN.PL : _ās ⇌ _ārum / X_

Il semble pourtant fallacieux de considérer que ROSA et DOMINUS ont des comportements distincts pour cette alternance. Un patron qui rend mieux compte des alternances entre mots pourrait décrire les deux alternances d'une même façon (5) :

(5) ACC.PL⇌GEN.PL : _s ⇌ _rum / XV_

Les analyses affixales manquent souvent des points de similarité entre mots formes dont les patrons d'alternances rendent compte naturellement. Prenons pour exemple quelques formes adjectivales du français, présentées dans le tableau 1.7.

Lexèmes	M.SG	F.SG/PL	M.PL
NORMAL	/nɔʁmal/	/nɔʁmal/	/nɔʁmo/
VERT	/vɛʁ/	/vɛʁt/	/vɛʁ/
BLEU	/blø/	/blø/	/blø/

TABLEAU 1.7 – Trois paradigmes adjectivaux français.

Nous présentons une analyse affixale de ces adjectifs dans le tableau 1.8. Celle-ci procède à un découpage global du paradigme, en retirant de chaque ligne la sous-chaîne commune à

chaque forme, c'est-à-dire le radical (Beniamine, Bonami et Sagot 2017). Cette segmentation identifie deux points de similarité entre les paradigmes de VERT et de BLEU, qui portent chacun un affixe zéro au masculin singulier et pluriel. Elle n'identifie aucun autre point de similarité entre les paradigmes.

Lexèmes	M.SG	F.SG/PL	M.PL
NORMAL	/-al/	/-al/	/-o/
VERT	—	/-t/	—
BLEU	—	—	—

TABLEAU 1.8 – Analyse affixale de trois paradigmes adjectivaux français.

Une analyse en patrons d'alternance se fonde au contraire sur des contrastes entre paires de formes, et produit donc une segmentation locale à cette paire de formes. Nous présentons dans le tableau 1.9 le résultat d'une telle analyse sur les formes du tableau 1.7. Nous notons « $\epsilon \rightleftharpoons \epsilon$ » le patron identité, et plus généralement employons le symbole ϵ pour dénoter la chaîne vide dans les patrons d'alternance. Ici, les patrons d'alternance identifient non seulement une similarité entre VERT et BLEU (avoir des formes identiques au masculin singulier et pluriel), mais également une similarité entre BLEU et NORMAL, que la segmentation affixale ignorait : tous deux présentent des formes identiques au masculin singulier et au féminin. Il existe donc des généralités dont l'analyse affixale ne rend pas compte, au contraire de l'analyse en patrons d'alternances. Il est notable que la grammaire traditionnelle du français ne traite justement pas l'alternance /-al//-o/ comme affixale.

Lexèmes	M.SG ∼ F.SG/PL	M.SG ∼ M.PL	F.SG/PL ∼ M.PL
NORMAL	$\epsilon \rightleftharpoons \epsilon$	_al ⇌ _o	_al ⇌ _o
VERT	_ ⇌ _t	$\epsilon \rightleftharpoons \epsilon$	_t ⇌ _
BLEU	$\epsilon \rightleftharpoons \epsilon$	$\epsilon \rightleftharpoons \epsilon$	$\epsilon \rightleftharpoons \epsilon$

TABLEAU 1.9 – patrons d'alternances binaires pour trois paradigmes adjectivaux français.

Ces différences ont des conséquences directes sur les structures des classes flexionnelles,

définie en termes de similarité. Les affixes sont des unités abstraites, dont le calcul nécessite une connaissance globale de l'ensemble des paradigmes d'un système. Les patrons constituent des unités de plus bas niveau, et ne sont pas des façons alternatives de noter des affixes. Ils manifestent l'ensemble des segmentations pertinentes pour chaque paire de cases. Par ailleurs, ils ne postulent pas d'allomorphie phonologiquement conditionnée. Ils exposent donc tous les contrastes de surface observables à travers les formes d'un même lexème, et ne masquent pas les similarités qui ne peuvent s'exprimer en termes d'affixes. En somme, ils constituent une meilleure base pour l'évaluation de la similarité entre paradigmes, et donc pour l'analyse en classes flexionnelles.

Le chapitre 2 sera consacré à l'inférence automatique non supervisée de tels patrons d'alternance à partir de tables de paradigmes contenant des mots formes en notation phonologique, et à l'étude des systèmes de microclasses qui en découlent. Ce faisant, nous prenons au mot Blevins (2006), qui écrit des classes flexionnelles : « elles contiennent des items qui présentent des patrons flexionnels communs [27] ».

1.3.4 Les macroclasses

Contrairement aux microclasses, les macroclasses ne peuvent se déduire uniquement d'une analyse en patrons d'alternances. Elles sont fondées sur un certain degré de similarité. Or la similarité est gradiente et multidimensionnelle. Afin de décider d'un seuil de similarité, il est généralement nécessaire de faire entrer en compte d'autres critères de décision. Nous en examinons ici quelques-uns.

Nous avons vu que les décisions de Carstairs (1987) pour réunir des microclasses reposaient sur trois principes : les microclasses d'un macroparadigme doivent présenter des similarités, elles ne doivent pas être en compétition (il existe un prédicteur permettant de les distinguer), et leur réunion mène à un système désirable du point de vue du NBP. Nous avons également discuté les objections de Blevins (2004) aux découpages en faveur du NBP qu'il compare au « charcutage électoral ». Le critère de prédictibilité seul ne permet donc pas de décider d'une unique partition de classes.

27. [En anglais dans le texte] « *they contain items that exhibit common patterns of inflection* ».

En général, les morphologues descriptifs motivent les classifications en macroclasses en fonction de propriétés discriminantes entre macroclasses. Le tableau 1.10 présente un extrait de paradigme de cinq verbes exemplaires du latin.

conjugaison	lexème	glose	INF	PRS.1SG	PRS.2SG
première	AMARE	'aimer'	amāre	amō	amās
seconde	TENERE	'tenir'	tenēre	teneō	tenēs
troisième	CURRERE	'courrir'	currere	currō	curris
troisième	CAPERE	'prendre'	capere	capiō	capis
quatrième	AUDIRE	'entendre'	audīre	audiō	audīs

TABLEAU 1.10 – Extraits de paradigmes verbaux latins.

Les conjugaisons latines sont caractérisées par la qualité et la longueur de la voyelle thématique de l'infinitif présent actif : -ā- pour la première conjugaison, -ē- pour la seconde, -e- pour la troisième et -ī- pour la quatrième. Cependant, au sein de chaque conjugaison, il existe de nombreuses façon distinctes de former le supin. Les conjugaison partagent également des points de similarités : Certains verbes de la troisième conjugaison comme CURRERE ont un indicatif présent première personne actif en -ō, similaire à ceux de la première conjugaison. D'autres, comme CAPERE, sont distincts de la première conjugaison pour cette case, mais identiques à la quatrième conjugaison. La tradition considère que le contraste à l'infinitif est plus important que celui de la première personne du singulier au présent, ou que ceux du supin.

Une telle stratégie pour motiver une classification en macroclasses présente deux problèmes. D'une part, il n'est pas facile de savoir si les propriétés fondent la classification, ou si elles sont sélectionnées *post-hoc* afin de contraster des classes pré-établies (par exemple pour des besoins pédagogiques). D'autre part, les raisons pour lesquelles certaines distinctions sont tenues pour prioritaires ne sont pas explicitées, et il est donc impossible de savoir leur nature.

Les études des systèmes de classes flexionnelles dans le cadre de la « Morphologie Naturelle » (Dressler et Thornton 1996 ; Kilani-Schoch et Dressler 2005 ; Dressler, Kilani-Schoch et

al. 2008), de par leur caractère particulièrement explicite, permettent de mettre le doigt sur les arbitrages difficiles à motiver entre critères dans l'établissement des macroclasses.

Pour les tenants de la « Morphologie Naturelle », les classes, et donc entre autres les macroclasses, sont définies par des implications nommées « conditions de structure paradigmatique [28] ». Nous présentons ci-dessous quelques conditions de structure paradigmatique établies par Kilani-Schoch et Dressler (2005) pour les classes verbales du français :

(6) Macroclasse I : Infinitif /X+e/ ⇒
- Participe passé = /X+e/
- Passé simple première personne du singuliers = /X+e/
- Présent singulier = /X/
- Présent indicatif troisième personne du pluriel = /X/
- Subjonctif présent = /X/

(7) Classe I.1 : Imparfait parl+ɛ, futur parl+ər+e.

(8) Classe II.2 : Infinitif /Xwar/ ⇒
- Participe passé en /y/
- Passé simple en /y/
- par défaut, /wa/ fait partie de l'infinitif

Remarquons tout d'abord que les conditions de structure paradigmatique sont de nature variable. Elles sont parfois formulées en terme de relations implicatives (Wurzel 1989 ; Ackerman, Blevins et Malouf 2009 ; Stump et Finkel 2013), comme c'est le cas pour la macroclasse I dont la condition de structure paradigmatique est reformulée en (6) ou pour la classe II.2, comme décrit en (8). Les implications sont parfois des relations entre les segmentations de deux cases, comme en (6), et parfois entre une case et une unité abstraite, comme en 8. Certaines classes, au contraire, sont définies par des exposants, comme c'est le cas en 7 pour la classe I.1.

Un autre principe organisateur est présent dans la classification verbale du français de Kilani-Schoch et Dressler (2005). Leur modèle de la flexion est fondé sur un double mécanisme (Clahsen 2006) et suppose qu'il existe une différence catégorique entre lexèmes réguliers et irréguliers, et que les lexèmes réguliers et irréguliers sont traités différemment par les locuteurs. Kilani-Schoch et Dressler (2005) considèrent que la bipartition entre réguliers et irréguliers est

28. [En anglais dans le texte] « *Paradigm Structure Conditions* ».

un critère décisif pour distinguer les macroclasses. Ils définissent donc deux macroclasses : la première regroupe l'ensemble des verbes productifs, caractérisés par un infinitif en /-e/. Elle correspond à la première conjugaison traditionnelle du français. La seconde réunit tous les autres verbes.

La notion de régularité est souvent employée pour justifier des classifications en macroclasses. Wurzel (1989, p. 64) propose une liste de propriétés permettant de juger de la « normalité ». Les classes normales, ou régulières, sont :

— des attracteurs pour les classes moins régulières,
— productives, tant pour les néologismes que les pseudo-mots,
— attractives, donc plus souvent sur-généralisées, et les erreurs résultantes semblent plus acceptables aux locuteurs que des erreurs résultant de la sur-généralisations de classes moins normales,
— moins touchées par les troubles aphasiques,
— plus rapidement acquises par les enfants et les apprenants en général.

La régularité d'un lexème, d'après ces critères, s'observe non en regardant la structure implicative d'un système flexionnel en synchronie, mais à travers l'étude diachronique et des expériences psycholinguistiques. Quelle que soit la position adoptée concernant les modèles à double mécanisme, ou la supposition que seule la première conjugaison traditionnelle des verbes français est productive (voir Bonami, Boyé, Giraudo et al. (2008)), il nous faut donc conclure que ce type de classification est entièrement distinct, et potentiellement orthogonal, à une classification fondée sur la similarité des patrons flexionnels.

La régularité d'une classe flexionnelle peut aussi être vue comme corrélée avec la fréquence de type (les classes plus grandes sont plus régulières), la simplicité formelle (notion elle-même floue et difficile à mesurer), et l'extensibilité à des items inconnus dans les expériences d'élicitation. Notons que le trio productivité, attractivité et extensibilité devraient corréler au moins en partie avec la généralité des contextes d'application d'un patron flexionnel. Le chapitre 5 explorera une façon de fonder l'inférence de macroclasses sur les similarités structurelles des comportements flexionnels des lexèmes. Nous évaluerons également comment ces classes sont distribuées en termes de généralité et de fréquence de type des patrons.

1.4 La canonicité des systèmes de classes flexionnelles

Au-delà de la question d'une partition remarquable du lexique en macroclasses, on peut s'intéresser à la structure des similarités entre les classes. Ainsi, Dressler et Thornton (1996), Kilani-Schoch et Dressler (2005) et Dressler, Kilani-Schoch et al. (2008) intègrent les macroclasses et microclasses dans un arbre dans lequel tout nœud peut être considéré comme une classe flexionnelle. Chacun de ces nœud hérite des propriétés de tous ses ancêtres selon un principe d'héritage monotone.

Corbett et Fraser (1993) et Brown et Hippisley (2012) proposent une théorie de la morphologie nommée Morphologie en Réseau (*Network morphology*) qui représente l'ensemble du système flexionnel par une arborescence d'héritage par défaut. Ces analyses sont de type constructif, et permettent de générer les formes d'un lexème en parcourant le chemin qui le relie à la racine. L'héritage par défaut a deux avantages dans ce contexte : d'une part il permet des représentations compactes en limitant les répétitions et le nombre de nœud dans la hiérarchie, et d'autre part il donne un statut naturel à la notion de régularité : un nœud qui réécrit un défaut fait exception à la règle définie par son parent.

Si partitions et arborescences peuvent être considérées comme des modèles concurrents pour représenter les systèmes de classes flexionnelles, on peut également voir les partitions comme un cas limite maximalement simple de classes flexionnelles : un système de classes formerait une véritable partition si les microclasses ne présentent aucune similarité entre elles.

Wurzel (1989, p. 63) décrit comme suit les propriétés d'un système de classes idéalement uniformes et indépendantes :

> Nous avons supposé qu'il existe une classe flexionnelle indépendante et uniforme dans une langue si chaque catégorie ou ensemble de catégorie dérivés sont formellement réalisés d'une façon uniforme pour un certain groupe de mots, et si les formes flexionnelles dérivées sont toutes formellement distinctes des formes dérivées pour tous les autres groupes de mots.
>
> La constitution des classes flexionnelles se fonde sur l'uniformité et la distinctivité des paradigmes, exactement comme chaque classification se fonde sur les propriété-

tés communes et distinctes des éléments à classer. [29]

Corbett (2009) systématise cette description, et fournit deux principes et neuf critères caractérisant les systèmes de classes flexionnelles canoniques, qui présentent :

(principe I) des classes **distinctives**, c'est à dire qu'ils présentent

(critère 1) à travers les classes, des paradigmes concrets distincts ;

(critère 2) mais des paradigmes abstraits identiques ;

(critère 3) à l'intérieur d'une classe, des paradigmes concrets identiques ;

(critère 4) une structure implicative plate, où aucune case n'est plus prédictive qu'une autre.

(principe II) des classes **indépendantes**, c'est à dire que « la distribution des items lexicaux à travers les classes flexionnelles canoniques n'est pas motivée en synchronie [30] ». En conséquence,

(critère 5) chaque classe comporte un nombre égal de lexèmes ;

(critère 6) les classes ne sont pas motivées phonologiquement ;

(critère 7) les classes ne sont pas motivées syntaxiquement ;

(critère 8) les classes ne sont pas motivées par la catégorie morphosyntaxique ;

(critère 9) les classes ne sont pas motivées pragmatiquement.

D'après cette description, les classes flexionnelles constituent des sous-systèmes entièrement distincts et indépendants. Corbett (2009) décrit le système du Burmeso (tableau 1.5) comme un rare cas de canonicité. Un système de classes canonique respecte parfaitement le principe d'économie du paradigme, tout comme le NBP, puisqu'il ne présente aucun exposant partagé entre classes. Ces critères ne distinguent pas entre micro- et macro-classes, puisqu'elles

29. [En anglais dans le texte] « *We assumed that there exists a uniform and independent inflectional class in a language if every derived category and/or bundle of categories is formally symbolized in a uniform way for some group of words and if the derived inflectional forms are all formally distinct from the derived inflectional forms of all other word groups.*

The constitution of inflectional classes is based on the uniformity and distinctiveness of paradigms, just as every classification is based on the common and distinct properties of the elements to be classified ».

30. [En anglais dans le texte] « *the distribution of lexical items over canonical inflectional classes is synchronically unmotivated* ».

sont identiques dans le cas canonique (critère 1 et 3). L'existence de similarités entre classes flexionnelles enfreint le critère 1 par définition. Un lexème est DÉFECTIF pour une case de paradigme s'il n'a pas de forme définie pour cette case. Son paradigme abstrait est donc distinct des paradigmes abstraits des autres lexèmes du système. À l'opposé, un lexème est SURABONDANT pour une case de paradigme s'il présente plus d'une forme pour cette case. Ces deux phénomènes enfreignent le critère 2. L'existence de sous-classes au sein de macroclasses enfreint le critère 3, car ces sous-classes dévient les unes des autres quoi qu'elles appartiennent à une même macroclasse. On parle d'HÉTÉROCLISE lorsqu'une classe présente des réalisations provenant de deux autres classes plus grandes. L'hétéroclise enfreint le critère 1 en raison de la similarité entre la classe hétéroclite et les deux autres classes, mais également le critère 5, puisqu'il s'agit d'une classe plus petite. De même, l'existence de paradigmes isolés au sens de Kilani-Schoch et Dressler (2005) ou de classes majoritaires enfreint le critère 5. Les critères 6 à 9 concernent les motivations synchroniques entres classes. Étudier la canonicité d'un système de classes flexionnelles peut donc consister, entre autres, à mesurer à quel point il s'écarte d'un système de partition. Le chapitre 6 propose une façon de quantifier cet écart de façon quantitative.

1.5 Conclusion

Dans ce chapitre, nous avons vu que les inventaires de flexionnelles pour un même système varient d'un auteur à l'autre. Nous avons remarqué qu'il existe une relation circulaire entre classes flexionnelles et segmentations affixales : les classes flexionnelles sont définies sur la base des affixes employés par chaque lexème, mais la segmentation en affixes nécessite de connaître les classes flexionnelles. Par ailleurs, dans une perspective constructive, les classes flexionnelles semblent le lieu d'une complexité gratuite qu'il est opportun de réduire au maximum. Il est donc fréquent de choisir une segmentation spécifique en vertu de la désirabilité des classes auxquelles elle mène. Pour sortir de cette circularité, nous avons proposé de renoncer à la segmentation en affixes et radicaux, et d'observer l'exponence au sein des paradigmes à travers les contrastes morphologiques manifestés par les paires de formes. Le chapitre 2 fournit

une méthode permettant d'identifier automatiquement les contrastes flexionnels au sein des paradigmes. Ces contrastes, exprimés sous la forme de patrons d'alternance, constituent une caractérisation du comportement flexionnel des lexèmes. Le chapitre 3 étudie la structure des paradigmes à travers les relations de prédictibilité fondées sur ces patrons d'alternance.

Par ailleurs, nous avons vu qu'il existe deux sens de la notion de classe flexionnelle : D'une part Kilani-Schoch et Dressler (2005) nomment microclasses des ensembles de lexèmes qui ont exactement le même comportement flexionnel. Cette notion est simple à opérationnaliser une fois que l'on dispose d'une caractérisation du comportement flexionnel des lexèmes. Le chapitre 4 présente une étude de la structure de similarité des microclasses fondées sur les patrons d'alternances. D'autre part, Kilani-Schoch et Dressler (2005) nomment macroclasses la notion qui correspond aux classes des descriptions traditionnelles. Celles-ci présentent de grandes classes peu nombreuses fondées sur la similarité des comportements entre lexèmes. L'inférence de macroclasses pose problème car la similarité est une notion continue. Comment évaluer le degré de similarité qui fonde une classe ? Le chapitre 5 propose de se fonder sur une mesure de longueur de description pour décider d'une partition optimale des microclasses en macroclasses. Nous nous y demandons à quel point les macroclasses fournissent une image fidèle des systèmes flexionnels. Enfin, nous montrons dans le chapitre 6 que la structure qui relie les microclasses est plus fidèlement représentée par une hiérarchie à héritage multiples que par une partition de macroclasse ou une structure arborée.

Première partie

Les alternances flexionnelles

Chapitre 2

Des formes aux patrons d'alternances

Le but de cette thèse est de classer les lexèmes selon leur comportement flexionnel. La première étape pour ce faire est de fournir une définition précise de leur comportement flexionnel, d'une façon qui se prête à l'observation des similarités entre ces comportements. Blevins (2006, p. 537) décrit ainsi la façon dont les modèles morphologiques fondés sur le mot décrivent le comportement flexionnel des lexèmes : « Contrairement à de nombreuses approches contemporaines, les modèles traditionnels n'imposent pas une séparation stricte entre "données" et "patrons", mais représentent les patrons morphologiques d'une langue par de véritables formes qui présentent ces patrons [1] ». Cette approche laisse cependant le lecteur intuiter lui-même l'abstraction des patrons à partir des formes. L'énumération des formes de surface brutes ne fournit pas en elle-même des unités permettant de fonder une classification, chaque paradigme de surface étant par définition distinct de tous les autres. Il nous faut donc extraire des paradigmes un ensemble d'abstractions et souscrire à une séparation stricte entre les données (décrites dans l'annexe A) et les représentations flexionnelles. Ce chapitre décrit l'inférence de ces représentations sous la forme de patrons d'alternance.

Nous avons discuté dans le chapitre 1 des raisons pour lesquelles une analyse en patrons d'alternance est préférable à une analyse affixale pour l'étude de la similarité dans les systèmes flexionnels. Dans la section 2.1, nous montrons qu'il est en fait impossible de fonder de façon satisfaisante une segmentation automatique des formes en radicaux et exposants. Dans la sec-

1. [En anglais dans le texte] « *Unlike many contemporary approaches, traditional models do not impose a radical separation of 'data' and 'patterns', but represent the morphological patterns of a language by actual forms that display those patterns* ».

tion 2.2, nous présentons des modèles computationnels de la flexion dont nous nous inspirons : d'une part les modèles de réinflexion, d'autre part le *Minimal Generalization Learner* d'Albright et Hayes (2002), et les modèles qui s'en inspirent. La section 2.3 présente notre modèle d'inférence automatique de patrons d'alternance. Enfin, la section 2.4 présente son application aux données des lexiques étudiés dans cette thèse.

2.1 Pour en finir avec la segmentation en radicaux et exposants

Nous avons montré (section 1.3.3) qu'il était préférable de poser les questions de prédictibilité et de complexité dans les paradigmes flexionnels à l'échelle du mot plutôt qu'à celle du morphème. Pour déterminer les alternances entre formes, il semble intuitif de se reposer sur une segmentation catégorique des formes en unités plus petites, radicaux et affixes, comme le font entre autres Stump et Finkel (2013). Nous avons discuté dans la section 1.3.3 du chapitre 1 du fait que cela conduit à sous-estimer les similarités entre le comportement flexionnel des lexèmes. Mais le problème de ces segmentations est bien plus considérable. Nous montrons ici que la segmentation de formes en radicaux et affixes est bien plus complexe qu'il n'y paraît, et argumentons que le problème de l'exponence catégorique ne peut trouver de réponse satisfaisante.

2.1.1 Le problème technique de la Segmentation

Les segmentations en affixes et radicaux des formes flexionnelles paraissent souvent intuitives. À partir de formes de surface, il semble au premier abord suffire de retirer de chaque forme son radical pour obtenir un paradigme d'affixes sur lequel fonder une théorie des classes flexionnelles : c'est la stratégie de Carstairs (1987), pour qui le comportement flexionnel d'un lexème est manifesté par un paradigme d'affixes.

On trouve rarement de description détaillée du processus nécessaire à l'inférence de ces segmentations. Comme nous l'avons mentionné dans la sous-section 1.2.4, il n'existe pas de procédure communément admise permettant d'inférer systématiquement des affixes à partir des paradigmes de formes, sans préconception sur les données, et de façon universelle, c'est à

dire applicables à toute langue sans préconception sur la nature des phénomènes morphophonologiques mis en jeu. C'est ce que Spencer (2012) nomme le Problème de la Segmentation :

> **Problème de la Segmentation** :
> En général, pour n'importe quel mot complexe de n'importe quelle langue, il n'y a (apparemment) aucun moyen d'établir une segmentation de façon algorithmique. [2]

Spencer identifie en fait deux problèmes distincts : d'une part, il s'agit d'un sujet peu étudié en lui-même ; d'autre part « lorsque l'on regarde la pratique des grammairiens, il apparaît clairement d'un examen superficiel des descriptions de langues morphologiquement complexes familières qu'il n'y a aucun consensus sur la segmentation même pour des langues extrêmement bien étudiées [3] ». En effet, les descriptions de segmentations élaborées manuellement par les linguistes sont le fruit de séries d'arbitrages qui varient entre auteurs, entre traditions théoriques, et d'une langue à l'autre.

Il existe bien des outils de segmentation non supervisée de formes de surface. Ceux-ci se fondent généralement sur l'optimisation de longueurs de description afin de fournir une segmentation économique des formes. C'est le cas de Linguistica (Lee et Goldsmith 2016) ou de certains programmes de la famille Morfessor (Virpioja et al. 2013 ; Grönroos et al. 2014). Cependant ces programmes ne sont pas capables d'inférer des segmentations spécifiquement flexionnelles, d'associer les segmentations opérées à des valeurs morphosyntaxiques (c'est à dire aux cases), de tirer parti de la structure paradigmatique, ou d'extraire des morphèmes discontinus. Ils ne permettent donc pas de segmenter des formes de paradigmes en radicaux et affixes, et ne sont pas utiles en présence de morphologie non-concaténative (Hammarström et Borin 2011, p. 68).

Afin de nous convaincre plus avant que le problème de segmentation ne peut être résolu simplement et que celle-ci ne permet pas l'identification des paradigmes types, nous proposons

2. [En anglais dans le texte] « ***Segmentation Problem***:
In general, for any complex word in any language, there is (apparently) no way to establish a segmentation algorithmically ».

3. [En anglais dans le texte] « *when we look at the practice of grammarians, it should be clear from a cursory glance through descriptions of familiar morphologically complex languages that there is no consensus on segmentation even for extremely well-studied languages* ».

d'examiner un exemple concret. Nous appliquons à la lettre deux descriptions du processus de segmentation, l'une portant sur les radicaux, l'autre sur les affixes, au petit ensemble de lexèmes nominaux du latin déjà présentés.

Kuryłowicz (1945, p. 220) propose d'extraire les radicaux, qu'il nomme *thèmes* par identification de la plus longue sous-chaîne commune à l'ensemble du paradigme d'un lexème :

> On trouve le thème en dégageant les éléments communs à toutes les formes casuelles du paradigme (quand il s'agit de la déclinaison). P.e *lup-us, -i, -o, -um, -orum, -is, -os* fondent le thème *lup-*. Le paradigme russe *trud, -'a, -'u, -'om, etc.* permet de dégager un thème *trud'-*.

Boyé (2000, p. 125) définit quant à lui les affixes comme la plus longue sous-chaîne commune à l'ensemble des formes d'une case :

> Établissons simplement les affixes flexionnels en utilisant la partie commune à tous les verbes pour chacune des formes.

Ces deux définitions, réunies ensemble, ont l'élégance de la symétrie : l'affixe d'une case est le matériel phonétique invariable dans les formes de cette case, le radical d'un lexème est le matériel phonétique invariable dans les formes de ce lexème. Le tableau 2.1 présente le résultat de cette analyse sur nos quelques noms latins [4]. Les exposants sont présentés pour chaque case sous l'intitulé « exp »., les radicaux dans la colonne « radical ». Les colonnes « reste » présentent le matériel que cette procédure ne définit ni comme radical ni comme exposant.

Le caractère problématique de la segmentation résulte du sort qui sera fait à ce « reste ». Les analyses proposent typiquement de l'assimiler soit aux exposants (on a alors des classes flexionnelles), soit aux radicaux (on a alors de l'allomorphie radicale). Ce choix n'est cependant pas aisé.

Observons tout d'abord la colonne NOM.SG. L'ensemble des affixes familiers des noms du latin (-a, -us, etc.) figurent non dans la colonne dans exposants, mais dans celle du reste. En

4. Nous extrayons la plus longue sous-chaîne commune (« longest common substring », ou LCS), possiblement discontinue, à toutes les formes d'un lexème pour trouver le radical, et la LCS à toutes les formes d'une case pour trouver l'exposant. Cette procédure est donc à même de découvrir des exposants et radicaux discontinus.

	radical	NOM.SG		VOC.SG		ACC.SG		GEN.SG		DAT.SG		ABL.SG	
		reste	exp.	reste	exp.	reste	exp.	reste	exp.	reste	exp.	reste	exp.
ROSA	ros-	-a	—	-a	—	-am	—	-ae	—	-ae	—	-a:	—
DOMINUS	domin-	-us	—	-e	—	-um	—	-i:	—	-o:	—	-o:	—
AGER	ag-r	-e-	—	-e-	—	-um	—	-i:	—	-o:	—	-o:	—
TEMPLUM	templ-	-um	—	-um	—	-um	—	-i:	—	-o:	—	-o:	—
REX	re-	-x	—	-x	—	-gem	—	-gis	—	-gi:	—	-ge	—
CONSUL	consul	—	—	—	—	-em	—	-is	—	-i:	—	-e	—
CORPUS	corp-	-us	—	-us	—	-us	—	-is	—	-i:	—	-e	—
CIVIS	civ-	-is	—	-is	—	-em	—	-is	—	-i:	—	-e	—
URBS	urb-	-s	—	-s	—	-em	—	-is	—	-i:	—	-e	—
MARE	mar-	-e	—	-e	—	-e	—	-is	—	-i:	—	-i:	—
MANUS	manu-	-s	—	-s	—	-m	—	-:s	—	-i:	—	-:	—
CORNU	cornu-	-:	—	-:	—	-:	—	-:s	—	-:	—	-:	—

	radical	NOM.PL		VOC.PL		ACC.PL		GEN.PL		DAT.PL		ABL.PL	
		reste	exp.	reste	exp.	reste	exp.	reste	exp.	reste	exp.	reste	exp.
RES	re-	-:s	—	-:s	—	-m	—	-:i:	—	-:i:	—	-:	—
ROSA	ros-	-ae	—	-ae	—	-a:s	—	-a:r-	-um	-i:-	-s	-i:-	-s
DOMINUS	domin-	-i:	—	-i:	—	-o:s	—	-o:r-	-um	-i:-	-s	-i:-	-s
AGER	ag-r	-i:	—	-i:	—	-o:s	—	-ro:-	-um	-i:-	-s	-i:-	-s
TEMPLUM	templ-	-a	—	-a	—	-a	—	-o:r-	-um	-i:-	-s	-i:-	-s
REX	re-	-ge:s	—	-ge:s	—	-ge:s	—	-g-	-um	-gibu-	-s	-gibu-	-s
CONSUL	consul	-e:s	—	-e:s	—	-e:s	—	—	-um	-ibu-	-s	-ibu-	-s
CORPUS	corp-	-a	—	-a	—	-a	—	—	-um	-ibu-	-s	-ibu-	-s
CIVIS	civ-	-e:s	—	-e:s	—	-e:s	—	-i-	-um	-ibu-	-s	-ibu-	-s
URBS	urb-	-e:s	—	-e:s	—	-e:s	—	-i-	-um	-ibu-	-s	-ibu-	-s
MARE	mar-	-ia	—	-ia	—	-ia	—	-i-	-um	-ibu-	-s	-ibu-	-s
MANUS	manu-	-:s	—	-:s	—	-:s	—	-u-	-um	-ib-	-s	-ib-	-s
CORNU	cornu-	-a	—	-a	—	-a	—	-u-	-um	-ib-	-s	-ib-	-s
RES	re-	-:s	—	-:s	—	-:s	—	-:r-	-um	-:bu-	-s	-:bu-	-s

Pour des raisons de lisibilité, nous notons ici l'allongement par le symbole de l'API « : ».

TABLEAU 2.1 – Problèmes dans la segmentation en morphèmes : exemple des noms du latin.

effet, la stratégie de Boyé (2000) échoue dès lors qu'il existe des classes flexionnelles, car elle présuppose que les affixes sont uniformes à travers tous les lexèmes. Les seuls affixes trouvés sont le -um communs à tous les génitifs pluriels, le -s commun à tous les datifs et ablatifs pluriels.

Il faudrait donc partitionner préalablement les lexèmes en classes, afin que cette stratégie fonctionne. Hélas ! Pour partitionner les lexèmes en classes, il nous faudrait déjà connaître leur comportement flexionnel.

Dans cet exemple, la solution consistant à reverser le « reste » vers les affixes pourrait nous tenter. Mais cela produit une segmentation peu convaincante en présence d'allomorphie radicale : on obtient l'affixe -x pour le nominatif et le vocatif de rex, et un ensemble de suffixes en -g pour les autres cas. Le résultat serait du même ordre en présence de supplétion : on analyserait un radical zéro et l'ensemble du matériel serait présent dans le reste.

Spencer (2012) propose à l'inverse de favoriser les radicaux dans ce type d'arbitrages :

> Le second principe est la Maximisation du Radical. Chaque fois que nous sommes confrontés avec le choix de traiter une sous-chaîne comme un affixe flexionnel ou une partie d'un radical morphomique, nous ne devrions opter pour la solution affixale que si cette solution est sans équivoque et mène à une déclaration simple de l'exponence. Sinon, nous supposons que la sous-chaîne fait partie du radical. Une conséquence de ce principe est que nous devrions traiter tout fragment (c'est à dire, sous-chaîne) comme partie d'un radical morphomique dès lors qu'elle réapparaît ailleurs dans le paradigme avec une fonction différente. [5]

Dans nos données, pour le lexème ROSA, la plupart des « restes » sont répétés entre nominatif et vocatif : faut-il pour autant tous les considérer comme parties du radical ? Aux génitif, datif et ablatif pluriels, on trouve des affixes uniques communs à toutes les classes, -um, -s. Faut-il renoncer à ce que -s fasse partie de l'affixe parce qu'il est répété entre datif et ablatif pluriel ?

5. [En anglais dans le texte] « *The second principle is Stem Maximization. Whenever we are confronted with a choice as to whether to treat a substring as an inflectional affix or as part of a morphomic stem we should only opt for the inflectional affix solution if that solution is unequivocal and leads to a simple statement of exponence. Otherwise, we assume that the substring is part of the stem. A consequence of this principle is that we should treat any partial (that is, substring) as part of a morphomic stem whenever that partial recurs elsewhere in the paradigm with a different function* ».

Est-il préférable de conserver ces affixes communs, plutôt que les habituels -arum, -orum, etc., au prétexte que cela mène à une définition plus simple des affixes ? Il est difficile de déterminer dans quels cas la solution affixale est « simple » et « sans équivoque ».

Les alternances non concaténatives posent encore un autre problème. Que faire par exemple face aux alternances de longueur des voyelles dans notre exemple ? Dans le paradigme de RES, où l'on peut observer une alternance e/e:, notre méthode a conservé la voyelle dans le radical, ce qui mène à des affixes du type -:s. Ceux-ci semblent artificiels, pourtant la solution inverse, qui consiste à toujours identifier e et e: comme faisant partie des exposants, manque de toute évidence une information sur la récurrence de la voyelle e dans les formes ce lexème, dont le radical serait alors réduit à r-.

Le problème technique de la segmentation que mentionne Spencer (2012) n'est donc pas, contrairement à sa suggestion, seulement apparent. Il est bien sûr possible de prendre une fois pour toutes une décision entre maximisation du radical et maximisation des exposants. Les deux mènent cependant à des segmentations peu satisfaisantes, même sur un exemple aussi peu controversé que l'ensemble des lexèmes exemplaires du latin analysés ici. Est-il seulement possible de résoudre ce problème d'une façon satisfaisante ?

2.1.2 Le problème de la segmentation catégorique

Afin de répondre à cette question, il nous faut déterminer comment juger de la qualité d'une segmentation. Nous proposons d'appeler SEGMENTATION CATÉGORIQUE la segmentation des mots d'un paradigme flexionnel en deux sous-chaînes complémentaires, le RADICAL et l'AFFIXE, de telle façon à ce que le premier soit l'expression exclusive de l'identité du lexème et l'autre l'expression exclusive des traits morphosyntaxiques d'une case de paradigme. Une segmentation catégorique peut être jugée en fonction de sa conformité à cette définition : les radicaux doivent exprimer toute l'identité lexicale et seulement elle, et les affixes toute l'identité flexionnelle et seulement elle.

Mais il arrive qu'un même matériel marque en partie l'identité lexicale et en partie des valeurs morphosyntaxiques. C'est le cas des formes supplétives, qui sont spécifiques à la fois à une case (ou quelques cases) et à un lexème. Dans une moindre mesure, il en va de même des

allomorphes d'un radical : la présence d'un allomorphe dans une forme, même si sa distribution est parfaitement régulière, donne des informations permettant de reconnaître la case à laquelle elle appartient, puisque certaines cases ne présentent pas cet allomorphe. De façon symétrique, lorsqu'un système présente des classes flexionnelles, les marqueurs de ces classes portent une valeur lexicale, car ils ne sont communs qu'à certains lexèmes. Plus la classe de ces lexèmes est petite, et plus ces marqueurs portent d'information lexicale.

Les éléments dits *thématiques* sont un cas par excellence de matériel ambigu : dire que le radical de ROSA n'inclut pas la voyelle thématique /a/, c'est ignorer que dans dix cas sur douze, cette voyelle est présente dans les formes de ce lexème, et qu'elle ne peut donc que lui être fortement associée. Cependant, dire qu'elle appartient aux radicaux revient à ignorer le fait que sa présence, comme les contrastes de longueur, contribue à l'expression casuelle. Il en va de même pour les verbes français du second groupe qui se caractérisent par un /s/ final à certaines personnes. Celui-ci fait partie de l'expression spécifique de ces lexèmes : en effet, eux seuls présentent une alternance du type fini/finis-, et celle-ci est systématique. L'identifier comme une allomorphie radicale permet d'analyser des affixes qui sont communs à la plupart des autres verbes du français (Boyé 2000 ; Bonami et Boyé 2003b). Pourtant, il est certain également que la présence de ce /s/ donne une indication sur la case du paradigme, puisqu'il n'est présent que dans 23 cases sur une cinquantaine. Par ailleurs, faire de cet élément une partie des affixes correspondants produit une analyse sans allomorphie radicale.

Radicaux et affixes fonctionnent comme des vases communicants pour l'analyse des paradigmes de flexion. Le matériel phonétique des formes fléchies n'exprime pas, en général, exclusivement du contenu lexical ou flexionnel. Il n'est pas possible de partager les mots formes en deux signes étanches et indépendants alors que l'on sait que leur matériel phonétique peut exprimer de façon simultanée des valeurs lexicales et flexionnelles. Il nous faut donc nous rendre à l'évidence que l'exponence étant de nature gradiente, il n'existe aucune réponse satisfaisante à la question de l'exponence catégorique.

Cette conclusion nous conduit à abandonner la segmentation en radicaux et exposants comme mode de description des comportements flexionnels.

2.2 Modèles computationnels de la flexion

Dans cette section, nous examinons l'état de l'art des modèles computationnels non affixaux de la flexion, dans le but d'identifier des stratégies pertinentes pour l'inférence automatique des comportements flexionnels. Plus précisément, nous décrivons ici les approches proposées par deux types de systèmes. D'une part, en traitement automatique des langues (TAL), la tâche dite de réinflexion, qui consiste à produire une forme fléchie à partir d'une autre forme déjà fléchie, est aujourd'hui très bien réussie par un ensemble de modèles. D'autre part, nous présentons un ensemble de modèles analogiques de la flexion dont le plus influent est le *Minimum Generalisation Learner* (MGL), et qui se fondent sur des patrons d'alternance.

2.2.1 La tâche de réinflexion

L'extraction de lexiques à partir de corpus produit des paradigmes incomplets en raison de la distribution zipfienne des formes fléchies des lexèmes (Blevins, Milin et Ramscar 2017). Afin de compléter ces lexiques, des efforts se sont tournés vers l'inférence des formes manquantes à partir des données déjà extraites. Il est en effet possible de compléter ces paradigmes automatiquement en se fondant sur la structure implicative des paradigmes pour tirer des inférences fiables, similairement à ce que font les locuteurs (Durrett et DeNero 2013 ; Ahlberg, Forsberg et Hulden 2014 ; Ahlberg, Forsberg et Hulden 2015 ; Nicolai, Cherry et Kondrak 2015). Ces travaux répondent donc à une manifestation concrète du PCFP (Ackerman, Blevins et Malouf 2009) dans le contexte de la construction de ressources lexicales.

Très récemment, la campagne d'évaluation SIGMORPHON (Cotterell et al. 2016) a formalisé une tâche de traitement automatique des langues correspondant à ce problème. Celle-ci consiste à prédire une forme cible à partir d'une forme déjà fléchie (ou d'une forme de citation). Les différentes variantes de cette tâche font varier les informations disponibles (forme seule, forme et sa case, etc.) ainsi que la richesse des données d'apprentissages (accès à des lexiques externes, etc.). La tâche 2 de la campagne SIGMORPHON est la plus proche de la formulation du PCFP : étant donnés une forme, sa case de paradigme, et une case cible, il est demandé aux systèmes de produire la forme cible appropriée. La campagne d'évaluation a permis de distinguer trois

familles d'approches.

Les approches par réseaux neuronaux sont les plus performantes (Kann et Schütze 2016). Le meilleur modèle évalué parvient à 95% d'exactitude sur les données restreintes (entraînement strictement sur les paires fournies pour la tâche, sans accès à d'autres ressources), montrant sur un large ensemble de langues typologiquement variées que l'information nécessaire à la résolution du PCFP peut être extraite de données éparses. Cependant, ces systèmes n'offrent pas d'interprétation claire des structures apprises en termes linguistiques.

Les autres systèmes apprennent des règles plus explicites.

La stratégie de Durrett et DeNero (2013) pour compléter les lexiques incomplets ressemble à celle de maximisation du radical, en favorisant une forme de base plus fréquente en corpus : ils alignent ainsi toutes les formes d'un lexème à une forme « de base » au moyen d'un algorithme itératif fondé sur les distances d'éditions. Dans leur lignée, certains modèles de réinflexion commencent par extraire des opérations d'édition de chaînes en alignant entre elles des formes de surface, puis entraînent des transducteurs à les appliquer.

Certains systèmes s'appuient sur des heuristiques plus directement inspirées de théories linguistiques : Taji et al. (2016) infèrent des segmentations affixales ternaires (préfixe, radical, suffixe). Ahlberg, Forsberg et Hulden (2015) et Sorokin (2016) construisent des « paradigmes abstraits », règles morphologiques formulées pour l'ensemble du paradigme (c'est-à-dire à nouveau par maximisation du radical).

Malgré le succès de ces entreprises pour la tâche de réinflexion, aucun de ces systèmes ne permet d'extraire des représentations explicites du comportement des lexèmes susceptibles de manifester les nombreux points de similarité entre ceux-ci. Ils ne sont donc pas directement utilisables comme point de départ pour une entreprise de classification flexionnelle.

2.2.2 Modèles fondés sur le MGL

Le *Minimal Generalization Learner* ou MGL (Albright et Hayes 1999, 2002 ; Albright et Hayes 2003 ; Albright et Hayes 2006) constitue le premier modèle quantitatif des tests psycholinguistiques d'élicitation morphologiques dits *wug test* (Berko 1958). Il procède à l'inférence de règles qui relient deux formes de surface. Le MGL, originellement conçu pour l'anglais, a

inspiré des travaux qui s'appuient sur des biais linguistiquement motivés pour inférer efficacement des règles d'alternance dans un ensemble de données spécifique. Les modèles de ce type échappent au Problème de la Segmentation, car ils peuvent prendre des décisions locales, ne supposent pas l'existence d'unités sous-jacentes dont la qualité est difficile à évaluer, et n'ont pas besoin de donner un statut de signe aux sous-chaînes qu'ils extraient.

2.2.2.1 Le Minimal Generalization learner

Le MGL infère incrémentalement, à partir de couples de formes, un ensemble de règles. Leur structure générale est celle de (9), familière depuis Chomsky et Halle (1968), qui se lit : « A devient B entre C et D ». Ainsi la règle (10), qui concerne les verbes anglais, se lit : « on ajoute /-əd/ final après /-t-/ et /-d-/ pour former le passé. »

(9) $A \rightarrow B \; / \; C\_D$

(10) $\varnothing \rightarrow \text{əd} \; / \; X \begin{bmatrix} \text{+coronal} \\ \text{+anterior} \\ \text{-nasal} \\ \text{-continuant} \end{bmatrix}\_$

Les étapes d'apprentissage sont les suivantes :

1. Le programme commence par inférer des règles spécifiques à une seule paire de formes.
2. Puis, il crée incrémentalement des règles de généralité croissante en fusionnant les règles initiales au fil de la découverte de nouvelles formes. Toutes les règles de généralité intermédiaire sont conservées.
3. L'ensemble des règles sont scorées pour l'application aux formes inconnues.

Recherche des alternances locales Afin de déterminer les règles spécifiques à une paire de formes, le programme cherche un changement unique dans les formes d'entrées, cherchant dans l'ordre un changement suffixal, préfixal puis interne (Albright et Hayes 2002) afin de déterminer les membres A, B, C et D de la règle [6] :

6. Albright et Hayes (2006) mentionnent un algorithme d'alignement sensible à la phonologie dans la généralisation des règles, mais n'est pas clair quant à sa possible utilisation pour détecter le changement lui-même. Le programme distribué semble suivre la description d'Albright et Hayes (2002).

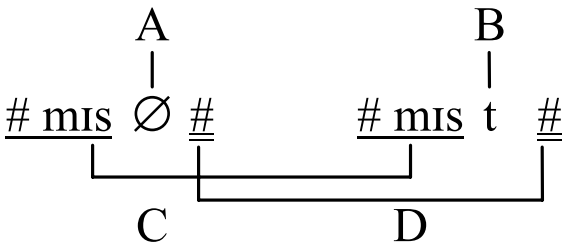


FIGURE 2.1 – Illustration du fonctionnement du MGL : recherche d'une règle. (Albright et Hayes 2002).

> Le modèle commence par trouver le plus long préfixe partagé par les deux formes (e.g., #mɪs), pour créer le terme C (contexte gauche). Le modèle examine ensuite le reste du matériel et trouve le plus long suffixe commun, pour créer le terme D (contexte droit). Le matériel restant est le changement ; le matériel non partagé de la première chaîne est le terme A, et celui de la seconde forme est le terme B [7].

La figure 2.1, tirée d'Albright et Hayes (2002), illustre l'alignement de la paire miss/missed, qui permet de déduire la règle en (11). De la même façon, la paire press/pressed produit la règle (12). Dans les règles de Albright et Hayes (2002), le symbole « ∅ » dénote la chaîne vide.

(11) ∅ → t / mɪs_#

(12) ∅ → t / prɛs_#

Généralisation incrémentale Les contextes de ces règles sont spécifiques à une paire seulement. Chaque fois qu'une nouvelle règle est rencontrée, il est généralisé avec chaque patron présentant une même alternance structurelle (A → B). Les fusions se font donc incrémentalement, et toujours deux à deux. Toutes les règles de généralité intermédiaire sont conservées.

Lors de la généralisation, on obtient une caractérisation du contexte par GÉNÉRALISATION MINIMALE en alignant les deux contextes phonème par phonème à partir du changement cen-

7. [En anglais dans le texte] « *the model first finds the maximal left-side substring shared by the two forms (e.g., #mɪs), to create the C term (left side context). The model then examines the remaining material and finds the maximal substring shared on the right side, to create the D term (right side context). The remaining material is the change; the non-shared string from the first form is the A term, and from the second form is the B term* ».

$$
\begin{array}{rcccccc}
 & \varnothing \rightarrow & \text{t} & / & \text{m} \quad \text{ɪ} & \text{s} \_\_ & \# \\
+ & \varnothing \rightarrow & \text{t} & / & \text{pr} \quad \text{ɛ} & \text{s} \_\_ & \# \\
\hline
= & \varnothing \rightarrow & \text{t} & / & \text{X} \quad \begin{bmatrix} \text{+syllabic} \\ \text{–low} \\ \text{–back} \\ \text{–tense} \\ \text{–round} \end{bmatrix} & \text{s} \_\_ & \#
\end{array}
$$

FIGURE 2.2 – Illustrations du fonctionnement du MGL : généralisation. (Albright et Hayes 2002).

tral. La généralisation d'une paire de phonèmes est définie par l'intersection de leurs traits phonologiques. La figure 2.2 montre comment ce processus aboutit à la règle (13). Albright et Hayes (2006) raffinent cette étape pour la modélisation de la phonologie, avec une procédure d'alignement des contextes plus subtile, et en introduisant des quantificateurs. Dans le contexte (14), qui signifie « où un segment non antérieur apparaît quelque part plus loin dans le mot [8] ». En effet, `[+seg]` caractérise n'importe quel segment phonologique, et * indique un nombre quelconque de répétitions, entre 0 et l'infini. Ces conventions sont celles des expressions régulières.

(13) $\varnothing \rightarrow \text{t} / \text{X} \begin{bmatrix} \text{+syllabic} \\ \text{-low} \\ \text{-back} \\ \text{-tense} \\ \text{-round} \end{bmatrix} \text{s\_\#}$

(14) _([+seg])* [-anterior]

Pour chaque paire de formes dans les données d'apprentissage, on obtient donc un ensemble de règles de généralité variable à même de les dériver, chacun reliant la paire à un ensemble d'autres paires de formes avec lesquelles elle entre en relation analogique.

Sélection par scorage L'ensemble de ces règles de granularité variable sont conservés en mémoire. Le but du programme est de simuler le comportement les locuteurs pour les *wug tests*. Après avoir appris les règles sur des paires de formes d'une langue, lorsque l'on présente une forme au système, il lui faut choisir une règle pour générer la forme cible. À cette fin, les règles obtiennent un score de fiabilité qui est le ratio entre le nombre de paires de formes

8. [En anglais dans le texte] « *where a non-anterior segment follows somewhere later in the word* ».

qu'elles dérivent correctement, ou SUCCÈS (*hits*) et leur PORTÉE (*scope*, c'est-à-dire le nombre d'items auxquels elles peuvent s'appliquer).

$$\text{fiabilité}(p) = \frac{\text{succès}(p)}{\text{portée}(p)}$$

Une règle décrivant une alternance entre les cases C_1 et C_2 est considérée comme APPLICABLE à une forme de la case C_1 si et seulement si la forme remplit la DESCRIPTION STRUCTURELLE de cette règle pour la case C_1, c'est-à-dire la combinaison de son contexte et du membre C_1 de son alternance. Par exemple, la règle de l'exemple (11) n'est applicable qu'aux formes de présent de la forme /mis#/, c'est-à-dire à un unique verbe. Sa fiabilité est de 1, car elle dérive correctement le passé de la seule forme à laquelle elle est applicable. La règle de l'exemple (13) est elle applicable à tous les présents de la forme (15). La description structurelle de chaque règle classifie donc les formes en deux catégories selon son applicabilité à ces formes. La fiabilité d'une règle est la précision de ce classifieur.

(15) X $\begin{bmatrix} \text{+syllabic} \\ \text{-low} \\ \text{-back} \\ \text{-tense} \\ \text{-round} \end{bmatrix}$ s#

Cependant, la qualité d'une règle dépend également de la taille de la portée, comme le remarquent Albright et Hayes (2002) :

> Une fiabilité fondée sur une large portée (par exemple, 990 prédictions correctes sur 1000) est préférable à une fiabilité fondée sur une portée étroite (par exemple, 5 sur 5) [9].

Afin de capter cela, ils ajustent les scores au moyen de limites de confiance. De cette façon, à fiabilité égale, on préférera la règle la plus générale, mais si une règle de petite portée est beaucoup plus fiable, elle peut être préférée à une règle plus générale. Ce système permet de capter ce que Albright et Hayes (2003) nomment des « îlots de fiabilité » : il se peut qu'une

9. [En anglais dans le texte] « *reliability based on high scope (for example, 990 correct predictions out of 1000) is better than reliability based on low scope (for example, 5 out of 5)* ».

règle applicable à un petit nombre d'items très homogènes soit très fiable, tandis que la même alternance représentée par une règle dont le contexte est plus général est moins fiable.

2.2.2.2 Modèles apparentés

Le principe de l'inférence de règles analogiques à partir de paires de formes a été repris par les tenants des analyses mot et paradigme quantitatives. Bonami, Boyé et Henri (2011), Bonami et Boyé (2014) et Bonami et Luís (2014) combinent l'idée directive du MGL et celle d'Ackerman, Blevins et Malouf (2009), en évaluant la prédictibilité au sein des paradigmes au moyen de l'entropie conditionnelle sur la base de patrons d'alternance bidirectionnels. Plutôt que de garder un ensemble de règles de granularités variables, un patron unique est associé à toute paire de cases. Les patrons fournissent donc une partition de l'ensemble des lexèmes. Ce choix permet de calculer des distributions de probabilité sur le choix d'un patron. Cette réduction du nombre de patrons a pour conséquence de drastiquement réduire le temps de calcul et les besoins en mémoire de l'algorithme. Cela constitue un avantage pratique important, car là où le MGL était prévu pour inférer une unique alternance, ces travaux calculent des distributions de patrons pour l'ensemble des paires de cases d'un paradigme de flexion. Garder l'ensemble des généralisations intermédiaires ne serait pas faisable dans cette perspective (pour exemple, les 51 cases du paradigme verbal du français donnent lieu à 1275 combinaisons [10]).

Pour identifier les alternances, les modèles qui s'inscrivent dans cette ligne s'appuient sur des biais adaptés à un ou quelques ensembles de données. Bonami et Boyé (2014) et Bonami et Luís (2014) cherchent un unique changement, d'abord suffixal, puis préfixal, sinon circonfixal. Bonami et Beniamine (2016) définissent des changements suffixaux, avec une éventuelle alternance interne au radical. Ces stratégies sont décidées selon les données, mais peuvent nécessiter la conception d'un nouveau programme *ad hoc* pour chaque nouveau lexique.

Nous proposons ici au contraire un modèle d'inférence de patrons bidirectionnels qui est à

10. En effet, dans le pire des cas, c'est-à-dire si toutes les paires de formes vues exemplifient la même alternance, le MGL doit faire $2^{n-1} - 1$ comparaisons et retenir en mémoire $2^n - 1$ patrons au terme de la procédure. La procédure de Bonami, Boyé et Henri (2011) ne retient en mémoire à terme qu'un patron par alternance, et fait au pire $n - 1$ comparaisons.

même d'identifier des alternances indépendamment du type d'exponence, et sans le connaître *a priori*.

2.3 Inférence automatique d'alternances morphologiques

Cette section présente notre modèle d'inférence de patrons d'alternance bidirectionnels. Notre système prend en entrée des paradigmes dont les formes sont transcrites en notation phonémique ainsi qu'une spécification sous forme de matrice de traits de la valeur des phonèmes utilisés en traits. Il calcule, pour chaque paire de cases possible parmi l'ensemble des cases du paradigme, un ensemble de patrons d'alternance qui relient les paires de formes. Chaque paire de formes est liée par un patron.

L'algorithme de recherche de patron suit un processus similaire à celui d'Albright et Hayes (2002), et procède en trois étapes : trouver, pour chaque paire de formes, l'ensemble des alignements localement optimaux, et en déduire des patrons élémentaires ; généraliser les patrons élémentaires en fusionnant les alternances structurellement identiques ; choisir les patrons selon leur pouvoir descriptif global pour l'ensemble des lexèmes.

2.3.1 Alignement

Le but de notre algorithme est d'étendre la stratégie de Bonami, Boyé et Henri (2011), Bonami et Boyé (2014) et Bonami et Luís (2014) afin d'identifier des alternances de tous types, sans savoir à l'avance si l'on a affaire à de la flexion préfixale ou suffixale, continue ou discontinue, avec des alternances de radical ou non, etc. Afin de déterminer le patron qui relie deux formes, il faut être capable de comparer les caractères de chaque forme un à un : si les deux caractères sont identiques, ils font partie du contexte, sinon ils font partie de l'alternance.

En conséquence, l'inférence d'un patron d'alternance repose principalement sur le choix d'un alignement de deux formes.

L'alignement optimal peut varier d'une langue à l'autre. Par exemple, si l'on sait qu'il n'y a aucune préfixation, il est possible d'aligner les formes à gauche, comme le montre l'exemple d'un verbe du français dans la figure 2.3. Les phénomènes préfixaux, par exemple la flexion

Français								
PRS.1SG ‘amène’	a	m	ɛ	n				
	|	|	|	|	|			
PRS.2PL ‘amenez’	a	m	ø	n	e			

Swahili								
FUT.1SG ‘je voudrai’	n	i	t	a	t	a	k	a
	|	|	|	|	|	|	|	|
PRS.2PL ‘tu veux’		m	n	a	t	a	k	a

Persan									
PRS.1SG ‘J’achète’	m	i	x	a	r			a	n
	|	|	|	|	|	|	|	|	|
PAS.PERF.1SG ‘J’ai acheté’			x	a	r	i	d	a	n

Arabe								
PFV.3SG ‘il a écrit’			k	a:	t	a	b	a:
	|	|	|	|	|	|	|	|
IPF.3SG ‘il écrit’	j	u	k	a:	t	i	b	u

FIGURE 2.3 – Les alignements optimaux dépendent du type d’exponence.

verbale du swahili peut être analysée au moyen d'un alignement à droite. L'exemple du persan montre que l'alignement permettant d'identifier une alternance peut être complexe et discontinue, même dans un système strictement concaténatif, si l'on doit comparer une forme préfixée à une forme suffixée. Dans le cas de l'arabe, la transfixation généralisée produit des discontinuités systématiques. Il ne serait donc pas suffisant de paramétrer le système pour lui indiquer s'il doit aligner les formes à gauche (pour les systèmes sans préfixes) ou à droite (pour les systèmes sans suffixes).

Localement à une paire de formes, nous sommes face à un problème d'alignement de chaînes de caractères (qui représentent des phonèmes). Nous proposons de le résoudre en calculant des distances d'édition.

La distance de Levenshtein (Levenshtein 1966), correspond au nombre minimal d'insertions, de suppressions et de substitutions nécessaires pour transformer une chaîne en une autre. La figure 2.4 détaille le calcul de la distance d'édition entre les formes /amɛn/ et /amøne/. Dans la figure, nous notons C la copie de caractère, qui est gratuite. Les trois autres opérations, substitution, ajout et suppression d'un caractère (notées respectivement S, A, D), coûtent chacune 1. En somme, la meilleure séquence d'opération est celle qui identifie au mieux les identités entre les deux chaînes. Dans le cas présent, nous ne somment pas intéressés *in fine* par la distance elle-même, mais par l'alignement qui est induit par la séquence d'opérations.

	a	m	ɛ	n		
	|	|	|	|	|		
	a	m	ø	n	e	
	C	C	S	C	A	
distance de Levenshtein	0	0	1	0	1	$= 2$

FIGURE 2.4 – Une distance d'édition.

2.3.1.1 Ambiguïtés virtuelles entre alignements optimaux

La distance de Levenshtein ne permet cependant pas toujours de distinguer localement les alignements les plus naturels morphophonologiquement parlant. Considérons le nom tchèque čivava (chihuahua), et plus précisément l'alternance entre son génitif pluriel čivav, prononcé /tʃɪvaf/ en raison d'un dévoisement final et son instrumental pluriel čivavami prononcé /tʃɪvavamɪ/. L'alignement intuitif pour un linguiste, en raison du dévoisement final, aligne /f/ avec le deuxième /v/ (figure 2.5). Cependant, il existe sept autres alignements qui présentent la même distance de Levenshtein. Par exemple, l'un de ces alignments suppose un infixe /-av-/ et qui présente la même distance de Levenshtein.

1.	GEN.PL 'chihuahua' (čivav)	ʧ	ɪ	v	a	f			
		|	|	|	|	|	|	|	|	
	INS.SG 'chihuahua' (čivavami)	ʧ	ɪ	v	a	v	a	m	ɪ	
	distance de Levenshtein	0	0	0	0	1	1	1	1	$= 4$
2.	GEN.PL 'chihuahua' (čivav)	ʧ	ɪ	v			a	f		
		|	|	|	|	|	|	|	|	
	INS.SG 'chihuahua' (čivavami)	ʧ	ɪ	v	a	v	a	m	ɪ	
	distance de Levenshtein	0	0	0	1	1	0	1	1	$= 4$

FIGURE 2.5 – Exemple d'alignements concurrents ayant la même distance de Levenshtein.

Cette ambiguïté d'alignement est virtuelle : il n'existe pas réellement huit bons alignements linguistiquement parlant. C'est notre distance d'édition qui est en cause. Pour corriger cette situation, il faut considérer que l'alignement /f/-/v/ est préférable à celui /f/-/m/ en raison de leur similarité phonétique. Il nous faut donc une opération de substitution sensible à la similarité phonologique. Suivant Albright et Hayes (2006), nous pondérons donc la substitution par la mesure de similarité phonologique proposée par Frisch, Pierrehumbert et Broe (2004). Celle-ci se fonde sur la proportion de classes naturelles partagées entre deux segments. Soit $C(a)$ l'ensemble des classes naturelles auxquelles appartient un segment a, la similarité entre deux

segments a et b se définit comme suit :

$$\text{sim}(a, b) = \frac{|C(a) \cap C(b)|}{|C(a) \cup C(b)|}$$

Cette similarité s'échelonne entre 0 et 1, on définit donc la distance entre a et b par $1 - \text{sim}(a, b)$. Dans ce schème, le coût de l'insertion est un paramètre. Idéalement, la substitution entre deux segments très similaires doit être préférée à une insertion et une suppression, tandis qu'une insertion et une suppression sont préférées à la substitution de segments dissimilaires. Nous avons donc choisi de fixer le coût de l'insertion à $\frac{1}{3}$ de la similarité moyenne au sein de l'inventaire de phonèmes. Soit I l'ensemble des phonèmes connus et $I \times I$ le produit cartésien sur cet ensemble, nous définissons A le coût moyen d'une substitution dans ce système :

$$A = \frac{1}{|I \times I|} \cdot \sum_{(a,b) \in I \times I} \text{sim}(a, b)$$

Le tableau 2.2 présente les coûts des opérations dans les deux schèmes utilisés.

Distance	Copie	Insertion ou suppression	Substitution de a par b
Levenshtein	0	1	0 si $a = b$ sinon 1
Similarité phonologique	0	$\frac{1}{3} \cdot A$	$1 - \text{sim}(a, b)$

Tableau 2.2 – Pondération des opérations d'édition.

L'implémentation du programme qui infère les patrons permet de choisir l'usage de l'une ou l'autre mesure.

2.3.1.2 Ambiguïtés réelles entre alignements optimaux

Il arrive qu'il existe non pas un unique alignement optimal reliant les deux formes mais un ensemble d'alignements également plausibles morphologiquement (Bonami 2014). L'ambiguïté entre ces alignements est réelle dans le sens où tous sont linguistiquement plausibles. Le tableau 2.3 reproduit l'exemple de Bonami (2014, pp.104-106). Il présente trois alignements susceptibles de décrire l'alternance entre les formes imaginaires baba/ba. Trois patrons en découlent, qui fournissent respectivement une analyse préfixale, suffixale ou infixale de cette

alternance. Il est impossible de décider de l'analyse correcte en se fondant strictement sur ces deux formes : c'est leur place dans le reste du système flexionnel qui permet de trancher.

Alignement					Patron
	b	a	b	a	
(i) Préfixe	_	_	b	a	$\epsilon \rightleftharpoons$ ba / _ba
(ii) Suffixe	b	a	_	_	$\epsilon \rightleftharpoons$ ba / ba_
(iii) Infixe	b	_	_	a	$\epsilon \rightleftharpoons$ ab / b_a

Tableau 2.3 – Alignments et patrons pour les formes imaginaires 'baba' and 'ba'.

Considérons les paradigmes nominaux de trois systèmes imaginaires présentés dans la figure 2.4. Pour chacun des systèmes A, B et C, une seule des trois analyses, respectivement infixe, préfixe et suffixe permet de rendre compte de l'ensemble des lexèmes. L'ambiguïté d'analyse locale peut donc être résolue par l'examen des autres lexèmes.

Notre solution est de produire localement tous les patrons issus des alignements optimaux, et de laisser à une étape ultérieure la sélection des patrons les plus adaptés au paradigme (c'est-à-dire les plus généraux).

A.	
SG	PL
ba	baba
to	tabo
ri	rabi
su	sabu
ne	nabe

B.	
SG	PL
ba	baba
to	bato
ri	bari
su	basu
ne	bane

C.	
SG	PL
ba	baba
to	toba
ri	riba
su	suba
ne	neba

Tableau 2.4 – Trois langages imaginaires.

2.3.1.3 Stratégie d'alignement

L'étape d'alignement procède comme suit. Pour chaque paire de cases de paradigme, et pour chaque paire de formes d'un même lexème appartenant à ces cases dans nos données, nous calculons l'ensemble des alignements optimaux de ces deux formes selon les distances d'édition pondérées par la similarité phonologique.

En raison des ambiguïtés réelles, nous cherchons et conservons, durant cette étape de l'algorithme, l'ensemble des alignements optimaux. Une étape ultérieure devra les départager selon leur adéquation au reste du paradigme.

Puisque chaque alignement détermine entièrement un patron, nous calculons immédiatement un ensemble de patrons concurrents pour chaque paire de formes, déduits de l'ensemble des alignements optimaux.

Ainsi l'alignement de /ba/ et /baba/ produit l'ensemble des trois patrons : {$\epsilon \rightleftharpoons$ ba / _ba, $\epsilon \rightleftharpoons$ ba / ba_, $\epsilon \rightleftharpoons$ ab / b_a}. Tandis que Albright et Hayes (2002) notent la chaîne vide par « ∅ », nous la notons par « ϵ », afin d'éviter la confusion avec le symbole de l'api /ø/. Ces patrons sont spécifiques aux formes dont ils proviennent, car leurs contextes contiennent des segments et non des généralisations phonologiques.

2.3.2 Généralisation

La seconde étape consiste à fusionner les patrons présentant la même alternance structurelle de façon à capter des généralisations sur les contextes d'application et sur les alternances. La généralisation des contextes suit globalement la procédure décrite en 2.2.2.1 (Albright et Hayes 2002), adaptée pour une généralisation n-par-n plutôt que 2 par 2.

Les contextes généralisés sont exprimés sous la forme d'expressions régulières où les classes de caractères, entre crochets, correspondent à des classes naturelles de segments phonologiques, et les quantifieurs « ?, +, * » rendent compte de la longueur des séquences concernées.

Étant donné un ensemble de patrons partageant une même alternance structurelle, on souhaite formuler des contraintes phonologiques sur l'application de cette alternance à partir de l'ensemble de leurs contextes. Il nous faut donc déterminer ce que ces contextes ont en commun.

On segmente chaque contexte sur les blancs, notés « _ », puis on généralise de la façon suivante :

i. Aligner les caractères des parties de contextes :
   — Avant un blanc (début de mot), aligner à droite.
   — Entre deux blancs, aligner aux deux bords avec la plus courte des parties de contextes. Les caractères supplémentaires sont alignés ensemble au centre.
   — Après un blanc (fin de mot), aligner à gauche.

ii. Calculer l'ensemble des caractères vus à chaque position d'alignement,

iii. Calculer un quantifieur,

iv. Fusionner ensemble les positions facultatives successives,

v. Associer à chaque ensemble de caractères la plus petite classe naturelle phonologique qui contienne ces sons. Cela revient à trouver l'ensemble des segments qui partagent les traits phonologiques communs à tous les segments de l'ensemble.

Dans l'implémentation, les étapes ii., iii. et iv. sont réalisées simultanément pour des raisons d'efficacité. Le tableau 2.5 détaille chacune de ces étapes pour l'alternance de quelques verbes du français entre les premières personnes du singulier et du pluriel au présent.

	Alternance	Contexte								Lexème
i.	ɛ _ ⇌ ə_ɔ̃ /			a	m	_	n	_		AMENER
	ɛ _ ⇌ ə_ɔ̃ /				p	_	l	_		PELER
	ɛ _ ⇌ ə_ɔ̃ /				ʒ	_	l	_		GELER
	ɛ _ ⇌ ə_ɔ̃ /	p	ʁ	o	m	_	n	_		PROMENER
	ɛ _ ⇌ ə_ɔ̃ /				s	_	m	_		SEMER
	ɛ _ ⇌ ə_ɔ̃ /				l	_	v	_		LEVER
	ɛ _ ⇌ ə_ɔ̃ /				p	_	z	_		PESER
	ɛ _ ⇌ ə_ɔ̃ /				s	_	v ʁ	_		SEVRER
ii.		{p}	{ʁ}	{o,a}	{m,p,ʒ,s,l}	_	{n,v,ʁ,z,l,m}	_		
iii.		{p}?	{ʁ}?	{o,a}?	{m,p,ʒ,s,l}	_	{n,v,ʁ,z,l,m}+	_		
iv.		{p,ʁ,o,a}*			{m,p,ʒ,s,l}	_	{n,v,ʁ,z,l,m}+	_		
v.	ɛ _ ⇌ ə_ɔ̃ /	[eəɛɔɔøœabdflpstvzʁ]*			[bdflmnpstvzɲʃʒ]	_	[lmnvzʁ] +	_		

TABLEAU 2.5 – Généralisation du contexte de trois patrons.

2.3.2.1 Opérations phonologiques régulières

Ci-dessus, nous avons parlé d'ALTERNANCE STRUCTURELLE. Dans la plupart des cas, deux patrons ont la même alternance structurelle s'ils ont une alternance strictement identique. Cependant, les alternances morphologiques prennent parfois la forme d'opérations phonologiques régulières, par exemple un voisement, une palatalisation ou un allongement. Notre système tente de reconnaître ces opérations comme appartenant à la même alternance structurelle.

Les exemples (16) à (18) présentent quelques-unes de ces opérations à l'œuvre dans nos données. Dans les exemples du navajo en (16), la voyelle longue avec ton haut de la base [11] à l'imperfectif alterne avec une voyelle courte sans ton haut au futur. Les formes du français présentées en (17) manifestent une alternance entre une voyelle finale et la semi-voyelle correspondante. Enfin, les formes fléchies des noms du russe présentent souvent des palatalisations, comme dans les exemples en (18).

11. Les formes correspondent à la deuxième base de chaque lexème, voir la discussion en section 2.4.1.

(16) (NAV)

Lexème	IPFV.1	FUT.1
ʾÁDIISHBÁÁH	/pá:h/	/pah/
ʾÁDIISHCHÍÍH	/ʧʰí:h/	/ʧʰih/
ʾAHISHSHÓÓSH	/ʃó:ʃ/	/ʃoʃ/
HAASHKʾĘ́Ę́H	/kʼę́:h/	/kʼęh/

(17) (FR)

Lexème	PRS.1.S	PRS.1.PL
HABITUER	/abity/	abitɥ̃ /
AFFILIER	/afili/	afiljɔ̃ /
TATOUER	/tatu/	/tatwɔ̃ /

(18) (RU)

Lexème	NOM.S	DAT.S
TEMNOTA	/tʲemnota/	/tʲemnotʲe/
TOLPA	/tolpa/	/tolpʲe/
VARKA	/varka/	/varkʲe/
NATURA	/natura/	/naturʲe/
DERZHAVA	/dʲerʒava/	/dʲerʒavʲe/
POLOVʾINA	/polovʲina/	/polovʲinʲe/
PASXA	/pasxa/	/pasxʲe/
PROGRAMMA	/programma/	/programmʲe/
GUBA	/guba/	/gubʲe/
KOSA	/kosa/	/kosʲe/
BUMAGA	/bumaga/	/bumagʲe/
NADEZHDA	/nadʲeʒda/	/nadʲeʒdʲe/
KUKLA	/kukla/	/kuklʲe/

Pour cela, lors de la création initiale de chaque patron, nous itérons sur les caractères des deux membres de l'alternance en évaluant s'il existe une opération phonologique réversible les reliant. Nous retenons, si elle existe, cette description abstraite en plus de la description stricte de l'alternance. Nous disons que deux patrons ont la même alternance structurelle s'ils ont la

même description abstraite, ou à défaut s'ils ont la même description stricte.

Lors de la généralisation, si l'on fusionne deux patrons qui partagent la même description abstraite, mais non la même description stricte, alors on ne conserve que la description abstraite. Si au contraire un patron qui comporte une description abstraite n'est fusionné qu'avec des patrons partageant la même description stricte, alors on ne conserve que la description stricte. En somme, on ne choisit la description abstraite que si elle offre un avantage descriptif.

Nous présentons en (19) à (21) les alternances des patrons inférés par notre programme respectivement pour les exemples (16) à (18). Le programme choisit d'écrire les classes naturelles de segments phonologiques sous la forme d'une liste de segments ou de traits de façon à obtenir la représentation la plus brève. Par ailleurs, nous écrivons C et C^{j} respectivement les consonnes [-PALATALISÉE] et [+PALATALISÉE] et V et V́: respectivement les voyelles courtes sans ton et les voyelles longues avec ton haut.

(19) (NAV) [aːáːéːíːóːǫːíːéːː] ⇌ [aeioąęįǫ]

(20) (FR) [iuy] ⇌ [jwų]ɔ̃

(21) (RU) Ca ⇌ Cʲe

L'identification de ces opérations est importante pour éviter d'inférer un nombre de microclasses inutilement haut, là où il est certain qu'il n'existe pas d'imprédictibilité pour les locuteurs. Par choix, cette procédure ne mène pas à généraliser les alternances qui sont régulières dans la langue mais non phonologiquement naturelles, par exemple l'atténuation en irlandais (Carnie 2008). Par ailleurs, elle ne généralise pas non plus des opérations comme la suppression d'une voyelle ou d'une consonne, car la bidirectionnalité des patrons impose une opération entièrement prédictible dans les deux sens. Il est à noter que la généralisation des opérations phonologiques régulières est très sensible aux choix opérés dans l'élaboration du système de traits distinctifs, puiqu'elle opère sur la géométrie définie par ce système.

2.3.2.2 Éviter la surgénéralisation

Telle que décrite ici, la procédure de généralisation peut mener à surgénéraliser les contextes. Il existe en effet des cas où l'alternance est identique en substance mais apparaît à des endroits

distincts dans le mot.

En français, le patron le plus fréquent (environ 4000 instances) qui lie la première personne du singulier et la seconde personne du pluriel au présent est une suffixation en -e, comme le montrent les exemples (22a,b). Le verbe ASSEOIR présente une alternance de la base en plus de cette suffixation, comme décrit en (22c).

L'analyse habituelle de cette alternance est asje+∅ alternant avec asej+e [12]. Cependant, du point de vue de l'alignement local des formes, aligner les deux /e/ finaux mène à une plus petite distance d'édition. Puisque cette alternance spécifique est unique dans les données (elle n'est partagée qu'avec le dérivé rasseoir), il n'existe aucune bonne raison de produire le patron correspondant à l'analyse habituelle plutôt que celui qui figure en (22c).

Les alternances en (22) peuvent toutes être décrites par la même description structurelle : $\epsilon \rightleftharpoons$ e. En a., b. et c. le changement est strictement suffixal, mais en c. il est interne à la base.

La fusion de l'ensemble des patrons présentant cette alternance suffixale produit le patron $\epsilon \rightleftharpoons$ e / X+_[Eij]*. L'application (gloutonne) de ce patron à la forme asje produit la forme incorrecte *asjee. En effet, malgré l'identité de l'alternance structurelle, les alternances en question sont en fait distinctes. Il est donc préférable de ne pas fusionner les trois patrons, malgré leur alternance commune.

(22)

	PRS.1SG	PRS.2PL	patron élémentaire
a. MANGER	mɑ̃ʒ	mɑ̃ʒe	$\epsilon \rightleftharpoons$ e / mɑ̃ʒ_
b. LAVER	lav	lave	$\epsilon \rightleftharpoons$ e / lav_
c. ASSEOIR	asje	aseje	$\epsilon \rightleftharpoons$ e / as_je

Ces problèmes sont toujours dûs à la présence de matériel alternant répété ailleurs dans le mot. Nous identifions donc à l'avance les cas potentiellement problématiques. Lorsqu'une alternance structurelle présente de tels patrons, nous procédons à une généralisation incrémentale, en vérifiant à chaque étape si la généralisation est correcte ou si elle produit un contexte trop général. Cette procédure est plus lente que la généralisation n par n, mais rarement nécessaire.

12. Pour les besoins de la démonstration, nous supposons ici une variété de français où asseyez se prononce /aseje/ plutôt que /asɛje/.

2.3.3 Sélection

La procédure d'alignement des formes puis de fusion et généralisation des patrons peut produire plusieurs patrons concurrents pour une même paire de formes, parmi lesquels il reste à choisir celui qui est le plus approprié au vu du reste du système. Nous opérons cette sélection en attribuant des scores aux patrons.

Suivant Albright et Hayes (2002), nous appelons SUCCÈS le nombre de lexèmes connus dérivés correctement par un patron. Nos patrons étant bidirectionnels, on considère qu'un patron d'alternance p est une paire de fonctions injectives dont la seconde est la réciproque de la première, (F, F^{-1}) de $\Sigma*$ dans $\Sigma*$. Une paire de mots (w, w') instancie un patron $p = (F, F^{-1})$ si et seulement si $F(w) = w'$.

Pour chaque paire de formes, nous cherchons l'ensemble des patrons qu'elles instancient. Ces patrons incluent nécessairement ceux qui ont été inférés à partir de cette forme, mais peuvent également comprendre des patrons inférés à partir d'autres formes et dont la généralisation se trouve être correcte pour celles-ci. Il s'agit d'une propriété importante de notre algorithme, car cela rend la recherche de patrons robuste face à de possibles obscurcissements locaux d'une alternance régulière. On peut donc trouver le bon patron même lorsqu'une paire de formes est telle que l'alignement localement optimal est peu satisfaisant au vu des autres paires de formes, à condition que ce patron ait été trouvé ailleurs et que la généralisation de son contexte le rende applicable à cette paire.

Parmi l'ensemble des patrons qui dérive correctement une paire de formes, nous choisissons celui qui a le meilleur succès. Dans le contexte des langages imaginaires A, B et C du tableau 2.4, le tableau 2.6 montre comment cette méthode permet de choisir le patron optimal pour décrire l'alternance entre les formes « ba » et « baba » selon le reste du paradigme.

Les ambiguïtés d'alignement arrivent en particulier lorsqu'une partie d'un mot répète du matériel affixal. C'est le cas par exemple du verbe français CRÉER, dont l'alternance entre présent première personne du singulier et seconde personne du pluriel est présentée dans l'exemple (23). Selon l'alignement des /e/, on peut supposer soit un changement infixal, soit un changement suffixal. Cependant l'immense majorité des verbes se comporte comme ceux de (22a) et

Alignement					Patron	succès		
						A	B	C
	b	a	b	a				
(i) Préfixe	_	_	b	a	$\epsilon \rightleftharpoons$ ba / _ba	1	4	1
(ii) Suffixe	b	a	_	_	$\epsilon \rightleftharpoons$ ba / ba_	1	1	4
(iii) Infixe	b	_	_	a	$\epsilon \rightleftharpoons$ ab / C_V	4	1	1

Tableau 2.6 – Scorage de trois patrons.

(22b), ce qui permet de choisir le patron suffixal. En russe, l'alternance entre le nominatif singulier et le datif pluriel décrite en (24a) se prête à deux alignements optimaux, l'un résultant en un patron suffixal en /-am/, l'autre un un patron infixal en /ma-/. Après généralisation, le patron suffixal est instancié par 681 paires de formes, tandis que le patron infixal ne l'est que par 26 formes. L'étape de sélection et scorage permet donc de choisir le patron suffixal. Il est notable que ce résultat est obtenu sans que l'algorithme introduise un biais en faveur de la suffixation et contre l'infixation : pour les données examinées, l'hypothèse d'une suffixation produit une meilleure généralisation.

(23)

	PRS.1SG	PRS.2PL	patrons possibles
a. FR : CRÉER	/kʁe/	/kʁee/	{$\epsilon \rightleftharpoons$ e / kre_ ,$\epsilon \rightleftharpoons$ e / kr_e}

(24)

	NOM.SG	DAT.PL	patrons possibles
a. RU : AERODROM	/aerodrom/	/aerodromam/	{$\epsilon \rightleftharpoons$ am / aerodrom_,
			$\epsilon \rightleftharpoons$ ma / aerodro_m}
b. RU : VOPROS	/vopros/	/voprosam/	{$\epsilon \rightleftharpoons$ am / vopros_}

Au terme de cette procédure, on obtient pour chaque lexème un ensemble de patrons d'alternance indexés par paire de cases.

2.3.4 Évaluation

La visée de note algorithme est de servir de base à la comparaison quantitative entre systèmes flexionnels. La pertinence des patrons inférés se jugera donc *in fine* aux généralisations translinguistiques qu'ils permettront de dégager. Cependant, leur capacité à capter des généralisations internes à un système peut d'ores et déjà être évaluée au travers d'une tâche de prédiction. Nous évaluons la capacité prédictive des patrons d'alternance pour l'ensemble des systèmes flexionnels étudiés dans cette thèse : les noms du russe, les verbes du français, de l'anglais, du portugais Européen, du chatino de Zenzontepec et de Yaitepec, de l'arabe classique et du navajo.

Nous menons une validation croisée en 10 plis. Pour chaque paire de cases, ou alternance, l'entraînement se fait sur 90% des paires de formes choisies aléatoirement. On évalue ensuite la prédiction en essayant de prédire la forme cible, dans les deux directions, sur les 10% de données restantes. Les données sont découpées aléatoirement en dix tranches de 10% qui tournent dix fois afin que chacune ait servi d'ensemble d'évaluation et d'ensemble d'entraînement.

Lors de l'entraînement, on apprend tout d'abord un ensemble de patrons sur les formes d'entraînement. Pour chaque forme d'entraînement, il existe un ensemble de patrons dont cette forme remplit les contraintes phonotactiques. Ces patrons sont APPLICABLES à cette forme. Suivant Bonami (2014), nous appelons CLASSE DE PATRONS APPLICABLES l'ensemble des patrons qui sont susceptibles d'être appliqués à une forme. Nous calculons sur l'ensemble d'entraînement la probabilité conditionnelle $P(\text{patron}|\text{classe})$ d'après les fréquences relatives observées des patrons et des classes. Nous calculons également la probabilité simple $P(\text{patron})$ pour servir de repli.

Lors de l'évaluation, nous considérons des formes uniques, dont nous connaissons la case de paradigme, et tentons de prédire la forme cible dans la case cible au moyen de l'ensemble de patrons et de classes de patrons applicables appris sur cette paire de cases. Pour chaque forme, nous calculons sa classe de patrons applicables. Si la classe est connue, nous effectuons la prédiction au moyen du patron qui maximise $P(\text{patron}|\text{classe})$. Si la classe est inconnue, nous choisissons le patron qui maximise $P(\text{patron})$. Si la prédiction est fausse, ou si aucun

patron connu n'est applicable, il s'agit d'un échec. Si la prédiction résulte en la forme correcte, il s'agit d'une réussite.

Enfin, nous comparons 5 versions de notre algorithme, selon la stratégie d'alignement. Celui-ci peut être fixe à gauche ou à droite, suivre la stratégie décrite dans Albright et Hayes (2002), ou enfin reposer sur des distances d'éditions soit simples (distance de Levenshtein) soit fondées sur la similarité phonologique. La phase de scorage n'est nécessaire que dans le cas des distances d'éditions, les autres méthodes ne générant pas d'ambiguïtés d'alignement. De même, la généralisation des alternances à des opérations phonologiques n'est effectuée qu'avec les distances d'édition.

Nous rendons compte dans le tableau 2.7 de l'exactitude moyenne (pourcentage de formes dérivées correctement) à travers l'ensemble des paires d'alternances, dans chaque direction, et à travers les 10 plis. Comme attendu, les stratégies d'alignement fixe à droite ou à gauche sont efficaces exclusivement pour certains ensembles de données mais ne permettent pas de capter de bonnes généralisations dans l'ensemble des langues. L'alignement qui permet de trouver un changement unique décrit par Albright et Hayes (2002) produit de bons résultats en anglais, français, portugais européen, et chatino, mais non en arabe. Notre stratégie d'alignement, qu'elle emploie des distances d'édition simples ou fondées sur la similarité phonologique, produit des scores plus élevés en chatino et en arabe, sans perte de performance significative dans les autres langues. En effet, elle permet de capter des alternances arbitrairement complexes sans présumer de leur forme.

Nous nous attendions à ce que les distances d'édition pondérées par la similarité phonologique produisent des patrons plus performants que les distances de Levenshtein. De façon surprenante, nos résultats ne montrent pas de différence importante entre les deux distances d'édition. Cela est dû à la robustesse de l'algorithme d'inférence des patrons : les distances de Levenshtein génèrent plus de patrons inutiles lors de l'étape d'alignement, mais ceux-ci sont écartés lors de l'étape de scorage. Cependant, cette surgénération a un coût computationnel, car lors du scorage, il faut évaluer chacun des patrons générés pour chacune des paires de formes disponibles. En conséquence, l'usage des distances d'éditions pondérées par la phonologie permet de raccourcir le temps d'exécution. Nous préférons donc pour la suite employer

	Lexèmes	Baseline	Préfixal	Suffixal	Levenshtein	Phono
Anglais	6064	93.7%	31.1%	93.7%	93.6%	93.6%
Arabe	1018	46.2%	26.1%	44.8%	80.2%	81.8%
Français	5249	94.4%	23.6%	94.5%	94.4%	94.4%
Portugais	1996	93.3%	18.1%	93.3%	91.3%	91.3%
Russe	1539	73.1%	28.0%	73.1%	74.8%	74.7%
Yaitepec	324	33.3%	29.6%	29.0%	35.5%	35.5%
Zenzontepec	392	56.0%	56.6%	25.1%	56.9%	56.9%
Navajo	2157	37.2%	32.4%	25.2%	42.0%	41.8%

TABLEAU 2.7 – Résultats de l'évaluation : pourcentage d'exactitude moyenne.

celles-ci.

Les performances sont très variables d'une langue à l'autre. Nous envisageons trois explications possibles : les lexiques verbaux du chatino sont particulièrement petits (moins de 400 lexèmes), et 90% des formes ne suffisent probablement pas, dans le cas général, à apprendre des généralisations couvrant les 10% restants. Par ailleurs, nous verrons au chapitre 3, qui étudie le problème de remplissage des cases de paradigmes dans ces données, que la prédiction au sein de certains paradigmes est plus difficile qu'au sein d'autres paradigmes. Enfin, nous discutons en section 2.4 discute d'une propriété des systèmes du russe, chatino et navajo qui contribue à expliquer des scores plus bas dans ces systèmes, et propose une solution alternative.

	Baseline	Préfixal	Suffixal	Levenshtein	Phono
Anglais	34.7	3743.3	34.5	34.1	34.1
Arabe	321.9	486.2	333.6	32.1	21.9
Français	25.3	3577.0	25.3	24.2	24.2
Portugais	18.0	1461.8	17.8	16.9	17.0
Russe	87.2	897.9	86.9	36.1	37.5
Yaitepec	143.3	174.2	165.8	112.3	112.3
Zenzontepec	70.8	47.9	227.2	49.0	49.9
Navajo	650.8	700.1	875.0	463.5	468.7

Tableau 2.8 – Résultats de l'évaluation : nombre moyens de patrons.

2.4 Application du modèle à nos lexiques

Pour les besoins de cette thèse, nous avons calculé les patrons d'alternance pour toutes les paires de cases de paradigme dans chaque langue. La combinatoire donne donc lieu à un grand nombre de paires de cases (près de 6000 paires de cases en arabe), comme indiqué dans le tableau 2.9.

	Cases de paradigme	Paires de cases
Portugais	69	2346
Zenzontepec	4	6
Arabe	109	5886
Russe	13	78
Anglais	8	28
Français	51	1275
Yaitepec	12	66

Tableau 2.9 – Nombres de paires de cases et nombre moyen de patrons.

Il est donc impossible de présenter dans ici l'ensemble des patrons obtenus. Cette section

présente un aperçu des patrons inférés. Tout d'abord, nous discutons la nécessité d'analyser certains systèmes flexionnels comme la conjonction de deux paradigmes, et les conséquences sur l'analyse en patrons d'alternance. Nous présentons ensuite un aperçu des patrons pour les systèmes flexionnels qui ne comportent qu'une dimension de variation.

2.4.1 Systèmes bipartites

L'évaluation des patrons révèle des scores particulièrement bas pour le navajo, le chatino et le russe. Or les descriptions de ces systèmes flexionnels distinguent chaque fois deux dimensions de variation : deux bases successives correspondant à la structure syllabique en navajo, un système tonal et segmental distincts en chatino, enfin, un système segmental et accentuel en russe. Dans cette section, nous proposons de modéliser les alternances entre formes dans ces langues par la conjonction de deux patrons opérant chacun sur une dimension.

2.4.1.1 Les deux bases du navajo

Les verbes du navajo [13] sont habituellement décrits comme structurés en trois « domaines » (Sapir et Hoijer 1967 ; Kari 1989) comme indiqué en (25).

(25) [(disjoint) conjoint - (classifieur) radical]

Le domaine disjoint, facultatif, contient des morphèmes aspectuels. Le domaine conjoint, obligatoire, contient des morphèmes qui marquent le mode et la personne du sujet. Le domaine du radical, obligatoire, coïncide toujours avec la dernière syllabe du verbe, et présente un marqueur de valence (le « classifieur ») souvent incorporé. L'exemple (26) illustre cette analyse.

(26) yish-cha
IPFV.1SG-pleurer.IPFV
Je pleure

McDonough et ses coauteurs (McDonough 1990 ; McDonough 1999 ; McDonough 2000 ; McDonough 2003 ; McDonough et Wood 2008 ; McDonough 2014) ont conduit une étude des

13. La section qui suit a été élaborée en collaboration avec Olivier Bonami et Joyce McDonough. Elle a fait l'objet d'une présentation au colloque ISMo 2017 (Beniamine, Bonami et McDonough 2017).

propriétés phonétiques, phonotactique et phonologiques des verbes des langues athabaskanes, et fournissent un ensemble d'arguments pour identifier un « verbe noyau » constitué de deux éléments distincts mais interdépendants, porteurs de la spécification morphosyntaxique minimale pour constituer un verbe bien formé.

Dans cette optique, le domaine conjoint constitue une base à laquelle se préfixe le domaine disjoint. Les verbes sont donc des sortes de mots composés formés de deux bases, que nous appellerons base$_1$ et base$_2$, comme en (27) [14]. La base$_2$ coïncide toujours avec la dernière syllabe du mot, si bien que la frontière entre les deux bases est toujours identifiable de manière déterministe.

(27) [[aspect-radical$_1$] [valence-radical$_2$]]

Les marqueurs de mode et personne du conjoint sont généralement combinés en un exposant cumulatif. Analyser la frontière entre base$_1$ et base$_2$ comme une frontière morphologique permet de rendre compte des groupes consonantiques qui apparaissent à cet endroit mais ne peuvent apparaître à l'intérieur d'un mot. Phonétiquement, les deux bases sont toujours saillantes pour les locuteurs, la seconde constituant toujours la dernière syllabe du composé. Tandis que les noms sont minimalement monosyllabiques en navajo, les verbes ne peuvent faire moins de deux syllabes.

Les bases$_2$ forment une classe fermée de 550 items (Young et Morgan 1987). Les bases$_1$ sont environ 330 et figurent dans Young et Morgan (1987) sous la forme de tables de paradigmes. Dans nos données, elles donnent lieu à 802 paradigmes de surface distincts (en prenant en compte les variations aspectuelles, ainsi que celles dues à l'harmonie consonantique ou la défectivité). Les entrées lexicales du dictionnaire sont des combinaisons de ces deux bases. Les entrées du dictionnaires retenues (annexe A) constituent 2113 verbes distincts, soit 12% des 181500 combinaisons possibles de bases$_1$ et bases$_2$. Les exemples (28) à (31) montrent comment deux bases$_1$ et deux bases$_2$ se combinent pour former quatre verbes distincts. Les combinaisons sont ici sémantiquement transparentes, mais il est fréquent qu'elles ne le soient pas.

(28) bits'a'níshkǫ́ǫh
éloigner.IPFV.1SG-nager.IPFV

14. Le groupe [disjoint conjoint] est parfois aussi appelé TAM, Mode, Pre-stem.

Je m'éloigne à la nage

(29) bits'a'níSh-'eeł
éloigner.IPFV.1SG-faire_flotter.IPFV
Je m'éloigne en bateau

(30) 'ahéé'níSh-kǫǫh
autour/en_cercle.IPFV.1SG-nager.IPFV
Je nage en rond

(31) 'ahéé'níSh-'eeł
autour/en_cercle.IPFV.1SG-faire_flotter.IPFV
Je rame/navigue en cercle

Les deux bases semblent constituer deux dimensions de variation distinctes. Les formes de surface complètes présentent un très grand nombre de patrons d'alternance distincts, suivant les nombreuses combinaisons d'alternances des deux bases. Lors de l'évaluation, il est donc très fréquent d'avoir besoin d'une combinaison de deux patrons qui n'a pas été vue à l'entraînement, même si chacun a pu être observé indépendamment.

Les exemples (32a) à (32e) présentent quelques alternances entre imperfectif et futur à la première personne du singulier. Dans les verbes de (32a) à (32c), l'alternance de la base$_1$ est la même : à l'imperfectif, la base présente un /i:/ qui alterne avec /ite:/ au perfectif. Dans les exemples (32d) et (32e), l'alternance de base$_1$ est différente : le premier ne présente pas le contraste de longueur, le second voit alterner /i:/ avec /ate:/. Dans les verbes de (32c) à (32e), l'alternance de la base$_2$ est la même : il s'agit d'un contraste entre une voyelle longue avec un ton haut et une voyelle courte sans ton haut. Les alternances de base$_2$ des deux premiers exemples sont différentes.

La combinaison de ces alternances produit cinq patrons distincts rendant compte des formes complètes, listées dans les exemples (33a) à (33e). Considérer chacun comme distinct revient à ignorer les similarités partielles que nous avons relevées. L'alternative évidente, que nous adopterons, consiste à analyser séparément les alternances de base$_1$, données en (34), et de base$_2$, données en (35).

(32) IPFV.1sg ⇌ FUT.1sg

a. yiishjʃʃh ⇌ yideeshjʃʃɬ
/ji:ʃʧʃ:h/ /jite:ʃʧʃ:ɬ/
je projette je projetterai

b. diish'eeɬ ⇌ dideesh'oɬ
/ti:ʃʔe:ɬ/ /tite:ʃʔoɬ/
je pars en bateau je partirai en bateau

c. 'ádiishbááh ⇌ 'ádideeshbah
/ʔáti:ʃpá:h/ /ʔátite:ʃpah/
je me rends gris je me rendrai gris

d. distsóós ⇌ dideestsos
/tistsʰó:s/ /tite:stsʰos/
je commence à porter je commencerai à porter

e. 'iishkááh ⇌ 'adeeshkah
/ʔi:ʃk͡xá:h/ /ʔate:ʃk͡xah/
je pars avec eux je partirai avec eux

(33) a. i:_h ⇌ ite:_ɬ

b. i:_e: ⇌ ite:_o

c. i:_V́: ⇌ ite:_V

d. _V́: ⇌ te:_V

e. i:_V́: ⇌ ate:_V

(34) a. i: ⇌ ite:

b. ɛ ⇌ te:

c. i: ⇌ ate:

(35) a. h ⇌ ɬ

b. e: ⇌ o

c. V́: ⇌ V

Nous proposons donc (Beniamine, Bonami et McDonough 2017) de modéliser les alternances entre formes verbales en navajo comme la conjonction de deux patrons, un pour chaque base. À cette fin, nous segmentons les formes de surfaces en deux ensembles de données, et calculons les patrons sur chacun indépendamment. La base$_2$ coïncidant toujours avec la dernière

syllabe de mot, nous avons dressé une liste de toutes les frontières de syllabe possibles, puis segmenté chaque forme en syllabes, et séparé la dernière syllabe ($base_2$) du reste des formes ($base_1$). Nous rendons compte dans le tableau 2.10 des résultats de l'évaluation de ces patrons. Nous évaluons séparément la généralisation sur chaque base seule, puis en combinaison. Pour évaluer les patrons en combinaison, nous prédisons un patron pour chaque base, et comptons une prédiction correcte si et seulement si les deux patrons sont corrects. Notons que pour la prédiction combinée des deux bases, nous n'indiquons pas de nombre de patrons moyens par paire de cases, puisque la prédiction s'appuie sur chaque ensemble de patrons. Le nombre de patrons moyens par paire de cases est donc d'une part celui des $bases_1$ et d'autre part celui des $bases_2$.

Navajo	Paradigmes uniques	Exactitude moyenne	Nombre de patrons moyen
$Base_1$	802	68.7%	74.7
$Base_2$	1494	72.9%	84.7
Formes entières	2157	41.8%	468.7
$Base_1$ & $base_2$	2157	81.3%	—

Tableau 2.10 – Évaluation pour la segmentation des formes du navajo.

Cette évaluation montre que la généralisation sur chaque base indépendamment est bonne, et que la prédiction sur les formes entières par deux patrons est bien meilleure que la prédiction à partir d'un patron holistique. Le nombre de patrons inférés par paire de cases en moyenne, bien inférieur pour les données segmentées, est également un indicateur de cette amélioration.

Pour donner une image concrète des causes du gain en exactitude documenté dans le tableau 2.10, nous proposons d'observer pour chaque base les patrons obtenus pour l'alternance IPFV.1SG ⇌ FUT.1SG. Le tableau 2.11 présente les patrons de $base_1$ pour ces verbes. Les contextes sont omis pour simplifier la lecture, et sur les 64 patrons trouvés, nous ne présentons que les 26 qui concernent plus de 5 lexèmes. Nous présentons à chaque fois un lexème illustrant l'alternance.

Le patron le plus fréquent est le plus simple formellement : il consiste à ajouter /te:/ avant le

/s/ ou /ʃ/ final de la base (le marqueur de première personne). Tous les autres patrons comportent l'addition au futur soit de /te:/, soit des variantes de /ti/. Le second patron le plus fréquent présente en plus une alternance de longueur pour la voyelle précédente. Notons qu'il existe 10 cas sur le modèle de « 'ahénásbąąs » pour lesquels l'alternance de longueur est inversée. De nombreux patrons présentent également une attraction de cette voyelle vers /i/, en conservant ou non le ton haut. Par ailleurs, les syllabes en /nV́/ et /nV/ sont susceptibles de se réduire à /ń/ ou de disparaître.

Le tableau 2.12 présente les alternances correspondantes pour la base$_2$. L'alternance la plus fréquente est une alternance de longueur avec perte d'un ton haut. L'alternance de longueur existe aussi pour les mots comme « ádíshch'iish » sans altérer la présence ou l'absence de ton haut. Les mots qui instancient le second plus fréquent patron présentent une alternance entre /h/ et /ɬ/ finaux. Le troisième plus fréquent patron est un patron identité. Outre des combinaisons des deux premiers patrons, on trouve également des alternances vocaliques et une alternance entre /t/ et /ɬ/ en finale, et la réduction de /n/ en voyelle nasale.

Comme le montrent les tableaux 2.11 et 2.12, la segmentation des verbes du navajo en deux bases sur un critère phonotactique permet de faire émerger des généralisations familières sur les types d'alternances dans un système qui apparaîtrait sinon remarquablement opaque. Nous pouvons donc conclure que la segmentabilité des verbes du navajo est avérée et constitue une caractéristique cruciale pour analyser ce système.

2.4.1.2 Système accentuel et système segmental en russe

Brown et Hippisley (2012) proposent une classification des noms du russe qui repose également sur la conjonction de deux systèmes distincts. Ils analysent d'une part des alternances affixales, et d'autre part des patrons d'accentuation :

> Le système accentuel nominal du russe fournit une bonne illustration de la façon dont l'information morphologique peut être distribuée à travers le réseau. Elle est paradigmatique (sensible aux propriétés morphosyntaxiques), non susceptible d'être directement associée à des propriétés métriques, et peut être vu comme une

IPFV.1 ⇌ FUT.1	Freq.	Lexème	IPFV.1	FUT.1
_ ⇌ te:_	102	ʾADISBĄĄS	ʔatis	ʔatite:s
V:_ ⇌ Vte:_	83	ʾADAASHJAAH	ʔata:ʃ	ʔatate:ʃ
ní_ ⇌ te:_	39	K'ÍNÍSTSʾIIH	k'ínís	k'íte:s
ni_ ⇌ te:_	30	ʾAHANDINISʾÉÉS	ʔahantinis	ʔahantite:s
_i_ ⇌ tí_ée_	28	ʾADINISHʾNÉÉH	ʔatiniʃ	ʔatitínéeʃ
_i:_ ⇌ tí_ée_	28	ʾÁDIʾNIISHCHʾÍÍD	ʔátiʔni:ʃ	ʔátitíʔnéeʃ
[a:aá:ą́:áą:ąą́]_ ⇌ [i:ií:íį́:įį:]te:_	27	ʾAHASŁÓÓS	ʔahas	ʔahite:s
ná_ ⇌ ńte:_	17	BINÁSHʾÉÉSH	pináʃ	pińte:ʃ
é_ ⇌ íte:_	15	ʾÁDÉSGAAS	ʔátés	ʔátíte:s
á_ ⇌ ite:_	14	HÁÁHÁSHCHʾIISH	há:háʃ	há:hite:ʃ
í_ ⇌ ite:_	13	ʾÁDÍSDÉÉS	ʔátís	ʔátite:s
i:_ ⇌ ate:_	13	ʾIISDIS	ʔi:s	ʔate:s
ji_ ⇌ te:_	10	YISHʾEEŁ	jiʃ	te:ʃ
Vn[áą́]_ ⇌ V:t[e:ę́:]_	10	ʾahénásbąąs	ʔahénás	ʔahé:te:s
a:_ ⇌ te:_	9	NAASHYEED	na:ʃ	nte:ʃ
_i:_ ⇌ ti_e:_	8	HAʾDIISBĄĄS	haʔti:s	hatiʔte:s
_ ⇌ te:_ ; ní_ ⇌ te:_	7	ʾAHÁNÍSGÉÉS	ʔahánís ; ʔahás	ʔaháte:s
a_ ⇌ te:_	7	BITAAʾASHʾAAŁ	pi͡kxa:ʔaʃ	pi͡kxa:ʔte:ʃ
nání_ ⇌ ńte:_	7	NÁNÍSHʾAAH	náníʃ	ńte:ʃ
V:_ ⇌ Vte:_ ; _iʃ ⇌ te:_	7	HAASHDLÓÓSH	ha:ʃ ; haʃiʃ	hate:ʃ
_í_ ⇌ ti_ée_	7	HADÍNÍSHCHÉÉH	hatíníʃ	hatitíneeʃ
_í_ ⇌ ti_e:_	6	BÁʾDÍSHCHʾIISH	páʔtíʃ	páti ʔte:ʃ
ji:_ ⇌ te:_	6	YIISHCHXOSH	ji:ʃ	te:ʃ
aná_ ⇌ á:te:_	6	HANÁSDZIIH	hanás	há:te:s
é_é_ ⇌ í_ite:_	6	CHʾÉHÉSDZį́į́S	ʧ'éhés	ʧ'íhite:s
ni_i_ ⇌ _te:_	6	NINISBĄĄS	ninis	nte:s

TABLEAU 2.11 – Quelques patrons d'alternance de la base$_1$ des verbes du navajo.

ipfv.1 ⇌ fut.1	freq.	Lexème	ipfv.1	fut.1
[á:ą́:é:í:ó:ǫ́:[:ę́:]_ ⇌ [aeioąęįǫ]_	108	ʾÁDIISHBÁÁH	pá:h	pah
h ⇌ ɬ	75	ʾADIISHDǪǪH	tǫ:h	tǫ:ɬ
_ ⇌ _	69	ʾÁDIISHGISH	kiʃ	kiʃ
[a:e:i:o:á:ą́:é:í:ó:ǫ́:ą:ę:[:į:ǫ:ę́:]_ ⇌ [aeioáéíóąęįǫ́ǫą]_	58	ʾÁDÍSHCHʾIISH	ʧʼi:ʃ	ʧʼiʃ
[-lat +voi -hto +son]h ⇌ [-lat +hto +voi +son]ɬ	34	ʾADISHʾAAH	ʔa:h	ʔá:ɬ
[a:e:i:o:á:ą́:é:í:ó:ǫ́:ą:ę:[:į:ǫ:ę́:]t ⇌ [aeioáéíóąęįǫ́ǫą]ɬ	22	ʾADIISTSʾǪǪD	ʦʼǫ:t	ʦʼǫɬ
é:_ ⇌ i_	18	ʾÁDÍSDÉÉS	té:s	tis
e:_ ⇌ i_	15	ʾÁDÍSHDLEESH	t͡ɬe:ʃ	t͡ɬiʃ
e:_ ⇌ o_	14	ʾADIISHʾEEŁ	ʔe:ɬ	ʔoɬ
é:_ ⇌ a_	13	ʾÁDÍSHDÉÉH	té:h	tah
t ⇌ ɬ	11	ʾÁDÍSHCHʾID	ʧʼit	ʧʼiɬ
[aeioáéíóąęįǫ́ą] ⇌ [a:e:i:o:á:ą́:é:í:ó:ǫ́:ą:ę:[:į:ǫ:ę́:]ɬ	11	ʾADIISHŁÉ	ɬé	ɬé:ɬ
[á:ą́:é:í:ó:ǫ́:[:ę́:]t ⇌ [aeioąęįǫ]ɬ	10	ʾAHISHʾÁÁD	ʔá:t	ʔaɬ
[aeioáéíóąęįǫ́ą]ʔ ⇌ [a:e:i:o:á:ą́:é:í:ó:ǫ́:ą:ę:[:į:ǫ:ę́:]ɬ	9	ʾAHISHTʾEʾ2	ʨeʔ	ʨe:ɬ
e:_ ⇌ a_	8	ʾAŁKʾIISHJEEH	ʧe:h	ʧah
á:_ ⇌ i_	6	ʾAHISHJÁÁH	ʧá:h	ʧih
in ⇌ [:ɬ	6	ʾÁDÍSHCHIN	ʧʰin	ʧʰ[:ɬ

Tableau 2.12 – Quelques patrons d'alternance de la base$_2$ des verbes du navajo.

couche paradigmatique additionnelle se superposant à la morphologie affixale [15].

Comme pour les bases du navajo, l'identification de ces deux dimensions est évidente pour les locuteurs, car elles coïncident avec des propriétés phonologiques orthogonales.

Brown et Hippisley (2012) comptent quatre classes fondées sur le matériel affixal (tout en reconnaissant l'existence de variations), et huit patrons accentuels, produisant 22 combinaisons attestées sur les 32 concevables.

Nous procédons donc de la même façon que pour le navajo, en constituant un lexique comportant uniquement l'accentuation et un lexique comportant des formes désaccentuées. Nous trouvons 88 microclasses accentuelles et 261 classes segmentales. Nous rendons compte de l'évaluation des patrons obtenus sur ces formes dans le tableau 2.13.

Russe	Paradigmes uniques	Exactitude	Nombre de patrons
Accents	159	62.8%	7.6
Segments	1530	78.5%	22.8
Formes entières	1539	74.7%	37.5
Accents & segments	1539	78.5%	—

Tableau 2.13 – Évaluation pour la segmentation des noms du russe en segments et accents.

La séparation en deux paradigmes augmente ici aussi la qualité de la généralisation des patrons, et mène à un plus petit nombre moyen de patrons par paire de cases. On trouve une exactitude de 78.5% tant pour la prédiction des segments seuls que pour la prédiction combinée des segments et des accents : il s'agit là d'une coïncidence frappante, mais dont il est difficile de tirer des conclusions à priori. Nous présentons dans les tableaux 2.14 et 2.15 l'alternance accentuelle et segmentale entre le nominatif singulier et le datif singulier.

Pour le matériel segmental, on retrouve parmi les patrons les plus fréquents (plus de 5

15. [En anglais dans le texte] « *The nominal stress system of Russian provides a good illustration of the way in which morphological information can be distributed around the network. It is paradigmatic (sensitive to morphosyntactic features), not susceptible to direct metrical assignment, and it can be viewed as an additional paradigmatic layer on top of the affixal morphology* ».

sg.nom ⇌ sg.dat	freq	Lexème	sg.nom	sg.dat
_ ⇌ u	593	Adjutant	adjutant	adjutantu
Ca ⇌ Cʲe	305	Arena	arʲena	arʲenʲe
e ⇌ u	201	Bedstv'ijo	bʲedstvʲije	bʲedstvʲiju
a ⇌ e	119	Agress'ija	agrʲessʲija	agrʲessʲije
_ ⇌ i	107	Bezopasnost'	bʲezopasnostʲ	bʲezopasnostʲi
o ⇌ u	96	Beljo	bʲeljo	bʲelju
o_ ⇌ _u	42	Belok	bʲelok	bʲelku
Cʲe_ ⇌ C_u	20	Amer'ikanec	amʲerʲikanʲets	amʲerʲikantsu
e_ ⇌ _u	8	Avstr'ijec	avstrʲijets	avstrʲijtsu
_ ⇌ _	7	B'uro	bʲuro	bʲuro
a ⇌ eni	6	Im'a	imʲa	imʲeni

Tableau 2.14 – Quelques alternances du russe entre nominatif et datif.

lexèmes) les alternances décrites par Brown et Hippisley (2012). La prise en compte des formes entières plutôt que des seuls affixes mène à distinguer les patrons avec ou sans alternances vocaliques et palatalisation.

le patron le plus fréquent correspond à la classe I de Brown et Hippisley (2012), caractérisée par un /u/ final au datif. Deux variantes de ce patron présentent en plus la disparition d'une voyelle interne à la base, et une variante cumule une palatalisation et la disparition d'un /e/ interne. Le second plus fréquent patron, ainsi que le quatrième, correspondent à la classe II, avec ou sans palatalisation de la dernière consonne. Les alternances e_ ⇌ _u et o ⇌ u correspondent à la classe IV. L'alternance ϵ ⇌ i recouvre leur classe III. Enfin, deux patrons plus rares ne correspondent à aucune de leurs classes flexionnelles : le patron identité et le patron a ⇌ eni.

Dans les données accentuelles, nous avons noté uniquement le nombre de voyelles, et distingué entre voyelles non accentuées et voyelles accentuées. On transcrit donc les alternances de la façon indiquée en (36) :

(36) zakon, nom.sg ⇌ dat.sg

a. /zakón/ ⇌ /zakónu/

b. AÁ ⇌ AÁA

Nous présentons dans le tableau 2.15 les patrons accentuels représentés par plus de 5 lexèmes.

sg.nom ⇌ sg.dat	Freq.	Lexème	sg.nom	sg.dat
⇌ / X+	805	Agress'ija	AÁAA	AÁAA
⇌ A / X+_A ?	647	Adjutant	AAÁ	AAÁA
⇌ A / X*_X+	66	Bik	Á	AÁ

Tableau 2.15 – Quelques alternances accentuelles du russe entre nominatif et datif.

Les deux patrons les plus fréquents correspondent à un accent fixe, avec ou sans ajout d'une syllabe au datif. D'après la classification de Brown et Hippisley (2012), ils correspondent à la classe d'accents A, B ou C. Le troisième patron consiste à ajouter une syllabe au datif, celle-ci prenant alors l'accent. L'accent reste ainsi toujours sur la dernière syllabe. Cela correspond à la classe B de Brown et Hippisley (2012). Il existe cinq autres patrons plus rares. Nos classifications diffèrent, car d'une part, notre exemple porte sur une paire de cases toutes deux au singulier, tandis que la classification de Brown et Hippisley (2012) prend en compte le pluriel ; et d'autre part parce qu'ils raisonnent en termes de radical et d'exposants, tandis que notre classification dépend du nombre de syllabes. Ainsi les deux alternances (37) correspondent respectivement aux classes A et B pour Brown et Hippisley (2012), car dans le premier cas l'accent est constant sur le radical, tandis que dans le second cas, il est constant sur l'affixe. Puisque notre programme ne fait pas de distinctions de ce type, il voit ces exemples comme deux instances du patron ϵ ⇌ / X+, c'est-à-dire d'une accentuation identique au nominatif et au datif singuliers.

(37) nom.sg ⇌ dat.sg

a. /bolót-o/ ⇌ /bolót-a/

b. /čert-á/ ⇌ /čert-é/

Au regard de l'évaluation de ces patrons, il apparaît qu'en russe, comme en navajo, il est

bénéfique de considérer deux types d'alternances distincts qui se combinent pour relier entre elles les formes de surface, plutôt que d'inférer des patrons holistiques.

2.4.1.3 Système tonal et système segmental en chatino

Campbell (2016) décrit la distinction entre morphologie affixale et morphologie tonale en chatino de Zenzontepec dans des termes très similaires à ceux employés par Brown et Hippisley (2012) pour le russe :

> Le système flexionnel en temps/aspect/mode du chatino de Zenzontepec est particulièrement complexe, car il met en jeu simultanément deux couches orthogonales, le système préfixal et les alternances tonales [16].

Campbell (2011) propose la première classification des verbes du chatino de Zenzontepec en classes flexionnelles. Il compte huit classes fondées sur le matériel affixal, regroupées en quatre macroclasses (A, B, C, D) qui correspondent à la description des classes du zapotec par Kaufman (1989). Nous présentons un lexème de chaque classe dans le tableau 2.16 [17]. Campbell (2011) reconnaît de la variation au sein des huit classes, qui ne sont pas exactement des microclasses, remarquant par exemple que « quelques verbes de la sous-classe Au semblent irréguliers, mais leur irrégularité peut être expliquée [18] ».

Campbell (2011) cite 10 classes fondées sur les tons, et montre qu'au sein d'une classe d'affixes, plusieurs patrons tonaux sont représentés. Le tableau 2.17 donne à titre d'exemple les patrons tonaux trouvés pour la classe d'affixes Bc.

De la même façon, pour une classe tonale donnée, on trouve diverses classes d'affixes. Le patron tonal le plus répandu à travers les classes consiste à conserver un ton bas uniforme pour les quatre aspects. Le tableau 2.18 en présente quelques exemples.

16. [En anglais dans le texte] « *The system of TAM inflection in Zenzontepec chatino is quite complex because there are two orthogonal layers, the prefixal system and the tone alternation system, simultaneously at play* ».

17. Comme il est de tradition dans la description des langues oto-mangues, les tons sont notés par un nombre en exposant immédiatement après chaque voyelle. Contrairement à Campbell (2011), nous notons explicitement par un « ^{0} » le ton bas.

18. [En anglais dans le texte] « *[s]ome subclass Au verbs appear irregular, but their irregularity can be explained* ».

Class	cpl	pot	hab	prog
Ac	nka⁰se⁰su⁰	ki⁰se⁰su⁰	nti⁰se⁰su⁰	nte⁰se⁰su⁰
Au	nka¹ra²	ku¹ra²	ntu¹ra²	nte¹ra²
A2	nkwi¹so²ʔ	ki⁰so¹ʔ	nti⁰so¹ʔ	nte⁰so¹ʔ
Bc	nku⁰hna²	ki⁰hna¹	nti⁰hna¹	nte¹hna²
Bt	nku⁰tye⁰hna¹	tye⁰hna¹	ntye⁰hna¹	nte⁰tye⁰hna¹
By	nkya²na¹	cha⁰na	ncha⁰na	nte⁰ya²na¹
Ca	ke²ʔ	ka¹ke²ʔ	nti¹ke²ʔ	ncha⁰ke¹ʔ
C2	ya⁰ku⁰	ka⁰ku⁰	nta⁰ku⁰	ncha⁰ku⁰

Tableau 2.16 – Classes flexionnelles des verbes du chatino de Zenzontepec d'après Campbell (2011).

cpl	pot	hab	prog
ki⁰nya⁰xę⁰ʔ	nti⁰nya⁰xę⁰ʔ	nte⁰nya⁰xę⁰ʔ	nku⁰nya⁰xę⁰ʔ
ki⁰la⁰kwa¹	nti⁰la⁰kwa¹	nte⁰la⁰kwa¹	nku⁰la⁰kwa¹
ki⁰ka⁰ʔne⁰	nti⁰ka⁰ʔne⁰	nte⁰ka²ʔne¹	nku⁰ka²ʔne¹
ki⁰ki⁰tę¹ʔ	nti⁰ki⁰tę¹ʔ	nte⁰ki¹tę²ʔ	nku⁰ki¹tę²ʔ
ki⁰su⁰	nti⁰su⁰	nte⁰su¹	nku⁰su¹
ki⁰ti⁰ta⁰	nti⁰ti⁰ta⁰	nte⁰ti⁰ta⁰	nku⁰ti⁰ta¹

Tableau 2.17 – Les patrons tonaux de la classe affixale Bc .

Class	cpl	pot	hab	prog
Ac	nka^{o}xi^{o}ti^{o}	ki^{o}xi^{o}ti^{o}	nti^{o}xi^{o}ti^{o}	nte^{o}xi^{o}ti^{o}
Au	nka^{o}xi^{o}kwą^{o}	ku^{o}xi^{o}kwą^{o}	ntu^{o}xi^{o}kwą^{o}	nte^{o}xi^{o}kwą^{o}
Bc	nku^{o}ki^{o}ʔi^{o}	ki^{o}ki^{o}ʔi^{o}	nti^{o}ki^{o}ʔi^{o}	nte^{o}ki^{o}ʔi^{o}
By	nkya^{o}ti^{o}ʔ	cha^{o}ti^{o}ʔ	ncha^{o}ti^{o}ʔ	nteya^{o}ti^{o}ʔ
C2	ya^{o}la^{o}ʔ	ka^{o}la^{o}ʔ	nti^{o}la^{o}ʔ	ncha^{o}la^{o}ʔ

Tableau 2.18 – Classes affixales pour le patron tonal bas uniforme.

Nous divisons les paradigmes du chatino de Zenzontepec en deux ensembles de données : l'un ne comporte que les informations tonales, l'autre ne comporte que les informations segmentales. L'évaluation des patrons (tableau 2.19) sur les segments seuls présente un score correct tandis que l'évaluation sur les tons révèle un score plutôt bas. Cela est lié au nombre de paradigmes tonaux distincts, 75, ce qui est très petit pour une évaluation croisée. Dans les deux cas, on trouve un petit nombre moyen de patrons d'alternance, ce qui est le signe d'une bonne généralisation. L'évaluation combinée des patrons de segments et des tons, comparée à l'évaluation des patrons sur les formes entières, présente une amélioration de plus de dix points.

Il n'existe pas d'étude similaire à celle de Campbell (2016) pour le chatino de Yaitepec. Rasch (2002) propose qu'« en sus de phonèmes segmentaux qui apparaissent au début des verbes fléchis, la morphologie aspectuelle implique des contrastes tonaux [19] ». Il remarque également que les contrastes tonaux sont parfois les seules marques aspectuelles. Par ailleurs, il commente la difficulté à prédire certaines formes : « Il ne semble pas y avoir de patron suffisamment fort pour permettre à un apprenant du chatino de prédire quel allomorphe de l'aspect potentiel sera utilisé pour une racine verbale donnée ; la sélection correcte doit être mémorisée [20] ». Cette information, combinée à la petite taille de nos données (324 verbes) pourrait expliquer les très faibles scores (35%) obtenus sur les formes entières. Afin de tester si l'existence de deux

19. [En anglais dans le texte] « *in addition to segmental phonemes that occur at the beginnings of inflected verbs, aspectual morphology involves tone contrasts* ».

20. [En anglais dans le texte] « *There appear to be no patterns strong enough to allow someone learning Chatino to predict which Potential Aspect allomorph will be applied to a given verb root; the correct selection must be memorized* ».

systèmes peut en être également la cause, nous évaluons des patrons tonaux et segmentaux distincts, comme en chatino de Zenzontepec (tableau 2.19). Comparée à l'évaluation sur les formes entières (35%), l'évaluation menée en combinant les patrons segmentaux et tonaux est plus de 20% meilleure. Ceci suggère fortement que, si la mauvaise qualité des généralisations est en partie attribuable au manque de données, ce problème peut largement être contourné en traitant tons et segments comme orthogonaux.

Zenzontepec	Lexèmes	Exactitude	Nombre de patrons
Tons	75	40.6%	15.7
Segments	370	68.6%	24.1
Formes entières	392	56.9%	49.9
Segments & Tons	392	70.4%	—

Yaitepec	Lexèmes	Exactitude	Nombre de patrons
Tons	130	55.2%	21.1
Segments	284	50.5%	67.7
Formes entières	324	35.5%	112.3
Segments & Tons	324	57.9%	—

Tableau 2.19 – Évaluation de la séparation des données en chatino.

Le tableau 2.20 présente les patrons segmentaux du chatino de Zenzontepec pour l'alternance entre complétif et potentiel. Comme précédemment, nous omettons les contextes, et ne présentons que les patrons instanciés par au moins six lexèmes. Remarquablement, les trois patrons les plus fréquents ont des effectifs assez proches. Par fréquence décroissante, les patrons du tableau correspondent aux classes Au, Bc, A2, By, Bt, C2,Au, Ca et Bt d'après Campbell (2011). Le dernier patron est irrégulier et non répertorié par Campbell (2011). Les variantes de Bt et Au sont dues à des effacements vocaliques.

Pour comparaison, nous présentons dans le tableau 2.21 l'alternance de première personne

CPL ⇌ POT	Freq.	Lexème	CPL	POT
n_a ⇌ _u	86	U^{0}SA^{1}NA^{2}	nkasana	kusana
n_u ⇌ _i	70	KI^{1}į^{2}	nkukiį	kikiį
nkw ⇌ k	61	I^{0}TYU^{0}SU^{1}ʔ	nkwityusuʔ	kityusuʔ
nky ⇌ ch	49	YA^{2}HA^{1}	nkyaha	chaha
nkut ⇌ ty	23	TA^{2}TZA^{0}	nkutatza	tyatza
y ⇌ k	23	U^{1}TĘ^{2}	yutę	kutę
nka ⇌	18	U^{1}TU^{2}KWI^{0}	nkatukwi	tukwi
n_a ⇌ _i	14	SU^{0}WI^{1}	nkasuwi	kisuwi
n_u ⇌ _a	7	A^{0}TZU^{0}	nkutzu	katzu
nku ⇌	6	TYU^{1}KWA^{2}	nkutyukwa	tyukwa
⇌ ka	6	A^{1}KWI^{2}	kwi	kakwi

TABLEAU 2.20 – Quelques patrons du chatino de Zenzontepec pour l'alternance entre CPL et POT.

entre les deux mêmes modes en chatino de Yaitepec. Il y apparaît que les patrons présentent diverses combinaisons d'un petit nombre de marqueurs, tous consonantiques.

1POT ⇌ 1CPL	Freq.	Lexème	1POT	1CPL
⇌ m̩	71	JKUʔN	hkũʔⁿ	m̩hkũʔⁿ
c ⇌ m̩ɟ	26	TYA	cɔ̃n	m̩ɟɔ̃n
k ⇌ n̩g	20	KILA	kʲlɔ̃n	n̩gʲlɔ̃n
ʃ ⇌ m̩s	19	SKWA	ʃkʷɔ̃n	m̩skʷɔ̃n
kʲ ⇌ n̩gw	16	LAʔ	kʲlɔ̃ʔⁿ	n̩gwlɔ̃ʔⁿ
c ⇌ m̩d	13	TJIN	chĩn	m̩dhĩn
kʷ ⇌ n̩gw	13	WE	kʷɛ̃n	n̩gwɛ̃n
t ⇌ m̩d	12	TA	tɔ̃n	m̩dɔ̃n
k ⇌ j	11	JOʔ	khũʔⁿ	jhũʔⁿ
kʷ ⇌ m̩	10	CHAʔ	kʷʧɔ̃ʔⁿ	m̩ʧɔ̃ʔⁿ
kʷ ⇌ j	8	LA	kʷlɔ̃n	jlɔ̃n
hʲ ⇌ m̩h	7	JKAʔ	hʲkɔ̃ʔⁿ	m̩hkɔ̃ʔⁿ

TABLEAU 2.21 – Quelques patrons du chatino de Yaitepec pour l'alternance entre 1POT et 1CPL.

En somme, l'évaluation des patrons nous indique qu'en chatino, la séparation des paradigmes en deux ensembles de patrons est bénéfique à la prédiction des formes de surface.

Dans la suite de ce travail, nous caractériserons par deux ensembles de patrons le comportement flexionnel des verbes du navajo, des noms du russe, et des verbes du chatino. Le problème que posent ces systèmes bipartites est bien sûr pertinent pour les locuteurs dans le cadre du PCFP : s'ils ne s'appuient pas sur la segmentation pour déterminer les formes inconnues, ils risquent de faire un très grand nombre d'erreurs. Nous investiguerons ce problème plus avant dans le chapitre 3.

2.4.2 Autres systèmes

Cette section présente quelques patrons d'alternance pour les autres systèmes flexionnels étudiés : le français, le portugais européen, l'anglais et l'arabe standard moderne.

En ce qui concerne le français, nous proposons d'observer les alternances entre la première personne du singulier du présent de l'indicatif et la première personne du pluriel du même temps. Nous montrons dans le tableau 2.22 ces patrons d'alternance, en ignorant ceux qui concernent moins de 6 lexèmes, qui sont au nombre de 15. Dans les données du français, les caractères /E/, /O/ et /Ø/ notent des voyelles dont la hauteur a été neutralisée (voir annexe A). La très grande majorité des lexèmes instancie une alternance suffixale en /-5̃ /. Il existe trois types de variantes de cette alternance : l'apparition au pluriel d'une consonne absente du singulier, des alternances morphophonologiques semi-régulières liées à la structure syllabique, et l'existence d'une alternance de radical.

PRS.1SG ⇌ PRS.1PL	Freq.	Lexème	PRS.1SG	PRS.1PL
⇌ 5̃ / X+_	4109	AHANER	aan	aan5̃
⇌ s5̃ / X+[Eai]_	362	ABÂTARDIR	abataʁdi	abataʁdis5̃
[iuy] ⇌ [jwɥ]5̃ / X*C_	279	HABITUER	abity	abitɥ5̃
⇌ j5̃ / X*[-syl][EOaiuyØ]_	107	ABOYER	abwa	abwaj5̃
E_ ⇌ Ø_5̃ / X*[-syl]_[+ant]_	98	HALETER	alEt	alØt5̃
⇌ z5̃ / X*[-syl -nas][EOaiuyØ]_	47	GÉSIR	ʒi	ʒiz5̃
⇌ d5̃ / X*[-nas -haut][Oaã5ʁ]_	45	APPENDRE	apã	apãd5̃
⇌ t5̃ / X*[-haut][EOaijluwyØã5œɥʁɛ̃]_	40	ABATTRE	aba	abat5̃
jɛ̃ ⇌ Øn5̃ / X*[bdfpstvz]_	26	APPARTENIR	apaʁtjɛ̃	apaʁtØn5̃
⇌ v5̃ / X*[EOaijluvwyzØɥʁʒ][EOaijluwyØɥʁ]_	21	ENSUIVRE	ã sɥi	ã sɥiv5̃
ɛ̃ ⇌ Eɲ5̃ / X*[bdflpstvzʁ]_	18	ASTREINDRE	astʁɛ̃	astʁEɲ5̃
ã ⇌ Øn5̃ / X*pʁ_	11	APPRENDRE	apʁã	apʁØn5̃
ɛ̃ ⇌ aɲ5̃ / X*w_	8	ADJOINDRE	adʒwɛ̃	adʒwaɲ5̃
wa ⇌ Øv5̃ / X*[dstz]_	7	APERCEVOIR	apEʁswa	apEʁsØv5̃
E ⇌ Øz5̃ / X*f_	6	DÉFAIRE	dEfE	dEfØz5̃

TABLEAU 2.22 – Patrons inférés en français pour l'alternance entre PRS.1SG et PRS.1PL (au moins 6 lexèmes).

En ce qui concerne l'anglais, nous proposons d'observer l'alternance entre présent et passé (tableau 2.23). Cette fois, il existe une soixantaine de patrons qui s'appliquent à moins de 6 lexèmes. Les trois plus fréquents patrons correspondent à la flexion usuellement dite « régulière ». On peut être surpris que le patron qui suffixe -t ne présente pas une restriction aux seules consonnes non voisées. Cela est dû à des alternances telles que dans le tableau 2.24. Il s'agit là d'un cas typique d'ilôt de prédictibilité, déjà mentionnés en section 2.2.2.1 : il existe un patron général qui n'est pas très fiable (il est applicable à de nombreuses paires de formes qui ne l'instancient pas). Une formulation moins générale du contexte pour la même alternance s'applique à seulement un sous-ensemble des formes qui instancient l'alternance, mais elle est beaucoup plus fiable (après une consonne non voisée, on trouve toujours /t/ pour les verbes réguliers). Nous faisons le choix de décrire ce type d'alternance par un unique patron, car notre algorithme cherche à maximiser les points de similarité possible entre lexèmes. Cette stratégie est distincte de celle d'Albright et Hayes (2002), qui conservaient des patrons plus fins afin de rendre compte de sous-régularités telles que l'ajout de -t après une consonne non voisée.

Les cas dits irréguliers rendent compte d'une très petite proportion des données. On y trouve trois principaux types de patrons : une absence d'alternance, des alternances vocaliques de la base, ou une alternance de voisement de la consonne dentale finale. Nous classifions également à part quelques lexèmes surabondants (patrons séparés par un point virgule).

En ce qui concerne le portugais, nous présentons les patrons issus de l'alternance entre infinitif et deuxième personne du singulier au présent. Ces patrons font alterner les marqueurs /ɾ/ et /ʃ/, et présentent des voyelles différentes en suffixe et des alternances vocaliques liées à un déplacement accentuel au sein des mots [21].

Enfin, nous présentons dans le tableau 2.26 les patrons d'alternance entre le perfectif (passé) et l'imperfectif (présent) troisième personne du singulier masculin en arabe. Les contextes étant particulièrement longs, nous les omettons pour faciliter la lecture des alternances. Les alternances discontinues sont correctement identifiées par le programme, et les alternances vocaliques de longueur sont généralisées lorsque c'est approprié. Le patron le plus fréquent est

21. L'alternance vocalique entre /-ɐ-/ et /-ə-/ est groupée avec celle entre /-ə-/ et /-ɛ-/, car l'analys en traits distinctifs adoptée ici réduit dans les deux cas le contraste à une opposition [+/−arrière].

PRES1S ⇌ PAST	Freq.	Lexème	PRES1S	PAST
⇌ d / X+_	2954	ABANDON	ɛbændɛn	ɛbændɛnd
⇌ ɪd / X+[dt]_	1659	ABATE	ɛbeˑɪt	ɛbeˑɪtɪd
⇌ t / X+C_	1136	FRENCH-POLISH	frɛntʃpɒlɪʃ	frɛntʃpɒlɪʃt
⇌ / X+	37	BESET	bɪsɛt	bɪsɛt
aˑɪ ⇌ əˑʊ / X*[bdlmnprtðθ]_[bdfpstvzðθ]	16	ARISE	ɛraˑɪz	ɛrəˑʊz
ɪ ⇌ ʌ / X*[-syll]_[bdklmnptŋg]k ?	13	CLING	klɪŋ	klʌŋ
iː_ ⇌ ɛ_t / X*[-syll]_[bdlmnpt]_	11	CREEP	kriːp	krɛpt
⇌ d / X+_ ; ⇌ t / X+C_	10	BURN	bɜːn	bɜːnd ;bɜːnt
r ⇌ d / X*[-syll][ɪˑəʊˑəɛɪʊʌiːɜːɔːɛˑəuː]_	10	BEAVER AWAY	biːvɛr	biːvɛd
eˑɪ ⇌ ʊ / X*[stʃʧθ]_k	9	BETAKE	bɪteˑɪk	bɪtʊk
⇌ / X+ ; ⇌ ɪd / X+[dt]_	9	BROADCAST	brɔːdkɑːst	brɔːdkɑːst ;brɔːdkɑːstɪd
ɪ ⇌ æ / X*[-syll]_[bdklmnptŋg]k ?	9	DRINK	drɪŋk	dræŋk
⇌ / X+ ;iː ⇌ ɛ / X*_[dt]	8	BLEED	bliːd	blɛd ;bliːd
ɛˑə ⇌ ɔː / X*[-syll]_	8	BEAR	bɛˑə	bɔː
d ⇌ t / X*[-low -lab][+son -lab]_	8	BUILD	bɪld	bɪlt
iː ⇌ ɛ / X*_[dt]	8	LEAD	liːd	lɛd
ɛ_ ⇌ əˑʊ_d / X*[stθ]_l_	7	FORETELL	fɔːtɛl	fɔːtəˑʊld
⇌ / X+ ;aˑɪ ⇌ aˑʊ / X*[-syll]_nd	7	BIND	baˑɪnd	baˑɪnd ;baˑʊnd
iː ⇌ əˑʊ / X*[-syll]_C	6	BESPEAK	bɪspiːk	bɪspəˑʊk
əˑʊ ⇌ uː / X*[bdkptxðgθ][lr]_	6	BLOW	bləˑʊ	bluː

TABLEAU 2.23 – Patrons inférés en anglais pour l'alternance entre présent et passé (au moins 6 lexèmes).

Lexème	PRES.1S	PAST
BURN	/bɜn/	/bɜnt/
DWELL	/dwɛl/	/dwɛlt/
LEARN	/lɜn/	/lɜnt/
MISSPELL	/mɪsspɛl/	/mɪsspɛlt/
INDWELL	/ɪndwɛl/	/ɪndwɛlt/

TABLEAU 2.24 – Verbes formant un îlot de régularité pour le passé en /t/.

INFINITIVO ⇌ PRESINDIC2	Freq.	Lexème	INFINITIVO	PRESINDIC2
aɾ ⇌ eʃ / X+[-low]_	912	FICAR	fikaɾ	fikeʃ
[eəɨ]_[ailɛ]ɾ ⇌ [ailɛ]_[eəɨ]ʃ / X*_C+_	321	PASSAR	pesaɾ	paseʃ
u_aɾ ⇌ ɔ_eʃ / X*[-low -round]_C+_	177	JOGAR	ʒugaɾ	ʒɔgeʃ
[eəɨ]_eɾ ⇌ [ailɛ]_əʃ / X*_C+_	122	FAZER	fezeɾ	fazəʃ
iɾ ⇌ əʃ / X+[bdfgkmnprstvzɲɾʃʒ]_	107	OUVIR	oviɾ	ovəʃ
[eəɨ]_iɾ ⇌ [ailɛ]_əʃ / [-lat]*_C+_	60	PARTIR	peɾtiɾ	paɾtəʃ
_aɾ ⇌ ɐ_eʃ / X+C_i_	53	NOMEAR	numiaɾ	numɐieʃ
eɾ ⇌ əʃ / X*V[-son]_	46	DEFENDER	dəfẽdeɾ	dəfẽdəʃ
ɾ ⇌ ʃ / X*[-low -nas][aeiɛ]_	38	VER	veɾ	veʃ
u_eɾ ⇌ ɔ_əʃ / [-lat]*[bdfgkmnprstvzɲɾʃʒ]_[bdflmnprstvzɬɲɾʎ]+_	30	DECORRER	dəkureɾ	dəkɔrəʃ
ə_aɾ ⇌ e_eʃ / X*C_[bdfgkmnprstvzɲɾʃʒ]_	20	CHEGAR	ʃəgaɾ	ʃegeʃ
oɾ ⇌ õĩʃ / [-lat]*p_	17	IMPOR	ĩpoɾ	ĩpõĩʃ
uaɾ ⇌ oeʃ / X*[bdfgklprstvzɬɾʃʎʒ]_	17	VOAR	vuaɾ	voeʃ
u_iɾ ⇌ ɔ_əʃ / [-lat]*C_C+_	15	SUBIR	subiɾ	sɔbəʃ
u_aɾ ⇌ o_eʃ / X*[bdfgklprstvzɬɾʃʎʒ]_[mnɲ]_	12	ABANDONAR	ebẽdunaɾ	ebẽdoneʃ
ɐ_ɾ ⇌ a_ʃ / [aeireəɛɨʃʒ]*[-low -lat]*[bdfgkprstvzɾʃʒ]_i_	11	CAIR	keiɾ	kaiʃ
eɾ ⇌ ẽĩʃ / [-lat]*t_	9	TER	teɾ	tẽĩʃ
[eo]_eɾ ⇌ [ɔɛ]_əʃ / [-lat]*_[bdfgklprstvzɬɾʃʎʒ]+_	9	RESOLVER	rəzolveɾ	rəzɔlvəʃ
u_ɾ ⇌ ɔ_ʃ / [bdfgkprstvzɔəɛɾʃʒ]*[begimouõĩũɔəɛiẽ]ʃtɾ_i_	6	CONSTRUIR	kõʃtɾuiɾ	kõʃtɾɔiʃ

TABLEAU 2.25 – Patrons inférés en portugais pour l'alternance entre infinitif et deuxième personne du présent (au moins 6 lexèmes).

IND.PRS.ACT.M.3.S ⇌ IND.PST.ACT.M.3.S	Freq.	Lexème	IND.PRS.ACT.M.3.S	IND.PST.ACT.M.3.S
[jw]a_[ii:]_[uu:] ⇌ [iu]_[aa:]_[aa:]	211	IBTAʾASA	jabtaʔisu	ibtaʔasa
ju_i_u ⇌ _a_a	210	ʾājara	juʔa:ʤiru	ʔa:ʤara
ju_[ii:]_[uu:] ⇌ a_[aa:]_[aa:]	127	ʾabdala	jubdilu	abdala
ja_u ⇌ _a	61	TAʾAkk̲ara	jataʔaxxaru	taʔaxxara
ja__u_u ⇌ _a_a_a	45	baraza	jabruzu	baraza
ja_[uu:]_[uu:] ⇌ _[aa:]_[aa:]	41	bāḥa	jabu:ħu	ba:ħa
ja__u ⇌ _a_a	39	baḥaṯa	jabħaθu	baħaθa
ja_u ⇌ i_a	27	IBTāʿa	jabta:ʕu	ibta:ʕa
ja_i: ⇌ i_a:	25	ittaqā	jattaqi:	ittaqa:
ja__i_u ⇌ _a_a_a	24	jalasa	jaʤlisu	ʤalasa
ja__[uu:] ⇌ _a_[aa:]	23	badā	jabdu:	bada:
ja__a_u ⇌ _a_i_a	23	baʾisa	jabʔasu	baʔisa
ja__i: ⇌ _a_a:	21	banā	jabni:	bana:
ju_i: ⇌ _a:	19	ʾaddā	juʔaddi:	ʔadda:
ja ⇌	18	taṯannā	jataθanna:	taθanna:
ja_[ii:]_[uu:] ⇌ _[aa:]_[aa:]	18	bāʿa	jabi:ʕu	ba:ʕa
ja__i_u ⇌ _a_a_a ;ja__u_u ⇌ _a_a_a	14	ḥaraṯa	jaħriθu ;jaħruθu	ħaraθa
ju_i: ⇌ a_a:	12	ʾaṯnā	juθni:	aθna:
ja__a: ⇌ _a_ija	7	baqiya	jabqa:	baqija
ju:_i_u ⇌ aw_a_a	7	ʾawtara	ju:tiru	awtara

Tableau 2.26 – Patrons inférés en arabe pour l'alternance entre le perfectif et l'imperfectif troisième personne du singulier masculin (au moins 6 lexèmes).

marqué [jw]a_[ii:]_[uu:] ⇌ [iu]_[aa:]_[aa:]. Les formes qui l'instancient présentent bien des alternances vocaliques qui conservent la longueur, mais présente toujours une alternance entre /ja/ et /i/, jamais entre /w/ et /u/. La généralisation nécessaire pour rendre compte des alternances vocaliques a mené à choisir la représentation abstraite et à surgénéraliser l'alternance /j/ /i/ de façon à inclure /w/ /u/. Une version ultérieure du programme pourrait choisir de conserver chaque opération phonologique indépendamment des autres.

2.5 Conclusion

Nous avons présenté un algorithme d'inférence de patrons d'alternance, permettant de rendre compte de l'ensemble des contrastes flexionnels de surface que présentent les formes d'un lexème. Celui-ci repose sur un alignement des formes phonologiquement motivé, et s'appuie sur la comparaison des alternances entre lexèmes pour choisir les règles les plus générales. Il permet également de rendre compte des alternances qui prennent la forme d'opérations phonologiques régulières. Ces patrons sont conçus pour servir d'unité de base dans la morphologie Item et Patron. Nous les évaluons sur une tâche distincte de prédiction (dans un contexte type PCFP). L'évaluation révèle que cette méthode est robuste d'une langue à l'autre, et constitue une amélioration comparativement aux heuristiques précédemment utilisées, qui consistaient à employer des alignements fixes, décidés en fonction des données.

Le programme d'extraction des patrons d'alternance nous a permis de trouver des patrons flexionnels linguistiquement pertinents dans les différentes langues étudiées, quel que soit le type d'exponence qu'elles présentent. Il rend compte en utilisant le même formalisme de toutes les alternances de surface, qu'elles soient affixales ou qu'elles concernent les bases, qu'elles soient concaténatives ou non, ou qu'elles consistent en des opérations morphophonologiques régulières ou non. La structuration du processus en trois étapes (alignement, généralisation, sélection) permet d'optimiser les patrons à la fois localement à une paire de formes et globalement à l'ensemble d'un sous-paradigme (paire de cases), rendant le processus robuste aux opacités locales. Puisqu'ils rendent compte d'alternances de surface, les patrons sont très précis, et généralement plus nombreux que les systèmes de marqueurs décrits par les grammaires formelles de ces langues.

Pour les systèmes habituellement décrits comme formés de deux systèmes orthogonaux, nous avons segmenté les lexiques en deux dimensions, et appris des patrons d'alternance sur chaque partie indépendamment. Inférer les patrons d'alternances à partir des formes de surface produit, pour chaque lexème, un (long) vecteur de patrons indexés par paire de cases. Toutes les paires sur les cases de paradigmes existantes sont considérées. Le vecteur de patron caractérise le comportement flexionnel d'un lexème.

Ces patrons constitueront les unités de comportement flexionnel sur lesquelles nous nous fonderons dans ce travail pour étudier la similarité entre paradigmes flexionnels.

Chapitre 3

Prédictibilité des propriétés flexionnelles : le PCFP

Depuis Ackerman, Blevins et Malouf (2009), l'étude quantitative de la structure des paradigmes de flexion s'est concentrée sur l'évaluation du problème de remplissage des cases de paradigme (PCFP, pour *Paradigm Cell Filling Problem*), que ces auteurs définissent ainsi :

> **Problème de remplissage des cases de paradigme** : Qu'est-ce qui autorise des inférences fiables concernant les formes de surfaces fléchies (ou dérivées) d'un lexème [1] ?

Ackerman, Blevins et Malouf (2009) et Ackerman et Malouf (2013) proposent de quantifier le PCFP au moyen d'une mesure d'entropie au sein des paradigmes flexionnels. Dans ce contexte, Bonami et Boyé (2014) ont conçu les patrons d'alternances comme un outil permettant de fonder les mesures sur des alternances entre mots plutôt qu'entre affixes. Les travaux de Bonami et Boyé (2014), Bonami et Luís (2014) et Bonami et Beniamine (2016) emploient des patrons suffixaux pour évaluer le PCFP dans les paradigmes du français et du portugais européen. L'algorithme que nous avons proposé au chapitre 2 permet d'appliquer leur méthodologie à tout système pour lequel on peut constituer un lexique flexionnel en transcription phonémique.

Ce chapitre combine et étend les résultats de deux travaux précédents. D'une part, Bonami

1. [En anglais dans le texte] « ***Paradigm Cell Filling Problem**: What licenses reliable inferences about the inflected (and derived) surface forms of a lexical item?* ».

et Beniamine (2016) étendent le modèle proposé par Bonami et Boyé (2014) à la prédiction d'une forme de paradigme à partir de la connaissance de plusieurs autres formes, et présentent les résultats sur les verbes du français et du portugais européen. D'autre part, Beniamine, Bonami et McDonough (2017) ont argumenté en faveur d'un modèle bipartite des verbes du navajo, sur la base de l'étude du PCFP dans un petit sous-ensemble des cases de paradigme. Ce chapitre présente l'application des méthodologies de Bonami et Boyé (2014), Bonami et Beniamine (2016) et Beniamine, Bonami et McDonough (2017) à l'ensemble des huit systèmes étudiés dans cette thèse. Nous suivons de près l'argumentation de Bonami et Beniamine (2016).

Nous présentons tout d'abord (section 3.1) les propriétés distributionnelles des formes qui donnent lieu au PCFP. Nous décrivons ensuite (section 3.2) la mesure d'entropie proposées par Ackerman, Blevins et Malouf (2009) et son extension aux patrons d'alternances par Bonami et Boyé (2014). Dans la section 3.3, nous présentons les résultats de ce calcul pour l'ensemble des huit systèmes étudiés dans cette thèse, en nous fondant sur les patrons d'alternance inférés au chapitre 2. La section 3.5 présente l'extension de la mesure d'entropie implicative à la prédiction depuis plusieurs formes (Bonami et Beniamine 2016). La section 3.5 présente les résultats de ce calcul sur nos systèmes flexionnels. Enfin, nous discutons (section 3.6) du lien entre la prédictivité n-aire et les parties principales.

3.1 La distribution zipfienne des formes de paradigme

Les locuteurs d'une langue ne sont jamais exposés à l'ensemble des formes des lexèmes qu'ils connaissent. Dans cette section, nous discutons de quelques travaux qui ont établi l'existence de cette propriété distributionnelle des formes fléchies qui donne lieu au PCFP.

Nous présentons dans le tableau 3.1 quelques mesures proposées par Chan (2008, chap. 4). Chan définit la SATURATION paradigmatique d'un corpus pour une catégorie morphosyntaxique donnée comme la proportion du paradigme réalisée en corpus pour le lexème qui maximise cette proportion. Cette mesure offre une borne haute du remplissage des paradigmes en corpus. Le tableau 3.1 établit clairement que dans les langues ayant un paradigme de taille conséquente, aucun lexème n'est attesté dans toutes ses formes. Chan (2008) avance qu'en effet, les données

morphologiques sont doublement éparses. D'une part les lexèmes ont une distribution qui suit la loi de Zipf, selon laquelle la fréquence d'un mot en corpus est inversement proportionelle à son rang, si bien que peu de lexèmes sont très fréquents, et de nombreux lexèmes sont très peu fréquents. D'autre part, la distribution des formes de chaque lexème est également déséquilibrée, avec quelques cases beaucoup plus fréquentes que d'autres.

Corpus	Taille (en millions)	Nombre de cases	Saturation
Anglais	1.2	6	1.000
Suédois	1.0	21	0.667
Basque	0.6	22	0.727
Slovène	2.4	32	0.750
Hébreux	2.5	33	0.697
Catalan	1.7	45	0.733
Espagnol	2.6	51	0.667
Italien	1.4	55	0.855
Tchèque	2.0	72	0.569
Hongrois	1.2	76	0.632
Grec	2.8	83	0.542
Finnois	2.1	365	0.403

Tableau 3.1 – Saturation des paradigmes pour des verbes d'un ensemble de corpus (adapté de Chan 2008, p. 79).

Cette étude n'établit pas cependant la réalité du PCFP. Premièrement, les corpus de 0.6 à 2.8 millions de tokens sont plutôt petits. Une estimation récente de Gilkerson et Richards (2009) fondée sur des enregistrements montre que les enfants américains de moins de 4 ans entendent en moyenne environ 6000 (10e centile) à 20000 (90e centile) tokens par jour prononcés par des adultes. En conséquence, même les enfants les moins exposés auront entendu autour de 9 millions de tokens à quatre ans, ce qui est bien supérieur aux corpus de Chan (2008). L'expérience du système flexionnel d'un locuteur adulte est donc bien plus étendue que les corpus examinés

par Chan (2008), et il est possible que la saturation soit plus élevée à de plus grandes tailles de corpus. Deuxièmement, la saturation mesure un maximum qui concerne uniquement le lexème le plus fréquent, et ne nous renseigne pas nécessairement sur la tendance majoritaire. Enfin, il ne suffit pas de montrer que les locuteurs ont une connaissance partielle des paradigmes pour justifier l'importance du PCFP : dans de nombreuses langues, il existe des cases de paradigme utilisées si rarement que leur prédiction n'est pas un problème fréquent pour les locuteurs.

Afin d'établir la réalité du PCFP, nous avons observé la distribution des formes fléchies dans le corpus *FrWaC* (Baroni et al. 2009), un corpus web de 1.6 milliards de tokens, ce qui est supérieur à l'exposition cumulée d'un locuteur adulte [2]. Le corpus comprend énormément de bruit (erreurs de segmentation, erreurs d'étiquetage), nous restreignons donc notre attention aux verbes documentés dans le lexique *Lefff* (Sagot 2010). En conséquence, nous ne pouvons rendre compte des néologismes récents. La figure 3.1 montre l'évolution du nombre moyen de forme par lexème tandis que l'on progresse dans le corpus. Comme il n'existe pas d'étiqueteur morphosyntaxique du français capable de discriminer les formes orthographiques syncrétiques avec une exactitude satisfaisante, nous comptons les formes orthographiques distinctes, et non directement les cases de paradigme. Les verbes du français présentent 51 cases de paradigme, mais les syncrétismes ramènent cette valeur à environ 36 formes distinctes par verbe en moyenne documentés dans le lexique *Lefff*.

Deux observations se dégagent de la figure 3.1 : d'une part, tandis que la taille du corpus croît, le nombre de formes par lexème en moyenne croît également. En conséquence, un locuteur est continuellement amené à rencontrer des nouvelles formes. D'autre part, même face à de très grands corpus, le nombre de formes par lexème qui sont attestées en corpus reste nettement inférieur au nombre de formes existantes. En français, il existe 12 cases de paradigme qui sont presque entièrement sorties de l'usage (le passé simple de l'indicatif et le subjonctif passé). En les excluant, on trouve toujours au maximum environ 18 formes par lexème sur les 24 formes utiles aux locuteurs. En conséquence, les paradigmes ne sont pas saturés en moyenne.

2. En extrapolant les résultats de Gilkerson et Richards (2009), si on suppose qu'un locuteur entend entre 10000 et 100000 mots par jour, 1.6 milliards de mots correspondent à une exposition cumulée se situant entre 44 et 440 années.

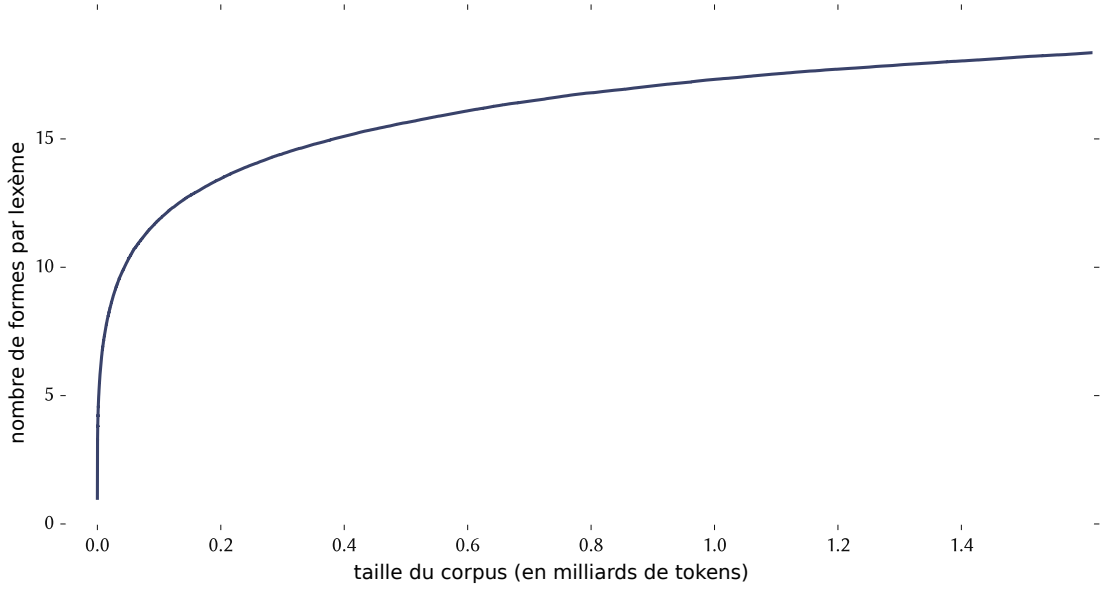


FIGURE 3.1 – Nombre de formes moyen par lexème en fonction de la taille du corpus dans *FrWaC*.

Il résulte de ces deux observations que les locuteurs sont bien confrontés au problème d'inférer des formes inconnues à partir d'autres formes connues.

Blevins, Milin et Ramscar (2017) observent également une distribution zipfienne sur les formes des noms du corpus *SdeWac* (Faaß et Eckart 2013). Les noms de l'allemand ont au maximum quatre cases de paradigme. Sur ce corpus de l'allemand Blevins, Milin et Ramscar (2017) observent qu'avec la croissance du vocabulaire, le nombre moyen de formes observées va jusqu'à décroître. Ils concluent que le PCFP ne disparaît jamais, quelle que soit la taille du corpus. Ils proposent que des pressions entre prédiction (PCFP) et discrimination des formes donne lieu à une dynamique qu'ils synthétisent comme suit :

LA DYNAMIQUE PRÉDICTION-DISCRIMINATION

a. Les systèmes morphologiques présentent des régularités car, en raison de la structure zipfienne des données, les locuteurs ne rencontrent jamais toutes les formes d'une langue et doivent prédire les formes nouvelles à partir d'échantillons incomplets.

b. Les formes irrégulières persistent car elles ont deux fonctions communicatives. En tant qu'expressions individuelles, elles sont bien discriminantes. En tant qu'items exceptionnels au sein d'un plus grand ensemble d'éléments, elles mettent en évidence des contrastes moins marqués dans les patrons réguliers.

c. Les voisinages de formes similaires compensent la pauvreté des échantillons. Tandis que les formes des items individuels ne sont pas toutes attestées, les patrons flexionnels qu'ils suivent sont attestés de façon robuste au sein des voisinages [3].

Le présent chapitre contribue à l'évaluation quantitative de la dernière de ces affirmations : à quel point est-il facile de prédire les patrons flexionnels appropriés au sein d'ensembles de formes similaires ? Cette difficulté varie-t-elle à travers des langues typologiquement variées ? Sur quoi se fondent de telles prédictions ?

3.2 La structure implicative des paradigmes

Wurzel (1989) avance que les paradigmes flexionnels présentent ce qu'il nomme *structure implicative* :

> Les paradigmes flexionnels sont, à proprement parler, maintenus ensemble par des implications. Il n'y a pas de paradigmes (si ce n'est des cas de supplétion extrêmes) qui ne soient pas fondés sur des implications valides au-delà d'un mot unique, si

3. [En anglais dans le texte] *« The prediction-discrimination dynamic*

a. *Morphological systems exhibit regularities because, given the Zipfian structure of the input, speakers never encounter all the forms of a language and must be able to predict new forms from partial samples.*

b. *Irregular formations are persistent because they serve two communicative functions. As individual expressions, they are well discriminated. As exceptional members of larger sets of alternating elements, they emphasize contrasts that are less saliently marked in regular patterns.*

c. *Lexical neighbourhoods compensate for input sparsity. Although the forms of individual items are partially attested, the inflectional patterns that they follow are robustly attested within their form neighbourhoods »*

> bien que nous pouvons dire à juste titre que les paradigmes flexionnels ont généralement une structure implicative, en ignorant les déviations de quelque cas particuliers [4]. Wurzel (1989, p. 114)

On peut décrire ces implications sous la forme de déclarations conditionnelles dans lesquelles l'antécédent mentionne des caractéristiques phonologiques de cases de paradigme, et le conséquent mentionne des caractéristiques phonologiques d'une unique case de paradigme distincte. Nous donnons en (38) quelques exemples pour la conjugaison de l'anglais :

(38) a. Si X est la forme de base d'un verbe, alors son participe présent est Xɪŋ.

b. Si X est la forme de base d'un verbe, alors sa forme de passé est Xd.

c. Si la forme de passé d'un verbe est Xd, alors sa forme de participe passé est Xd.

d. Si la forme de base d'un verbe est Xɪŋ et sa forme de passé Xæŋ, alors sa forme de participe passé est Xʌŋ.

e. Si la forme de base d'un verbe est Xɪŋ et sa forme de participe passé est Xʌŋ, alors sa forme de passé est Xæŋ.

Dans le cas unaire, ces implications sont très similaires à ce que Matthews (1991) nomme « transformations morphologiques » (voir chapitre 1) et elles peuvent s'exprimer sous la forme de patrons d'alternance unidirectionnels comme en (39) :

(39) a. base $\rightarrow$ participe présent : $\epsilon \rightarrow$_ɪŋ/X

b. base $\rightarrow$ passé : $\epsilon \rightarrow$_d/X

c. passé $\rightarrow$ participe passé : $\epsilon \rightarrow \epsilon$/$X$d

La fonction de ces implications est de fournir des stratégies pour pallier au PCFP : c'est pour contrer le caractère lacunaire des connaissances des locuteurs que les paradigmes ont besoin d'être « maintenus ensemble ». Ackerman, Blevins et Malouf (2009) proposent d'évaluer

4. [En anglais dans le texte] « *The inflectional paradigms are, as it were, kept together by implications. There are no paradigms (except highly extreme cases of suppletion) that are not based on implications valid beyond the individual word, so that we are quite justified in saying that inflectional paradigms generally have an implicative structure, regardless of deviations in the individual cases* ».

la cohésion d'un paradigme en quantifiant la robustesse des implications qui lui donnent sa cohésion.

3.2.1 Entropie des paradigmes

Ackerman, Blevins et Malouf (2009) et Ackerman et Malouf (2013) proposent une mesure informationnelle pour quantifier, au sein des paradigmes entiers, la certitude avec laquelle il est possible de prédire une forme fléchie à partir d'une autre forme fléchie. Nous illustrons la façon dont ils calculent l'entropie paradigmatique sur un petit paradigme partiel des noms russes présenté dans le tableau 3.2. Les exemples sont choisis comme représentatifs de trois classes flexionnelles majeures du russe, ainsi que de la classe des indéclinables.

	NOM.SG	NOM.PL
ZAKON	/zakón/	/zakóni/
KARTA	/kárta/	/kárti/
BOLOTO	/bolóto/	/bolóta/
APACHI	/apát͡ɕi/	/apát͡ɕi/

	NOM.SG	NOM.PL
ZAKON	—	/-i/
KARTA	/-a/	/-i/
BOLOTO	/-o/	/-a/
APACHI	—	—

TABLEAU 3.2 – Paradigmes partiels et analyse affixale de quelques noms du russe.

Les nominatifs singuliers sont séparés en trois groupes, selon qu'ils présentent un affixe zéro, qu'ils finissent en /-a/ ou /-o/. Les nominatifs pluriels sont également partagés en trois groupes, selon qu'ils prennent un affixe zéro, /-a/ ou /-i/. Il existe un lien entre les affixes du nominatif singulier et du nominatif pluriel : si le singulier est en /-a/ ou /-o/, le pluriel est respectivement en /-i/ ou en /-a/. Si le singulier présente un affixe zéro, le pluriel peut soit être identique, soit prendre un affixe /-i/. Si l'on suppose que chaque classe a une fréquence identique, cela signifie que si l'on connaît une forme de nominatif singulier, il existe au pire deux possibilités pour former le pluriel, chacune de probabilité $\frac{1}{2}$.

La notion d'entropie conditionnelle permet de quantifier formellement l'incertitude d'un locuteur face à cette prédiction. Rappelons que l'entropie d'une variable aléatoire X se définit comme suit :

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x)$$

Supposons que l'on tire aléatoirement une forme de nominatif singulier russe parmi l'ensemble des noms du tableau 3.2. On peut définir une variable aléatoire qui associe aux formes l'exposant correspondant. Il est possible également de faire de même pour le tirage d'une forme de nominatif pluriel. Nous présentons dans le tableau 3.3 ces variables aléatoires, ainsi que leur probabilités.

NOM.SG	P(NOM.SG)
/-a/	$\frac{1}{4}$
/-o/	$\frac{1}{4}$
—	$\frac{1}{2}$

NOM.PL	P(NOM.PL)
/-i/	$\frac{1}{2}$
/-a/	$\frac{1}{4}$
—	$\frac{1}{4}$

TABLEAU 3.3 – Variables aléatoires pour les exposants du tableau 3.2.

On peut calculer pour chaque distribution son entropie de la façon suivante :

$$\begin{aligned}
H(\text{NOM.SG}) &= -\begin{pmatrix} P(\text{NOM.SG} = \text{/-a/}) \log_2(P(\text{NOM.SG} = \text{/-a/})) \\ +P(\text{NOM.SG} = \text{/-o/}) \log_2(P(\text{NOM.SG} = \text{/-o/})) \\ +P(\text{NOM.SG} = -) \log_2(P(\text{NOM.SG} = -)) \end{pmatrix} \\
&= -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{2}\log_2\frac{1}{2}\right) \\
&= -\left(\frac{1}{4} \times -2 + \frac{1}{4} \times -2 + \frac{1}{2} \times -1\right) \\
&= -1.5
\end{aligned}$$

$$
\begin{aligned}
H(\textsc{nom.pl}) &= -\left(\begin{array}{r} P(\textsc{nom.pl}=-)\log_2(P(\textsc{nom.pl}=-)) \\ +P(\textsc{nom.pl}=\text{/-a/})\log_2(P(\textsc{nom.pl}=\text{/-a/})) \\ +P(\textsc{nom.pl}=\text{/-i/})\log_2(P(\textsc{nom.pl}=\text{/-i/})) \end{array}\right) \\
&= -\left(\frac{1}{4}\log_2\frac{1}{4}+\frac{1}{4}\log_2\frac{1}{4}+\frac{1}{2}\log_2\frac{1}{2}\right) \\
&= -1.5
\end{aligned}
$$

L'entropie $H(\textsc{nom.sg})$ vaut 1.5, de même que $H(\textsc{nom.pl})$, car il existe trois affixes pour chaque, avec les mêmes distributions. Ces entropies mesurent la difficulté à deviner dans l'absolu la terminaison d'une forme de nominatif singulier ou de génitif singulier.

Ce type de tirage revient à deviner une forme sans aucune information préalable, ce qui ne constitue pas un modèle utile du PCFP. En effet, les locuteurs peuvent toujours s'appuyer sur leur connaissance d'une forme de paradigme connue pour informer leur prédiction. Ackerman, Blevins et Malouf (2009) ont donc plutôt recours à l'entropie conditionnelle, qui vise à évaluer l'incertitude restante concernant la valeur d'une variable aléatoire Y lorsque la valeur d'une autre variable X est connue :

$$
H(Y \mid X) = -\sum_{x\in X} P(x) \sum_{y\in Y} P(y \mid x)\log_2 P(y \mid x)
$$

Nous pouvons à présent utiliser l'entropie conditionnelle pour évaluer la difficulté à deviner la forme de NOM.PL d'un lexème lorsque l'on connaît sa forme de NOM.SG. Si le nominatif présente un affixe /-a/ et /-o/, il n'y a chaque fois qu'une possibilité pour le NOM.PL. S'il présente un affixe zéro, il y a deux possibilités. L'entropie est donc de 0.5 :

$$
\begin{aligned}
&H(\textsc{nom.pl} \mid \textsc{nom.sg}) \\
&= -\left(\begin{array}{c}
\frac{1}{2}\left(\begin{array}{c} P(\textsc{nom.pl} = \text{/-i/} \mid \textsc{nom.sg} = -) \log_2(P(\textsc{nom.pl} = \text{/-i/} \mid \textsc{nom.sg} = -)) \\ +P(\textsc{nom.pl} = - \mid \textsc{nom.sg} = -) \log_2(P(\textsc{nom.pl} = - \mid \textsc{nom.sg} = -)) \end{array}\right) \\
+\frac{1}{4}\left(P(\textsc{nom.pl} = \text{/-a/} \mid \textsc{nom.sg} = \text{/-o/}) \log_2(P(\textsc{nom.pl} = \text{/-a/} \mid \textsc{nom.sg} = \text{/-o/}))\right) \\
+\frac{1}{4}\left(P(\textsc{nom.pl} = \text{/-i/} \mid \textsc{nom.sg} = \text{/-a/}) \log_2(P(\textsc{nom.pl} = \text{/-i/} \mid \textsc{nom.sg} = \text{/-a/}))\right)
\end{array}\right) \\
&= -\left(\frac{1}{2}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{4}\left(1\log_2 1\right) + \frac{1}{4}\left(1\log_2 1\right)\right) \\
&= -\left(\frac{1}{2}\log_2\frac{1}{2}\right) \\
&= 0.5
\end{aligned}
$$

Nous obtenons exactement le même résultat pour la prédiction du pluriel vers le singulier. En effet, il existe également une seule possibilité si le singulier est /-a/ ou —, et deux possibilités équiprobables si le singulier est /-i/ :

$$
\begin{aligned}
&H(\textsc{nom.pl} \mid \textsc{nom.sg}) \\
&= -\left(\begin{array}{c}
\frac{1}{2}\left(\begin{array}{c} P(\textsc{nom.sg} = - \mid \textsc{nom.pl} = \text{/-i/}) \log_2(P(\textsc{nom.sg} = - \mid \textsc{nom.pl} = \text{/-i/})) \\ +P(\textsc{nom.sg} = - \mid \textsc{nom.pl} = -) \log_2(P(\textsc{nom.sg} = - \mid \textsc{nom.pl} = -)) \end{array}\right) \\
+\frac{1}{4}\left(P(\textsc{nom.sg} = \text{/-o/} \mid \textsc{nom.pl} = \text{/-a/}) \log_2(P(\textsc{nom.sg} = \text{/-o/} \mid \textsc{nom.pl} = \text{/-a/}))\right) \\
+\frac{1}{4}\left(P(\textsc{nom.sg} = \text{/-a/} \mid \textsc{nom.pl} = \text{/-i/}) \log_2(P(\textsc{nom.sg} = \text{/-a/} \mid \textsc{nom.pl} = \text{/-i/}))\right)
\end{array}\right) \\
&= -\left(\frac{1}{2}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{4}\left(1\log_2 1\right) + \frac{1}{4}\left(1\log_2 1\right)\right) \\
&= -\left(\frac{1}{2}\log_2\frac{1}{2}\right) \\
&= 0.5
\end{aligned}
$$

Ce petit exemple illustre la façon dont l'entropie conditionnelle peut permettre de mesurer la difficulté à prédire une forme de paradigme à partir d'une autre. Ackerman et Malouf (2013) définissent l'ENTROPIE PARADIGMATIQUE d'un système flexionnel comme la moyenne des entropies conditionnelles à travers toutes les paires de cases de paradigme. Cette mesure permet

d'évaluer la difficulté générale, dans un système, à prédire une case sélectionnée aléatoirement depuis une autre case sélectionnée aléatoirement.

Ackerman et Malouf (2013) trouvent, sur un ensemble varié de langues et de systèmes, que l'entropie paradigmatique est beaucoup plus basse que ce qui pourrait être attendu au vu de la diversité des réalisations flexionnelles pour chaque classe. Ils proposent en conséquence la CONJECTURE DE BASSE ENTROPIE, selon laquelle « les paradigmes flexionnels semblent organisés de telle façon que leur entropie conditionnelle moyenne soit relativement basse [5] ». Ces résultats rejoignent l'intuition de Wurzel (1989) selon laquelle les implications « maintiennent ensemble » les paradigmes, mais également l'observation de Carstairs (1987) selon laquelle le nombre de classes flexionnelles est beaucoup plus bas que ce que la combinatoire des réalisations disponibles dans chaque case de paradigme autoriserait. Ackerman et Malouf (2015) montrent également que le No Blur Principle de Carstairs (1987) discuté au chapitre 1 constitue une façon parmi d'autres de maintenir une entropie paradigmatique basse.

3.2.2 Limitations et améliorations

Le modèle de Ackerman, Blevins et Malouf (2009) et Ackerman et Malouf (2013) présente des limitations qui tiennent principalement au fait qu'il s'appuie sur des grammaires descriptives pour calculer l'entropie paradigmatique d'un système. Or ces grammaires fournissent généralement des descriptions simplifiées, dans lesquelles certaines opacités ainsi que certains patrons rares sont omis, ce qui mène à sous-estimer la difficulté du PCFP. Par ailleurs, l'absence d'information de fréquence sur chaque classe est également problématique. Ackerman et Malouf (2013) sont conscients de ces limitations :

> nos calculs d'entropie sont fondés sur des descriptions de systèmes flexionnels tirés de grammaires écrites. Ces descriptions omettent souvent des détails cruciaux, et, de façon plus importante pour déterminer les véritables complexités intégratives des systèmes flexionnels, elles ne reflètent pas les fréquences de type ou de token

5. [En anglais dans le texte] « *inflectional paradigms appear to be organized in such a way that their average conditional entropies are relatively low* ».

> associées avec les classes flexionnelles. [6] Ackerman et Malouf (2013, p. 17)

De plus, Bonami (2014) et Bonami et Boyé (2014) montrent que les segmentations de ces grammaires sont informées par l'ensemble du système, et intègrent des informations qui ne sont pas disponibles aux locuteurs lorsqu'ils doivent résoudre le PCFP. Nous présentons dans le tableau 3.4 un petit sous-paradigme des verbes du français. Dans l'analyse affixale du français, les formes infinitives /finiʁ/ et /diʁ/ prennent un affixe /-ʁ/, tandis que /vəniʁ/ prend un affixe /-iʁ/. Cette segmentation est choisie précisément parce qu'elle permet de réduire l'incertitude dans les paradigmes : tandis que les verbes à l'infinitif en /-iʁ/ forment par exemple leur PRS.1PL en /-ɔ̃/, les verbes en /-ʁ/ présentent un PRS.1PL avec une amplification du radical en /-sɔ̃/ ou /-zɔ̃/.

	INF		PRS.1PL	
Lexème	Mot forme	Affixe	Mot forme	Affixe
VENIR	/vəniʁ/	/-iʁ/	/vənɔ̃/	/-ɔ̃/
FINIR	/finiʁ/	/-ʁ/	/finisɔ̃/	/-sɔ̃/
DIRE	/diʁ/	/-ʁ/	/dizɔ̃/	/-zɔ̃/

TABLEAU 3.4 – Extraits de paradigmes verbaux français.

Or, face à l'infinitif d'un verbe inconnu se terminant en /-iʁ/, les locuteurs n'ont pas de moyen de savoir s'ils sont face à un affixe /-iʁ/ ou /-ʁ/. En conséquence, calculer l'entropie paradigmatique sur des affixes résultant d'une optimisation à l'échelle du paradigme entier mène à sous-estimer la complexité du PCFP.

Enfin, s'appuyer sur des grammaires descriptives pose également problème d'un point de vue typologique. D'une part, il n'est pas garanti que les choix descriptifs faits par les spécialistes de chaque langue étudiée mènent à des descriptions comparables. D'autre part, il n'est alors

6. [En anglais dans le texte] « *our entropy calculations are based on the descriptions of morphological systems culled from written grammars. These descriptions often omit crucial details, and, more importantly for determining the veridical I-complexities of inflectional systems, they do not convey the relative type or token frequencies associated with inflectional classes* ».

pas possible d'étudier la structure paradigmatique de langues peu documentées, pour lesquelles une telle grammaire n'est pas disponible.

En nous appuyant au contraire sur des patrons d'alternance inférés automatiquement à partir de grands jeux de données, nous pouvons fournir une amélioration sur chacun de ces points. Calculés sur un grand nombre de lexèmes, les patrons d'alternance n'omettent aucune opacité à laquelle un locuteur pourraît être confronté. Ils fournissent des informations de fréquence de type pour toutes les microclasses. Ils n'encodent que de l'information locale à chaque paire de cases, rendant compte fidèlement de l'incertitude au sein d'une alternance, sans intégrer d'information provenant du reste du paradigme. Les descriptions qu'ils fournissent sont strictement comparables d'une langue à l'autre s'ils ont été inférés par un procédé identique qui ne présume pas de biais distinct en fonction de la langue (chapitre 2). Enfin, ils sont applicables à n'importe quel lexique disponible au format d'entrée : nous pouvons ainsi calculer la difficulté à résoudre le PCFP pour les verbes du navajo ou du chatino de Yaitepec pour lesquels il n'existe pas de grammaire descriptive complète dans une forme exploitable selon la méthodologie de Ackerman et Malouf (2013).

3.2.3 Entropie implicative

Bonami et Boyé (2014) ont proposé un amendement à la méthodologie de Ackerman, Blevins et Malouf (2009) et Ackerman et Malouf (2013) permettant de mesurer l'entropie moyenne au sein des paradigmes en se fondant sur des patrons d'alternance plutôt que sur des descriptions affixales. Nous présentons ici leur méthodologie sur l'exemple du russe introduit dans le tableau 3.2, et en nous fondant sur la formulation améliorée de Bonami et Beniamine (2016).

Tout d'abord, pour toute paire de cases de paradigme (A, B), définissons une variable aléatoire « $A \rightleftharpoons B$ » qui classe chaque lexème selon le patron d'alternance qu'il instancie pour ces cases. Rappelons que puisque les patrons sont bidirectionnels, « $A \rightleftharpoons B$ » et « $B \rightleftharpoons A$ » sont strictement équivalents. Étant donnée une forme de paradigme de la case A, seul un sous-ensemble des patrons de « $A \rightleftharpoons B$ » lui est applicable. Nous rappelons ci-dessous la définition de l'APPLICABILITÉ d'un patron à une forme (chapitre 2) :

(40) Un patron décrivant une alternance entre les cases A et B est considéré comme AP-

plicable à une forme de la case A si et seulement si la forme remplit la description structurelle de ce patron pour la case A, c'est-à-dire la combinaison de son contexte et du membre A de son alternance.

Afin de caractériser les propriétés morphophonologiques des formes de la case A, nous classons chaque lexème en fonction des patrons qui sont applicables à sa forme A. Nous notons la variable aléatoire correspondante « $A_{A \rightleftharpoons B}$ ». Celle-ci fournit, pour chaque forme de la case A, l'ensemble des patrons parmi lesquels il est possible de choisir pour produire une forme de la case B. Contrairement à la variable aléatoire précédente, notons que « $A_{A \rightleftharpoons B}$ », qui caractérise les formes de la case A, n'est pas équivalent à « $B_{A \rightleftharpoons B}$ », qui caractérise les formes de la case B.

Le tableau 3.5 illustre ces deux variables aléatoires sur l'extrait de paradigme nominal russe déjà présenté. Les patrons présentés sont des simplifications inférées manuellement.

	Lexème	NOM.SG	NOM.PL	NOM.SG $\rightleftharpoons$ NOM.PL	$\text{NOM.SG}_{\text{NOM.SG} \rightleftharpoons \text{NOM.PL}}$
(i)	ZAKON	/zakón/	/zakóni/	_ϵ $\rightleftharpoons$ _i / XC_	{_ϵ $\rightleftharpoons$ _i / XC_}
(ii)	KARTA	/kárta/	/kárti/	_a $\rightleftharpoons$ _i / XC_	{_a $\rightleftharpoons$ _i / XC_ }
(iii)	BOLOTO	/bolóto/	/bolóta/	_o $\rightleftharpoons$ _a / XC_	{_o $\rightleftharpoons$ _a / XC_}
(iv)	APACHI	/apátɕi/	/apátɕi/	ϵ $\rightleftharpoons$ ϵ / XCi	{ϵ $\rightleftharpoons$ ϵ / XCi}

Tableau 3.5 – Variables aléatoires pour la prédiction NOM.SG $\Rightarrow$ NOM.PL dans les noms du tableau 3.2.

Remarquons qu'il n'existe plus d'ambiguïté entre les classes (i) et (iv), dont les nominatifs étaient précédemment analysés identiquement (affixe zéro). En effet, le nominatif consonantique de /zakón/ est discriminé du nominatif en /-i/ de /apátɕi/ par la variable aléatoire $\text{NOM.SG}_{\text{NOM.SG} \rightleftharpoons \text{NOM.PL}}$, et ce même si ces terminaisons ne sont pas morphémiques. Puisque chaque ensemble de $\text{NOM.SG}_{\text{NOM.SG} \rightleftharpoons \text{NOM.PL}}$ ne comporte qu'un unique patron applicable, qui est toujours le patron instancié, il est facile de voir qu'il n'existe aucune incertitude pour inférer le nominatif pluriel à partir du singulier.

Bonami et Boyé (2014) formulent le problème de la prédiction d'une case de paradigme B à

partir d'une case A comme le choix d'un patron de $A \rightleftharpoons B$, étant donnée la connaissance de la classe de patron applicable à la case A. En conséquence, ils mesurent la difficulté du PCFP en se fondant sur l'entropie conditionnelle suivante, qu'ils nomment ENTROPIE IMPLICATIVE UNAIRE DE A À B :

$$H(A \Rightarrow B) = H(A \rightleftharpoons B \mid A_{A \rightleftharpoons B})$$

Dans le cas présent, nous obtenons une entropie implicative de 0, tandis que la méthode de Ackerman, Blevins et Malouf (2009) menait à une entropie de 0.5, car l'analyse affixale ne fournissait pas toute l'information disponible aux locuteurs.

$$H(\text{NOM.SG} \Rightarrow \text{NOM.PL}) = -\left(\frac{1}{4}\log_2 1 + \frac{1}{4}\log_2 1 + \frac{1}{4}\log_2 1 + \frac{1}{4}\log_2 1\right) \qquad = 0$$

	Lexème	NOM.SG	NOM.PL	NOM.SG $\rightleftharpoons$ NOM.PL	NOM.PL$_{\text{NOM.SG}\rightleftharpoons\text{NOM.PL}}$
(i)	ZAKON	/zakón/	/zakóni/	_$\epsilon \rightleftharpoons$ _i / XC_	{_$\epsilon \rightleftharpoons$ _i / XC_, _a $\rightleftharpoons$ _i / XC, $\epsilon \rightleftharpoons \epsilon$ / XCi}
(ii)	KARTA	/kárta/	/kárti/	_a $\rightleftharpoons$ _i / XC_	{_$\epsilon \rightleftharpoons$ _i / XC_, _a $\rightleftharpoons$ _i / XC, $\epsilon \rightleftharpoons \epsilon$ / XCi}
(iii)	BOLOTO	/bolóto/	/bolóta/	_o $\rightleftharpoons$ _a / XC_	{_o $\rightleftharpoons$ _a / XC_}
(iv)	APACHI	/apátʂi/	/apátʂi/	$\epsilon \rightleftharpoons \epsilon$ / XCi	{_$\epsilon \rightleftharpoons$ _i / XC_, _a $\rightleftharpoons$ _i / XC, $\epsilon \rightleftharpoons \epsilon$ / XCi}

TABLEAU 3.6 – Variables aléatoires pour prédire le NOM.SG depuis le NOM.PL pour les noms du tableau 3.2.

Considérons à présent la prédiction dans l'autre directio. Les variables aléatoires correspondantes sont présentées dans le tableau 3.6. Lorsque l'on classe les formes de NOM.PL selon les alternances qu'elles pourraient potentiellement instancier, les trois classes (i), (ii) et (iv), qui présentent un nominatif pluriel en /-i/, sont chacune susceptible d'instancier les patrons {_$\epsilon \rightleftharpoons$ _i / XC_, _a $\rightleftharpoons$ _i / XC, $\epsilon \rightleftharpoons \epsilon$ / XCi}. La classe (iii) se distingue de ces trois classes, car il n'existe qu'une possibilité si le nominatif pluriel est en /-a/. Nous avons donc à présent deux classes de taille différente, l'une déterminant entièrement le choix du patron, l'autre permettant trois patrons :

$$
\begin{aligned}
H(\textsc{nom.pl} \Rightarrow \textsc{nom.sg}) &= -\left(\frac{1}{4}\log_2 1 + \frac{3}{4}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3}\right)\right) \\
&= -\left(\frac{9}{4}\left(\frac{1}{3}\log_2\frac{1}{3}\right)\right) \\
&\approx 1.18
\end{aligned}
$$

Dans cet exemple spécifique, on trouve une entropie implicative de 1.18, supérieure à celle que l' on obtenait par la méthode de Ackerman, Blevins et Malouf (2009), et qui valait 0.5. Ici l'analyse affixale menait à sous-estimer la difficulté de prédire depuis le nominatif pluriel. Face à un nominatif pluriel en /-i/, en l'absence de connaissances supplémentaires, un locuteur ne peut pas déterminer si la forme connue appartient aux classes (i), (ii) ou (iv). Sur cet exemple, nos résultats diffèrent donc de ceux obtenus par la méthode Ackerman, Blevins et Malouf (2009) pour la prédiction dans les deux directions.

3.3 Implications unaires : résultats empiriques

Nous appliquons la méthodologie décrite ci-dessus à l'ensemble des systèmes étudiés dans cette thèse, de façon à déterminer l'entropie implicative pour prédire chaque case de paradigme à partir de chacune des autres cases.

Le tableau 3.7 présente les entropies implicatives unaires au sein des paradigmes étudiés. Les valeurs obtenues confirment la conjecture d'entropie faible. Sagot (2013) remarque que cette mesure tend à être plus haute dans les petits paradigmes, ce qu'il considère comme contre-intuitif. Cette tendance se confirme dans nos données (la valeur la plus haute étant celle du chatino de Zenzontepec). Quoique peu intuitive, nous pensons qu'il s'agit bien d'une propriété du PCFP dans les petits systèmes. Il est possible que dans ces systèmes, le PCFP se pose plus rarement, soit que les paradigmes soient mieux remplis en moyenne [7], soit que les locuteurs puissent en mémoriser une plus grande proportion. Il n'en est pas moins qu'en moyenne, lorsqu'il se pose, le PCFP peut être plus difficile que dans des paradigmes qui présentent de nom-

7. Cependant, ce n'est pas ce que trouvent Blevins, Milin et Ramscar (2017) pour l'allemand.

breuses occasions de syncrétismes ou d'interprédictibilité [8].

Langues	Entropie implicative moyenne
Anglais	0.1785
Arabe	0.3165
Chatino Y.	0.6847
Chatino Z.	0.7030
Français	0.1843
Navajo	0.3862
Portugais	0.1670
Russe	0.6367

Tableau 3.7 – Entropies implicatives unaires moyennes au sein des paradigmes.

Afin d'observer l'ensemble des distributions d'entropies pour toutes les paires de cases des paradigmes, nous présentons dans la figure 3.2 la silhouette lissée de chaque distribution (les lignes en pointillés représentent la médiane et les quartiles de chaque distribution). Ici également, les systèmes se distinguent selon leur taille. Ceux comportant peu de cases, à l'exception de l'anglais [9], présentent des distributions longilignes. Les distributions des paradigmes plus grands sont plus étalés vers le bas, indiquant un grand nombre de cases très interprédictibles, et ce même lorsque la pointe de la distribution est assez haute (arabe, navajo). La silhouette des verbes français, avec trois élargissements, est conforme aux observations de Bonami (2014).

Nous avançons l'hypothèse que dans les paradigmes de grande taille, le « danger » posé par le PCFP est plus grand, et qu'en conséquence, une plus grande prédictibilité est nécessaire

8. Idéalement, on souhaiterait pondérer la moyenne par la fréquence des tokens de chaque case. Il est malheureusement très difficile, même dans les langues bien documentées, d'obtenir des données de fréquence fiables par case. Cela est dû d'une part à une différence entre les fréquences de tokens par case à l'oral et à l'écrit (en particulier dans le texte journalistique qui constitue l'essentiel des corpus richement annotés), et d'autre part aux fréquents syncrétismes orthographiques qui rendent difficile l'usage de données non annotées manuellement

9. Le cas de l'anglais produit une entropie basse due aux grandes zones d'interprédictibilité créées par la surdifférenciation du verbe « to be », à cause duquel nous comptons huit cases de paradigme plutôt que cinq. Cette surestimation serait compensée si la moyenne était pondérée par la fréquence de type des cases.

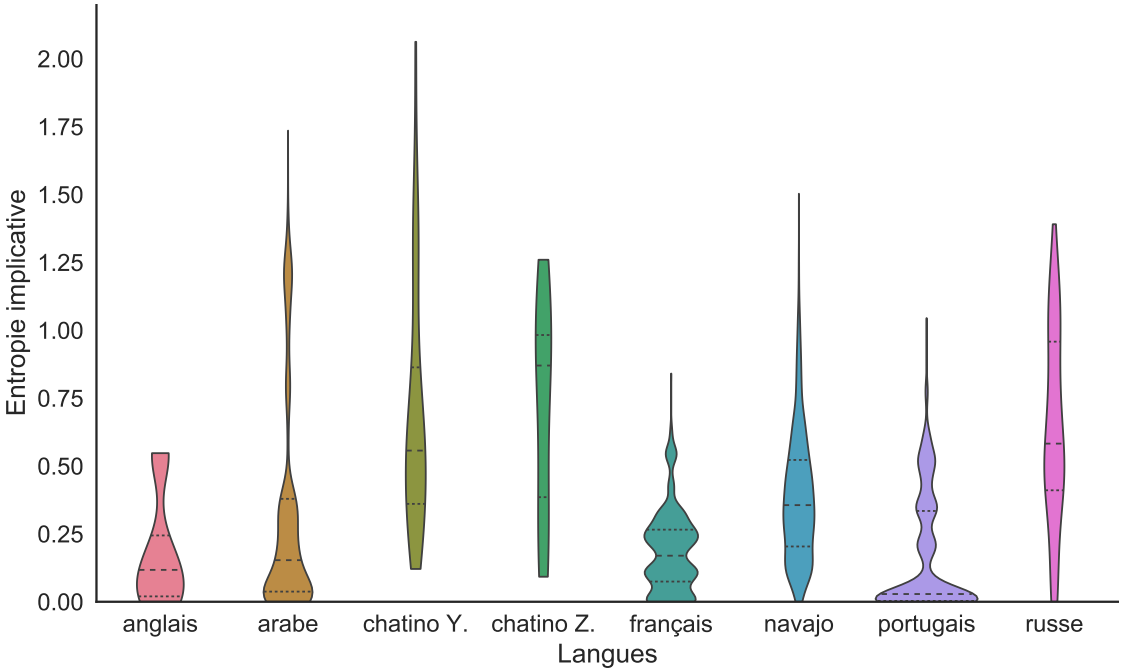


FIGURE 3.2 – Distributions des entropies implicatives pour chaque système flexionnel étudié.

afin de *maintenir ensemble* les paradigmes.

3.3.1 Zones d'interprédictibilité

Bonami (2014) s'appuie sur les entropies implicatives unaires afin de calculer des zones d'interprédictibilité totales entre cases de paradigme : il s'agit d'une partition des cases du paradigmes en zones. Une zone est un sous-ensemble de cases au sein de laquelle chaque paire de cases a une entropie implicative de 0 dans les deux sens. La figure 3.3 présente les zones que nous obtenons pour les verbes du français. Dans cette figure, chaque zone est identifiée par un entier et une couleur. Par exemple, la zone 7, en vert clair, réunit l'ensemble des personnes du conditionnel et du futur.

En français, nous trouvons quatorze zones, soient deux de moins que Bonami (2014). En effet, l'algorithme de Bonami (2014) infère des patrons ayant des contextes phonologiques d'application légèrement différents, ce qui a des conséquences sur la prédictibilité absolue. Les verbes DIRE, REDIRE et DÉDIRE, quoique formés similairement, diffèrent par leur seconde per-

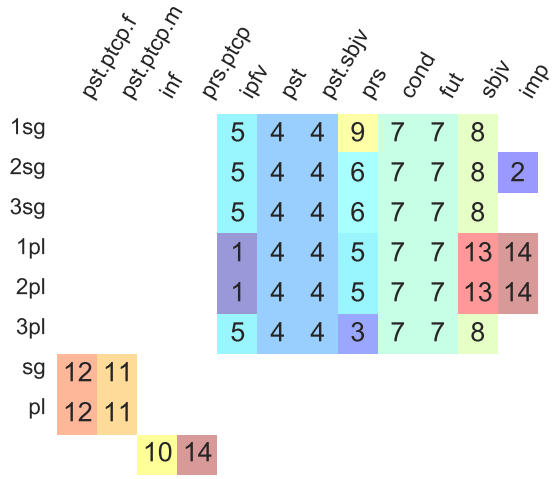


FIGURE 3.3 – Zones d'interprédictibilité dans les paradigmes du français.

sonne pluriel de présent (et d'impératif qui lui est syncrétique), respectivemnt /dit/, /ʁədit/ et /dedize/. Bonami (2014) obtient des patrons Xzɔ̃ ⇌ Xt et Xɔ̃ ⇌ Xe concurrents, impossibles à départager. De même, il existe des verbes comme PESER qui suivent le patron Xɔ̃ ⇌ Xe (/pøzɔ̃/, /pøze/), tandis que FAIRE suit un patron Xøzɔ̃ ⇌ Xɛt (/føzɔ̃/, /fɛt/). Ces deux patrons sont à nouveau applicables aux deux verbes. Pour Bonami (2014), ces patrons sont également impossibles à prédire avec une certitude entière. Notre algorithme produit cependant des patrons dont le contexte est plus précis, présentés dans les exemples 41, 42 et 43).

(41) zɔ̃ ⇌ t / [EOalØʁ]*di_

(42) ɔ̃ ⇌ E / X+_

(43) Øzɔ̃ ⇌ Et / X*f_

Les contextes de ces patrons suffisent à discriminer entre FAIRE et les verbes du type PESER, ainsi qu'entre DÉDIRE et les autres composés de DIRE. Nous obtenons donc que ces cases sont entièrement interprédictibles.

Ces variations dues à des subtilités de l'algorithme d'inférence des patrons nous laissent penser qu'une partition catégorique est trop sensible aux variations dans l'inférence des patrons d'alternance. Il semble raisonnable de penser que deux cases sont interprédictibles en pratique dès lors que l'entropie est extrêmement basse, et que les locuteurs s'accommodent de

mémoriser quelques exceptions qui ne concernent que quelques lexèmes très fréquents. Suivant cette intuition, nous calculons les zones pour lesquelles toutes les paires de cases sont prédictibles avec une entropie inférieure à 0.005. En français, cet ajustement produit des zones identiques à celles de la figure 3.3.

La figure 3.4 présente les treize zones obtenues pour les paradigmes du portugais [10]. L'interprédictibilité y définit de grandes zones de paradigme, à travers les temps et les personnes.

En arabe (figure 3.5), on trouve 35 zones, ce qui constitue une réduction importante à partir des 117 cases de paradigme initiales. On observe en particulier de grandes zones d'interprédictibilité au passif, ainsi qu'entre le jussif et le subjonctif actif. En navajo, les 75 cases de paradigme donnent lieu à 65 zones (figure 3.6). Les interprédictibilités semblent opérer principalement au sein de chaque mode.

En anglais, les zones trouvées réunissent simplement les cases sur-différenciées pour le verbe « to be », c'est à dire d'une part les présents première personne, autres personnes, et l'infinitif, et d'autre part les deux cases du passé. En russe, seuls le datif et le locatif pluriel sont entièrement interprédictibles, et l'instrumental pluriel s'y ajoute si l'on augmente la tolérance à 0.005. Dans les deux paradigmes du chatino, aucune zone d'interprédictibilité n'apparaît.

Le fait que l'interprédictibilité soit beaucoup plus présente dans les paradigmes présentant un grand nombre de cases va dans le sens de l'hypothèse que nous formulions plus haut, selon laquelle les implications paradigmatiques doivent nécessairement être plus fortes au sein de ceux-ci.

En français et en portugais les interprédictibilité réduisent le nombre élevé de cases de paradigmes à quelques très grandes zones qui sont relativement cohérentes d'un point de vue des traits morphosyntaxiques (temps, personnes). Ce sont les seuls systèmes pour lesquels on

10. Nous obtenons une zone de plus que Bonami (2014), selon la partition présentée en figure 3.4. Cette différence est due à l'implémentation de l'alignement des contextes qui produit un alignement suboptimal pour les patrons identité de faible couverture. Par coïncidence, la différence tient à nouveau à la seconde personne du pluriel au présent de l'indicatif et à l'impératif, que Bonami (2014) tient pour interprédictibles, tandis que ce n'est pas le cas dans nos calculs à cause des deux lexèmes REQUERER et QUERER qui sont identiques pour ces cases, respectivement /rəkəɾeiʃ/ et /kəɾeiʃ/. Un alignement plus subtil est implémenté dans les versions plus récentes du script d'inférence des patrons.

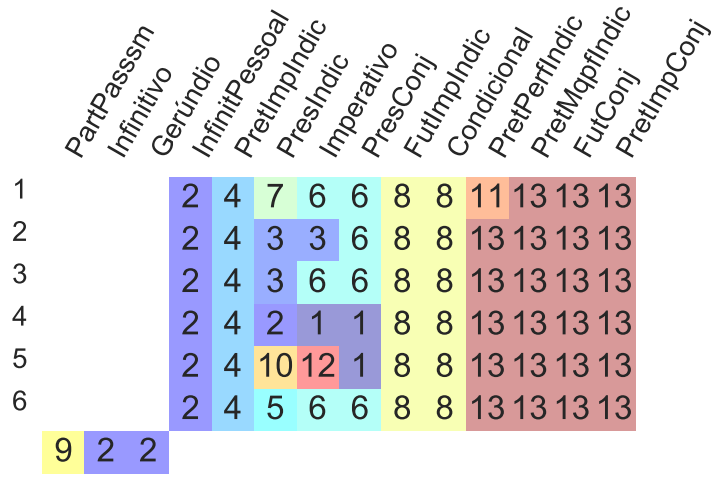


FIGURE 3.4 – Zones d'interprédictibilité dans les paradigmes du portugais ($H \leq 0.005$).

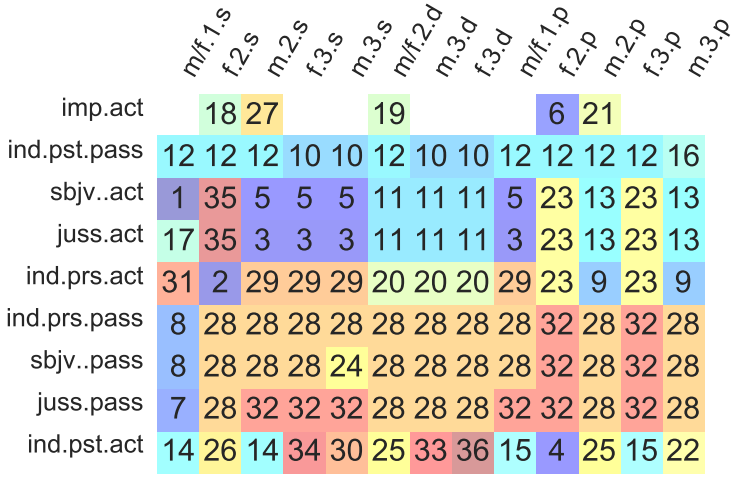


FIGURE 3.5 – Zones d'interprédictibilité dans les paradigmes de l'arabe ($H \leq 0.005$).

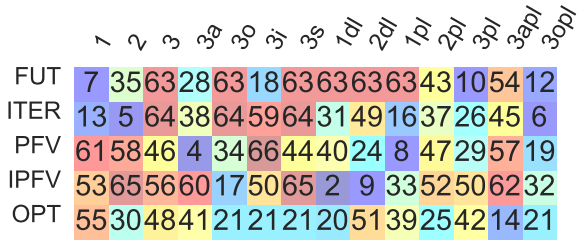


FIGURE 3.6 – Zones d'interprédictibilité dans les paradigmes du navajo ($H \leq 0.005$).

trouve ce type d'organisation. Ce type de situation est celui qui donne sa plausibilité à des analyses fondées sur des « espaces thématiques » (Aronoff 1994 ; Bonami et Boyé 2003a). L'organisation des zones d'interprédictibilité est très différente pour les autres langues observées, ce qui met en doute la pertinence d'une analyse en « espaces thématiques » pour celles-ci.

3.3.2 Détail des entropies implicatives

Les figures 3.7 à 3.14 présentent les entropies implicatives pour les huit systèmes étudiés. Pour rendre les figures plus lisibles, nous nous concentrons sur ce que Stump et Finkel (2013) appellent une DISTILLATION du paradigme, c'est à dire à un ensemble de cases dont chacune est représentative d'une zone d'interprédictibilité. Les cases indiquées en ligne sont les prédicteurs (formes connues), et les cases indiquées en colonne sont les cases prédites (formes inconnues). Les colonnes et les lignes portent donc les mêmes étiquettes, dans le même ordre. Leur affichage est partiel lorsque la place n'est pas suffisante. La couleur indique l'entropie, et va du bleu clair (entropies les plus basses) au violet foncé (entropies les plus élevées). Le premier constat est l'omniprésence du bleu clair dans ces cartes thermiques. Les cases d'entropie haute s'organisent principalement en ligne, c'est à dire qu'elles distinguent des mauvais prédicteurs au sein des paradigmes. En général, les cases concernées ont tendance à être plus prédictibles que prédictrices : c'est le cas par exemple du progressif en chatino de Zenzontepec, dont la colonne est nettement plus claire que la ligne, du IPFV.1PL du français, du PRESINDIC1 du portugais, du datif en russe, des passifs en arabe, et des formes de futur en navajo.

Afin de mieux observer cette relation, nous calculons pour chaque case deux moyennes : nous appelons PRÉDICTIVITÉ d'une case l'entropie moyenne pour prédire chaque autre case à partir de celle-ci, et PRÉDICTIBILITÉ de cette case l'entropie moyenne pour la prédire à partir de chacune autre case. Ces mesures correspondent respectivement aux moyennes en ligne et en colonne dans les cartes thermiques. La figure 3.15 dessine la relation entre prédictivité et prédictibilité pour chaque langue : chaque point représente une case de paradigme, sa position en abscisse correspond à sa prédictivité, et sa position en ordonnée à sa prédictibilité. On observe que dans la plupart des langues, certaines cases sont de beaucoup moins bons prédicteurs que d'autres, et ces cases peu informatives sont systématiquement facile à prédire. Cette relation

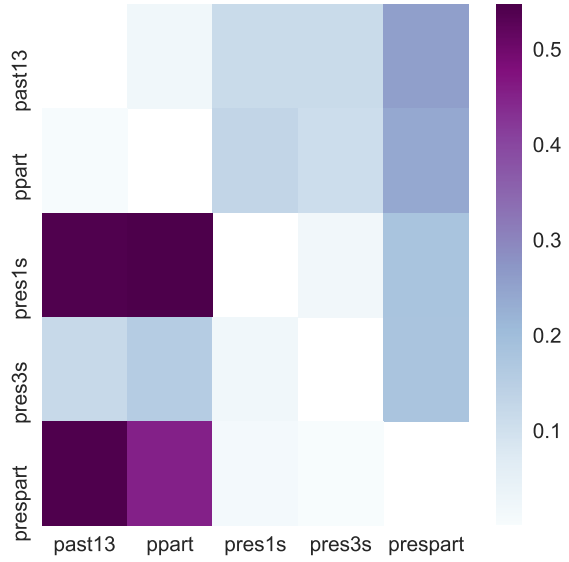


FIGURE 3.7 – Entropies implicatives au sein des paradigmes de l'anglais.

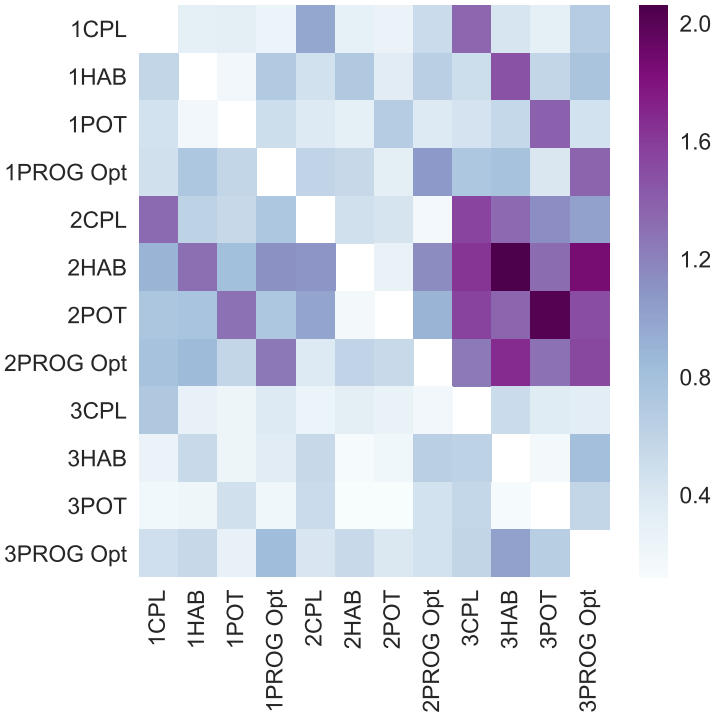


FIGURE 3.8 – Entropies implicatives au sein des paradigmes du chatino de Yaitepec.

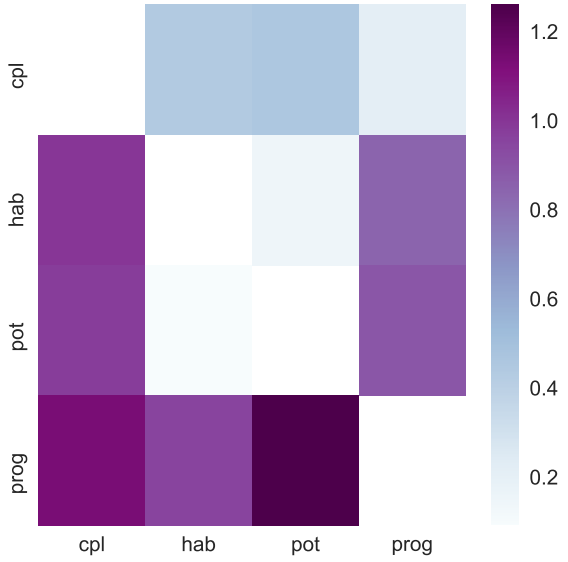


FIGURE 3.9 – Entropies implicatives au sein des paradigmes du chatino de Zenzontepec.

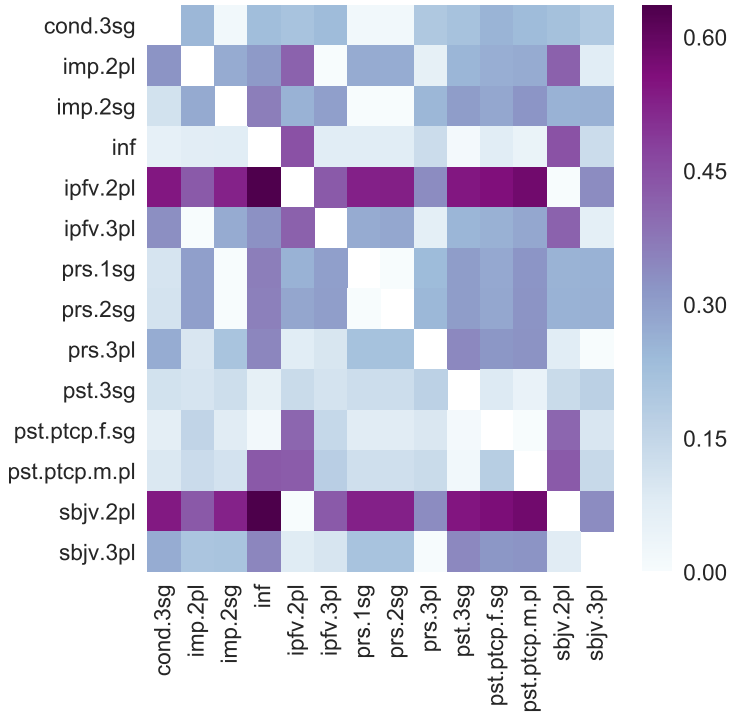


FIGURE 3.10 – Entropies implicatives au sein des paradigmes du français.

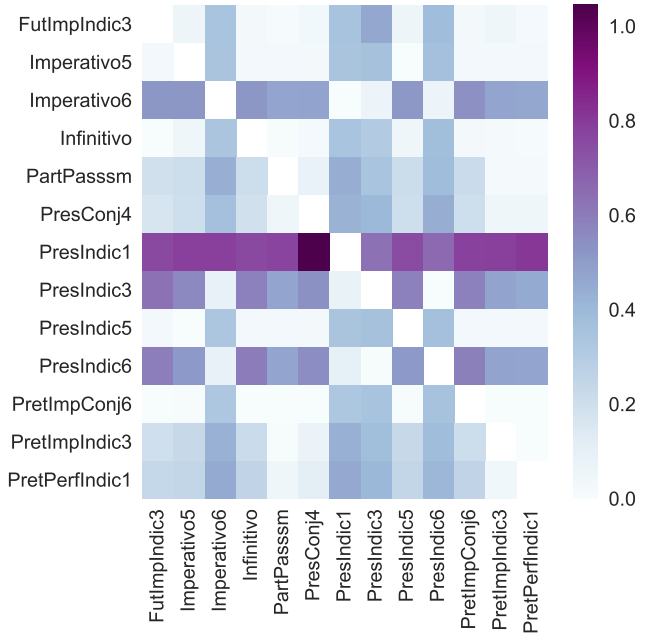


FIGURE 3.11 – Entropies implicatives au sein des paradigmes du portugais.

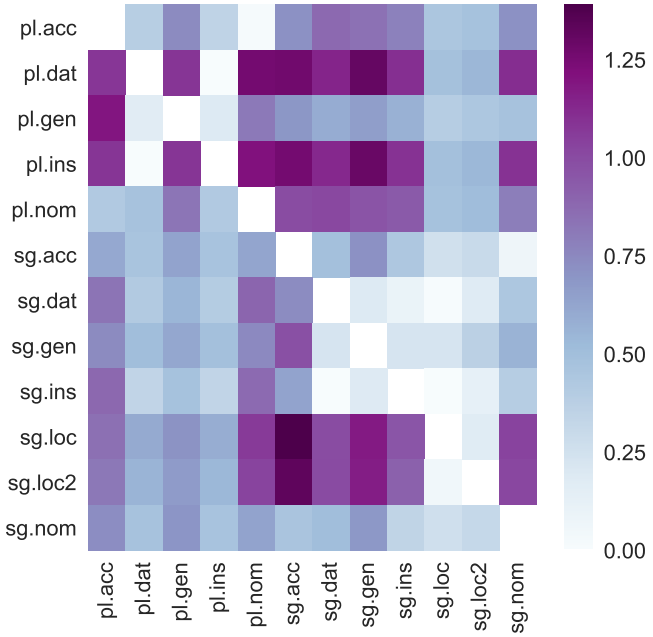


FIGURE 3.12 – Entropies implicatives au sein des paradigmes du russe.

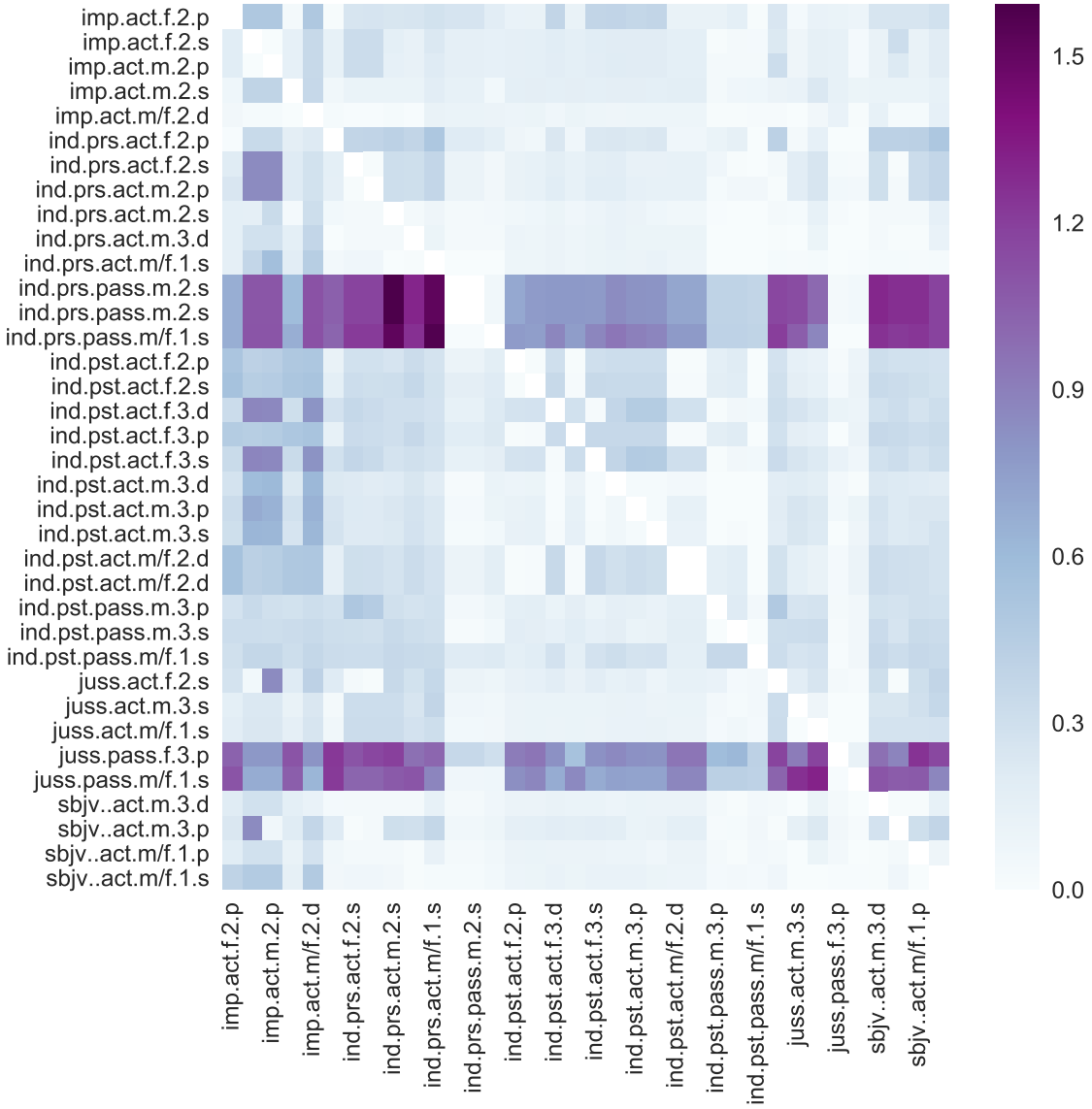


FIGURE 3.13 – Entropies implicatives au sein des paradigmes de l'arabe.

Figure 3.14 – Entropies implicatives au sein des paradigmes du navajo.

ne tient pas pour le portugais, dont l'entropie implicative moyenne est particulièrement basse.

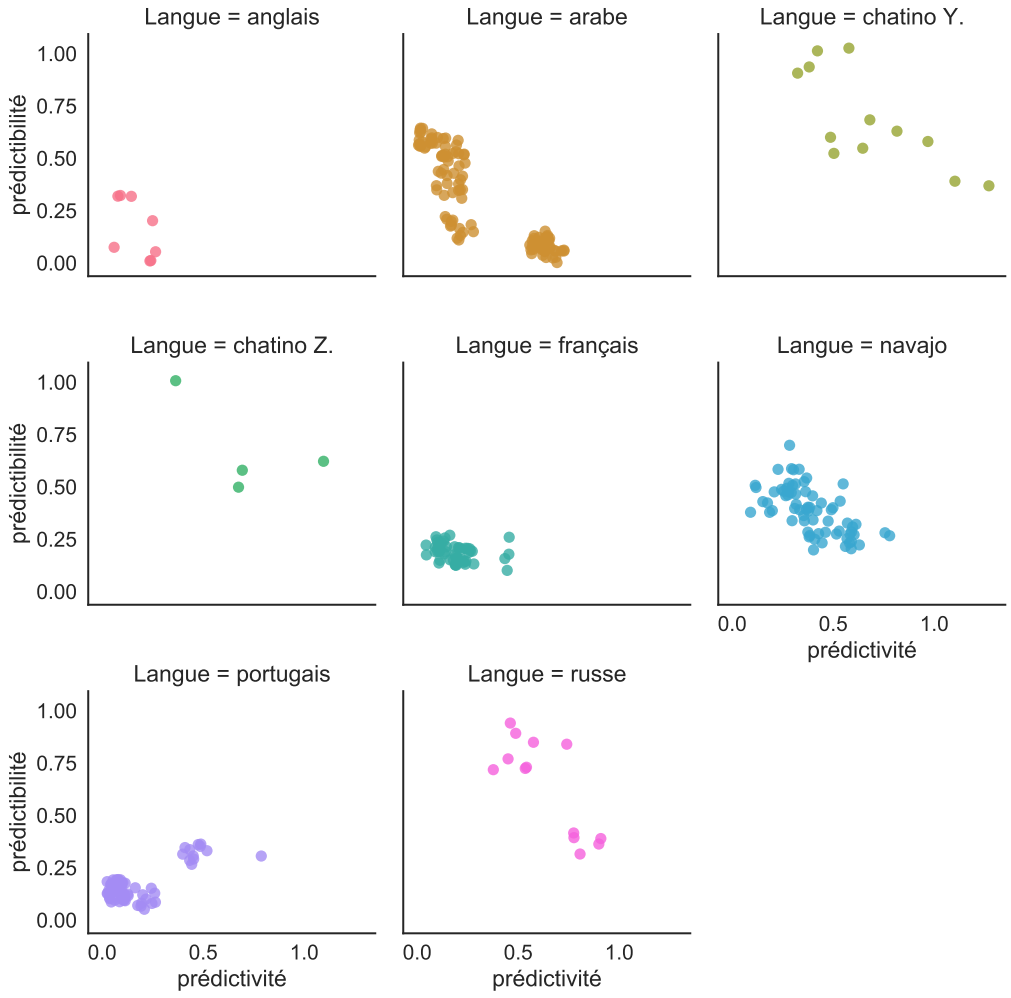


FIGURE 3.15 – Relations entre entropies des cases prédites et prédictrices.

3.3.3 Le cas des systèmes bipartites

Au chapitre 2, avons défini certains des systèmes que nous étudions (verbes du navajo et du chatino, noms du russe) comme bipartites, c'est à dire constitué de deux systèmes distincts. Dans ces systèmes, la prédiction doit se faire conjointement sur les deux systèmes. Prédire une

forme de la case B à partir d'une forme de la case A revient donc à prédire chacune des deux alternances $A_1 \Rightarrow B_1$ et $A_2 \Rightarrow B_2$. Dans les résultats présentés ci-dessus, nous avons employé l'entropie implicative jointe des deux systèmes :

Dans un système bipartite,

$$\begin{aligned} H(A \Rightarrow B) &= H(A_1 \Rightarrow B_1, A_2 \Rightarrow B_2) \\ &= H(A_1 \rightleftharpoons B_1, A_2 \rightleftharpoons B_2 \mid A_{A1 \rightleftharpoons B1}, A_{A2 \rightleftharpoons B2}) \end{aligned}$$

Un locuteur du russe observant un nom au nominatif singulier et voulant prédire un génitif singulier doit prédire à la fois les changements affixaux et le patron accentuel. Pour ce faire, il dispose de façon concomitante d'informations sur chaque dimension de la forme connue.

Afin d'estimer à quel point chaque système est informatif sur l'autre, nous définissons également l'entropie de chaque sous-système. Lorsqu'un locuteur doit prédire la forme de l'un des sous-systèmes il a toujours à disposition également les informations concernant l'autre sous-système pour la forme connue. Par exemple, un locuteur du chatino de Zenzontepec devant prédire la forme segmentale de l'optatif à partir d'une autre forme connaît nécessairement les propriétés tonales et segmentales de cette forme connue. En conséquence, nous définissons : Dans un système bipartite,

$$H(A \Rightarrow B_1) = H(A_1 \rightleftharpoons B_1 \mid A_{A1 \rightleftharpoons B1}, A_{A2 \rightleftharpoons B2})$$

$$H(A \Rightarrow B_2) = H(A_2 \rightleftharpoons B_2 \mid A_{A1 \rightleftharpoons B1}, A_{A2 \rightleftharpoons B2})$$

L'entropie du système entier définie plus haut est l'entropie jointe de ces deux distributions. Nous pouvons évaluer le point auquel chaque système nous informe sur l'autre au moyen de l'information mutuelle $I(A \Rightarrow B_1, A \Rightarrow B_2)$. L'information mutuelle peut se déduire des entropies de chaque distribution et de l'entropie jointe :

$$I(X, Y) = H(X) + H(Y) - H(X, Y)$$

Dans les systèmes bipartites, nous avons défini $H(A \Rightarrow B)$ par les entropies jointes $H(A \Rightarrow B_1, A \Rightarrow B_2)$. On a donc :

$$I(A \Rightarrow B_1, A \Rightarrow B_2) = H(A_1 \Rightarrow B_1) + H(A_2 \Rightarrow B_2) - H(A \Rightarrow B)$$

Cette information mutuelle vaut 0 au minimum, si les systèmes sont indépendants. Suivant Fred et Jain (2003) et Wagner et Wagner (2007), nous normalisons cette mesure de façon à ce qu'elle s'échelonne entre 0 et 1. Nous définissons l'information mutuelle normalisée *NMI* comme suit :

$$NMI(A_1 \Rightarrow B_1, A_2 \Rightarrow B_2) = \frac{2I(A_1 \Rightarrow B_1, A_2 \Rightarrow B_2)}{H(A_1 \Rightarrow B_1) + H(A_2 \Rightarrow B_2)}$$

Cette mesure nous permet de comparer, d'un système à l'autre, l'informativité de chaque sous-système sur l'autre. Le tableau 3.8 compare les valeurs d'entropie implicatives des systèmes bipartites aux entropies implicatives de chaque sous-système. Les valeurs d'information mutuelle indiquent que dans les quatre cas, les sous-systèmes ne sont pas indépendants. En chatino de Zenzontepec et en russe, l'information mutuelle entre les distributions des deux systèmes est particulièrement marquée.

Langue	$H(A_1 \Rightarrow B_1)$	$H(A_2 \Rightarrow B_2)$	$H(A \Rightarrow B)$	I	NMI
Chatino	tons	segments			
de Yaitepec	0.2561	0.5141	0.6847	0.0858	0.2264
de Zenzontepec	0.4209	0.4370	0.7030	0.1546	0.3681
Russe	accents	segments			
	0.4843	0.3017	0.6367	0.1495	0.3200
Navajo	base_1	base_2			
	0.2056	0.2179	0.3862	0.0372	0.1361

TABLEAU 3.8 – Entropies des sous-systèmes dans les paradigmes bipartites.

En conséquence, dans les systèmes bipartites, quoique l'analyse nécessite une forme de ségrégation des données (segmentation dans le cas du navajo, séparation de dimensions segmentales et suprasegmentales en russe et en chatino), la prédiction du PCFP est toujours plus

aisée lorsqu'elle se fait sur les deux systèmes conjointement, c'est à dire sur les mots entiers, que sur chaque sous-système indépendamment.

3.4 Implications *n*-aires

Nous avons étudié la force des implications unaires au sein des paradigmes. Les implications discutées en introduction comprenaient également des déclarations dont l'antécédent se fonde sur plus d'une forme, comme celles en (38d-38e) que nous répétons en (44a-44b)

(44) a. Si la forme de base d'un verbe est Xɪŋ et sa forme de passé Xæŋ, alors sa forme de participe passé est Xʌŋ.

b. Si la forme de base d'un verbe est Xɪŋ et sa forme de participe passé est Xʌŋ, alors sa forme de passé est Xæŋ.

Dans Bonami et Beniamine (2016), nous avons proposé d'étendre la méthodologie de Bonami et Boyé (2014) afin d'évaluer également les implications n-aires. Cette étude nous a permis d'appliquer la méthodologie de Ackerman, Blevins et Malouf (2009) à des questions posées par les travaux de Stump et Finkel (2013).

Nous commençons par justifier le cas des prédictions n-aires dans le cadre du PCFP, puis nous présentons une stratégie permettant d'évaluer l'entropie implicative n-aire sur la base de patrons d'alternances binaires. Enfin, nous appliquons cette stratégie aux huit systèmes étudiés dans cette thèse.

3.4.1 Motivations

Dans cette section, nous motivons la pertinence, pour les locuteurs, de résoudre le PCFP en se fondant sur plus d'une case de paradigme connue. Dans le corpus *FrWaC*, nous avons observé qu'au fil du corpus les paradigmes se remplissent lentement, sans jamais atteindre saturation. Nous en avons conclu que les locuteurs sont continuellement amenés à produire des formes qu'ils n'ont pas encore rencontrées. Nous souhaitons à présent savoir combien de formes d'un même lexème les locuteurs disposent pour faire une telle prédiction.

La figure 3.16 montre la proportion de lexèmes trouvés pour au moins k formes en fonction de la taille du corpus. Le nombre de lexèmes pour lesquels on trouve au moins deux formes atteint une valeur maximale supérieure à 90% dès 100 millions d'occurences, puis demeure constante.

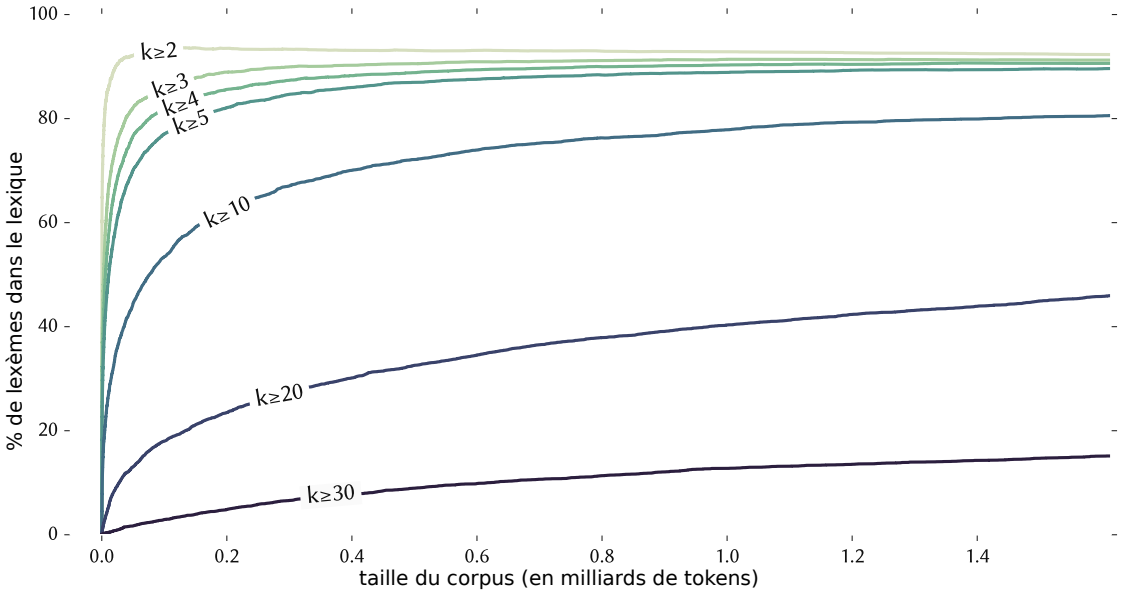


FIGURE 3.16 – Proportion des lexèmes attestés dans au moins k formes en fonction de la taille du corpus dans *FrWaC*, pour diverses valeurs de k.

Cette observation indique que les locuteurs sont massivement exposés à plusieurs formes de la plupart des lexèmes : leur expérience linguistique leur donne les moyens de fonder des inférences sur au moins deux formes d'un même lexème. La même observation tient pour un nombre de formes plus élevées. Cependant, plus on augmente le nombre de formes, plus la croissance et la valeur maximale de la courbe baissent.

Le fait que les locuteurs aient accès à plusieurs formes de lexèmes dont ils ne connaissent pas le paradigme entier ne prouve pas l'utilité de cette information. Il se pourrait que la connaissance d'une seconde forme de lexème n'améliore pas la prédictibilité. Afin d'illustrer la façon dont cette information peut contribuer à une meilleure prédiction, nous nous appuyons sur un petit sous paradigme du français présenté dans le tableau 3.9. Ces données illustrent les alternances principales entre l'infinitif, le présent de l'indicatif et le participe passé des verbes

français.

Classe	Taille	Exemple	INF	PRS.3SG	PRS.3PL	PST.PTCP
(i)	4108	LIVRER ‘deliver’	/livʁe/	/livʁ/	/livʁ/	/livʁe/
(ii)	210	RELIER ‘link’	/ʁəlje/	/ʁəli/	/ʁəli/	/ʁəlje/
(iii)	22	RATISSER ‘rake’	/ʁatise/	/ʁatis/	/ʁatis/	/ʁatise/
(iv)	327	BÂTIR ‘build’	/batiʁ/	/bati/	/batis/	/bati/
(v)	37	TENIR ‘hold’	/təniʁ/	/tjɛ̃/	/tjɛn/	/təny/
(vi)	8	OUVRIR ‘open’	/uvʁiʁ/	/uvʁ/	/uvʁ/	/uvɛʁ/
(vii)	1	MOURIR ‘die’	/muʁiʁ/	/mœʁ/	/mœʁ/	/mɔʁ/

TABLEAU 3.9 – Sous-paradigmes exemplaires de verbes français.

Comme illustré dans ce tableau, le participe passé peut être prédit catégoriquement à partir de l’infinitif pour les verbes dont l’infinitif se termine en /-e/. Cependant, pour les verbes dont l’infinitif est en /-iʁ/, la prédiction est plus difficile. Il existe un patron majoritaire correspondant à la classe (iv), c’est à dire la seconde conjugaison dans la description traditionnelle, mais également une minorité de patrons suivant (v), (vi) et (vii). L’implication en (45) n’est exacte que dans 90% des cas :

(45) Si l’infinitif d’un verbe est /Xiʁ/, alors son participe passé est /Xi/.

Aucune case de paradigme seule ne suffit à fournir une prédiction parfaite du participe passé de ces verbes. Au présent troisième personne, la principale difficulté provient des formes qui se terminent en /-i/ et pourraient appartenir soit à la classe (ii), avec un participe passé en /-je/, soit à la classe (iv), avec un participe passé en /-i/. Au présent troisième personne du pluriel, les formes qui se terminent en /-is/ alternent soit avec une forme de participe passé en /-ise/ (iii), soit en /-i/ (iv). Aucune des autres cases de paradigme présentées dans cette table n’est un prédicteur catégorique du participe passé.

La situation est différente si l’on considère des combinaisons de cases. Supposons qu’un locuteur ait une connaissance jointe de l’infinitif et d’une forme de présent singulier d’un verbe.

Si l'infinitif se termine en /-e/, le participe passé est identique. Sinon, le présent désambiguïse entre les classes (iv), (v), (vi) et (vii). Spécifiquement, l'implication en (46) qui identifie les verbes de la classe (iv) est catégorique.

(46) Si l'infinitif d'un verbe est /Xʁ/ et son présent 3SG est /Xi/, alors son participe passé est /Xi/.

Cet exemple montre qu'il existe parfois des prédictions n-aires utiles dans les systèmes flexionnels. Le besoin de telles prédictions peut même être systématique, comme le montrent Bonami et Luís (2014) et Bonami et Beniamine (2016) pour le portugais européen.

Nous ne connaissons pas d'étude psycholinguistique discutant de l'usage d'implications binaires par les locuteurs. En l'absence d'études expérimentales, Bonami et Beniamine (2016) s'appuient sur l'observation des erreurs de conjugaison des locuteurs afin d'appuyer l'existence de prédiction n-aires dans le comportement des locuteurs. Les erreurs de régularisation sont fréquentes lors de l'acquisition d'une langue native, d'une langue seconde, et elles sont présentes dans une moindre mesure dans la parole des locuteurs natifs adultes. Elles fournissent des indices de l'application de patrons productifs par les locuteurs (par opposition à une mémorisation des formes). Le tableau 3.10 fournit quelques exemples de régularisations fréquentes dans la conjugaison du français, choisis dans la liste établie par Kilani-Schoch et Dressler (2005).

	Lexème	Case de paradigme	Forme correcte	Régularisation
(i)	DIRE	PRS.2PL	/dit/	/dize/
(ii)	FAIRE	PRS.2PL	/fɛt/	/fəze/
(iii)	PRÉVOIR	PST.PTCP	/pʁevy/	/pʁevwaje/
(iv)	OUVRIR	PST.PTCP	/uvɛʁ/	/uvʁi/
(v)	PRENDRE	PST.PTCP	/pʁi/	/pʁɑ̃dy/
(vi)	PEINDRE	PST.PTCP	/pɛ̃/	/pɛ̃dy/
(vii)	MOURIR	PST.PTCP	/mɔʁ/	/muʁy/

TABLEAU 3.10 – Erreurs de régularisations fréquentes dans la conjugaison du français.

Les exemples de DIRE et FAIRE mettent en évidence le fonctionnement de la régularisation.

Ces verbes sont les 3e et 4e plus fréquents d'après *Lexique* (New, Brysbaert et al. 2007), et les seuls verbes à l'exception du verbe *être* à avoir un PRS.2PL en /-t/ plutôt que /-e/. Il n'existe pas de cas documenté de généralisation erronée du /-t/ final aux formes en /-e/. En conséquence, c'est la fréquence de type, et non la fréquence de token, qui détermine les erreurs de régularisation. Les locuteurs régularisent les patrons utilisés par de nombreux lexèmes, non les patrons utilisés par les lexèmes fréquents.

Observons à présent les erreurs de régularisation qui concernent le participe passé. La plupart d'entre elles peuvent être attribuées à des patrons unaires. L'exemple (iii) aligne PRÉVOIR avec les verbes de la première conjugaison qui présentent une alternance -wa ⇌ -waje entre présent et participe passé, plutôt que le véritable patron -wa⇌-y, instancié par très peu de verbes. L'exemple (iv) aligne OUVRIR avec les verbes de la seconde conjugaison qui présentent une alternance -iʁ⇌-i entre infinitif et participe passé, plutôt qu'avec la petite classe de verbes qui suivent un patron -iʁ⇌-ɛʁ. Les exemples (v) et (vi) alignent PRENDRE et PEINDRE sur le patron fréquent -Cʁ⇌-Cy entre infinitif et participe (voir tableau 3.9).

Dans l'exemple (vii), MOURIR présente un très rare participe passé en /-ɔʁ/, en faisant un très bon candidat à la régularisation. Mais la régularisation attendue de cette forme serait /muʁi/ [11], /muʁe/ ou /mœʁe/, selon la case employée pour fonder l'analogie. Les seules cases pouvant donner lieu à une forme /muʁy/ sont le passé simple et le subjonctif passé, mais celles-ci sont trop peu fréquentes pour que cette hypothèse soit défendable. Cependant, il existe une implication binaire très fiable dans le reste du paradigme : la très vaste majorité de verbes dont l'infinitif est en /-iʁ/ et qui ont un présent sans /-i/ final ont un participe passé en /-y/, comme pour COURIR, PRS.1SG /kuʁ/, PST.PTCP /kuʁy/.

3.4.2 Extension de l'entropie implicative au cas à prédicteurs multiples

Afin d'évaluer quantitativement l'utilité des implications n-aires, Bonami et Beniamine (2016) définissent une mesure d'entropie fondée sur les patrons binaires mais évaluant la pré-

11. Kilani-Schoch et Dressler (2005) écrivent que des sources anciennes document une régularisation en /muʁi/. Nous n'en trouvons aucune trace. Dans le corpus *FrWac*, nous trouvons 115 occurrences de *mouru* (/muʁy/), mais aucune de *mouri* (/muʁi/).

diction à partir de plusieurs formes. Ils proposent que l'entropie implicative à partir de cases A et B vers une case C soit :

$$H(A, B \Rightarrow C) = H(A \rightleftharpoons C, B \rightleftharpoons C \mid A_{A \rightleftharpoons C}, B_{B \rightleftharpoons C}, A \rightleftharpoons B)$$

L'intuition qui motive cette définition est la suivante : nous voulons estimer à quel point il est difficile de prédire C à partir de la connaissance de A et B. Nous connaissons donc la forme de A et de B, et donc la valeur des variables aléatoires $A_{A \rightleftharpoons C}$ et $B_{B \rightleftharpoons C}$. Puisque ces deux formes sont connues, le patron qui les lie est également connu. Celui-ci est indiqué par la valeur de la variable aléatoire $A \rightleftharpoons B$. Enfin, il nous faut prédire la variable aléatoire jointe « $A \rightleftharpoons C, B \rightleftharpoons C$ », étant donné la connaissance de la classification jointe des cases A et B.

Nous illustrons la définition de l'entropie implicative jointe par un exemple concret de prédiction au sein des données du tableau 3.9. Le tableau 3.11 présente les variables aléatoires nécessaires à la prédiction du participe passé à partir de la connaissance jointe de l'infinitif et du présent 3PL, c'est à dire :

- Variables dont la valeur est prédite :
  - INF $\rightleftharpoons$ PTCP, qui classe l'alternance entre infinitif et participe passé
  - 3PL $\rightleftharpoons$ PTCP, qui classe l'alternance entre présent 3PL et participe passé
- Variables dont la valeur est connue :
  - $\text{INF}_{\text{INF} \rightleftharpoons \text{PTCP}}$, qui classe les formes d'infinitif
  - $\text{3PL}_{\text{3PL} \rightleftharpoons \text{PTCP}}$, qui classe les formes de présent 3PL
  - INF $\rightleftharpoons$ 3PL, qui classe l'alternance entre infinitif et présent 3PL

Il est apparent dans le tableau 3.11 qu'à chaque combinaison des valeurs de variables connues correspond une unique combinaison de valeurs de variables à prédire. En conséquence, la distribution jointe des formes d'infinitif et de présent 3PL détermine entièrement la forme de participe passé. L'entropie est de 0 [12].

12. Dans cet exemple spécifique, la variable aléatoire INF $\rightleftharpoons$ 3PL ne semble pas nécessaire, car on a :

$$H(A \rightleftharpoons C, B \rightleftharpoons C \mid A_{A \rightleftharpoons C}, B_{B \rightleftharpoons C}, A \rightleftharpoons B) = H(A \rightleftharpoons C, B \rightleftharpoons C \mid A_{A \rightleftharpoons C}, B_{B \rightleftharpoons C})$$

Cela n'est pas le cas en général. Dans une situation où A est un prédicteur parfait de C, mais non B, on a : $H(A \rightleftharpoons C, B \rightleftharpoons C \mid A_{A \rightleftharpoons C}, B_{B \rightleftharpoons C}, A \rightleftharpoons B) = 0$

Classe	Taille	Lexème exemplaire	INF ⇌ PTCP	3PL ⇌ PTCP	$\text{INF}_{\text{INF}\rightleftharpoons\text{PTCP}}$	$\text{3PL}_{\text{3PL}\rightleftharpoons\text{PTCP}}$	INF ⇌ 3PL
(i)	4108	LIVRER (/livʁe/, /livʁ/, /livʁe/)	$p_1 =$ $\_ \rightleftharpoons \_/X\text{e}$	$p_6 =$ $\_ \rightleftharpoons \_\text{e}/X\_$	$\{p_1\}$	$\{p_6\}$	$p_{12} =$ $\_\text{e} \rightleftharpoons \_/X\_$
(ii)	210	RELIER (/ʁəlje/, /ʁəli/, /ʁəlje/)	p_1	$p_7 =$ $\_\text{i} \rightleftharpoons \_\text{je}/X\_$	$\{p_1\}$	$\{p_6, p_7\}$	$p_{13} =$ $\_\text{je} \rightleftharpoons \_\text{i}/X\_$
(iii)	22	RATISSER (/ʁatise/, /ʁatis/, /ʁatise/)	p_1	p_6	$\{p_1\}$	$\{p_6, p_8\}$	p_{12}
(iv)	327	BÂTIR (/batiʁ/, /batis/, /bati/)	$p_2 =$ $\_\text{ʁ} \rightleftharpoons \_/X\text{i}\_$	$p_8 =$ $\_\text{s} \rightleftharpoons \_/X\text{i}\_$	$\{p_2, p_3\}$	$\{p_6, p_8\}$	$p_{14} =$ $\_\text{ʁ} \rightleftharpoons \_\text{s}/X\text{i}\_$
(v)	37	TENIR (/təniʁ/, /tjɛn/, /təny/)	$p_3 =$ $\_\text{iʁ} \rightleftharpoons \_\text{y}/X\_$	$p_9 =$ $\_\text{jɛn} \rightleftharpoons \_\text{əny}/X\text{n}\_$	$\{p_2, p_3\}$	$\{p_6, p_9\}$	$p_{15} =$ $\_\text{əniʁ} \rightleftharpoons \_\text{jɛn}/X\_$
(vi)	8	OUVRIR (/uvʁiʁ/,/uvʁ/,/uvɛʁ/)	$p_4 =$ $\_\text{ʁiʁ} \rightleftharpoons \_\text{ɛʁ}/X\_$	$p_{10} =$ $\_\text{ʁ} \rightleftharpoons \_\text{ɛʁ}/X\_$	$\{p_2, p_3, p_4\}$	$\{p_6, p_{10}\}$	$p_{16} =$ $\_\text{iʁ} \rightleftharpoons \_/X\_$
(vii)	1	MOURIR (/muʁiʁ/,/mœʁ/,/mɔʁ/)	$p_5 =$ $\_\text{u}\_\text{iʁ} \rightleftharpoons \_\text{ɔ}\_/\text{m}\_\text{ʁ}$	$p_{11} =$ $\_\text{œ}\_ \rightleftharpoons \_\text{ɔ}\_/X\text{m}\_\text{ʁ}$	$\{p_2, p_3, p_4, p_5\}$	$\{p_6, p_{11}\}$	$p_{17} =$ $\_\text{u}\_\text{iʁ} \rightleftharpoons \_\text{œ}\_/\text{m}\_\text{ʁ}$

TABLEAU 3.11 – Variables aléatoires pour prédire le PST.PTCP depuis INF et PRS.3PL dans le tableau 3.9.

Cette situation contraste avec la prédiction depuis une case unique : si l'on prédit depuis l'infinitif seul, la variable aléatoire qui classe les formes d'infinitif regroupe ensemble les classes (iv) et (v) qui instancient pourtant des patrons différents, menant à une entropie implicative unaire non nulle, $H(\textsc{inf} \Rightarrow \textsc{ptcp}) \approx 0.0064$. De même, lorsque l'on prédit depuis le 3PL, la variable aléatoire classant les formes de 3PL groupe ensemble les classes (iii) et (iv) qui instancient des patrons distincts, menant à une entropie implicative unaire non nulle $H(\textsc{3pl} \Rightarrow \textsc{ptcp}) \approx 0.0251$.

Cette définition de l'entropie implicative n-aire permet donc bien de rendre compte de la situation que nous observions sur les données du français. Ici, l'entropie implicative binaire tombe à 0 tandis que chaque entropie unaire est strictement positive. En général, la propriété suivante est une conséquence triviale de la définition de Bonami et Beniamine (2016) :

$$H(A, B \Rightarrow C) \leq \min\{H(A \Rightarrow C), H(B \Rightarrow C)\}$$

En effet, la connaissance de deux formes d'un lexème ne peut en aucun cas être moins prédictive que la connaissance de l'une seule des formes. Il n'est pas cependant nécessaire que la connaissance d'une case supplémentaire soit plus prédictive (il se peut que la prédictibilité soit égale). La comparaison de l'entropie implicative binaire aux entropies implicatives unaires permet donc d'évaluer si, empiriquement, la connaissance jointe est utile dans la prédiction dans les paradigmes flexionnels.

La définition de l'entropie implicative jointe s'étend aisément au cas général de la prédiction depuis n cases. Soient n cases de paradigme, $A^1, \ldots, A^n$, nous notons $[A^1 \rightleftharpoons \cdots \rightleftharpoons A^n]$ la variable aléatoire jointe de tous les $A^i \rightleftharpoons A^j$ pour $i, j \in \{1, ..., n\}$. Nous définissons alors l'entropie implicative n-aire comme suit [13]. :

---

mais :

$H(A \rightleftharpoons C, B \rightleftharpoons C \mid A_{A \rightleftharpoons C}, B_{B \rightleftharpoons C}) = H(B \rightleftharpoons C \mid B_{B \rightleftharpoons C}) > 0$.

13. Les définitions proposées dans cette section se généralisent également au cas des prédictions dans les systèmes bipartites en substituant à toutes les prédictions $A \Rightarrow B$ leur décomposition en variables aléatoires jointes $A_1 \Rightarrow B_1, A_2 \Rightarrow B_2$ exposées précédemment.

$$H(A^1, \ldots, A^n \Rightarrow B) =$$
$$H\left(A^1 \rightleftharpoons B, \ldots, A^n \rightleftharpoons B \mid A^1_{A^1 \rightleftharpoons B}, \ldots, A^n_{A^n \rightleftharpoons B}, [A^1 \rightleftharpoons \cdots \rightleftharpoons A^n]\right) \quad (3.1)$$

3.5 Implications *n*-aires : résultats empiriques

Nous appliquons cette méthodologie à l'ensemble des systèmes étudiés. En principe, nous voudrions calculer l'entropie implicative moyenne n-aire pour chaque système flexionnel et toutes les valeurs possibles de n. En pratique, le nombre de combinaisons de n prédicteurs pour prédire 1 case de paradigme croît extrêmement vite [14], ce qui rend le calcul trop complexe pour des valeurs élevées de n. Nous nous arrêtons à $n = 2$ pour l'arabe et le navajo, $n = 3$ pour le portugais, français et chatino de Zenzontepec (qui n'a que 4 cases) et $n = 4$ pour les autres systèmes. Le tableau 3.12 indique les entropies implicatives brutes en fonction du nombre de prédicteurs [15], et la figure 3.17 en propose une comparaison visuelle.

Il apparaît clairement qu'en moyenne, dans chaque langue, la prédiction à partir de deux formes est considérablement plus facile qu'à partir d'une forme, et l'ajout de cases de paradigme supplémentaires continue à réduire drastiquement la difficulté du PCFP. Ces résultats établissent avec certitude l'utilité de la prédiction jointe dans le cadre du PCFP. Ce résultat observé sur les moyennes se reflète nettement dans les distributions entières : les figures 3.2, 3.18, 3.19 et 3.20 montrent un écrasement progressif des distributions vers les entropies nulles.

Ces résultats confirment ceux de Bonami et Beniamine (2016), et fournissent un argument supplémentaire en faveur de la conjecture d'entropie basse (Ackerman, Blevins et Malouf 2009) : non seulement les entropies implicatives moyennes à partir d'un prédicteur unique sont gé-

14. En général, pour n cases de paradigme et k prédicteurs, il y a $\frac{n!}{k!(n-k-1)!}$ combinaisons. Avec trois prédicteurs, il y a $999,600$ entropies à considérer pour les 51 cases du français, $3,458,004$ pour les 69 cases du portugais, $3,667,580$ pour les 70 du navajo, et $22,253,004$ pour les 109 cases de paradigme des verbes arabe. Pour chacun de ces calculs, il faut itérer sur l'ensemble du lexique, qui est pour ces langues de l'ordre de quelques milliers de lexèmes.

15. Puisque le chatino de Zenzontepec ne comporte que quatre cases de paradigmes, il est impossible de calculer une prédiction à quatre prédicteurs. La case correspondante est grisée dans les tableaux 3.12, 3.13 et 3.14

	$n=1$	$n=2$	$n=3$	$n=4$
Anglais	0.1786	0.0999	0.0544	0.0314
Arabe	0.3165	0.1097	—	—
Chatino Y.	0.6848	0.1631	0.0546	0.0253
Chatino Z.	0.7030	0.2574	0.1105	
Français	0.1844	0.0582	0.0224	0.0095
Navajo	0.3862	0.0999	0.0423	—
Portugais	0.1671	0.0606	0.0375	—
Russe	0.6367	0.2630	0.1276	0.0688

TABLEAU 3.12 – Entropie implicative n-aire moyenne pour diverses valeurs de n.

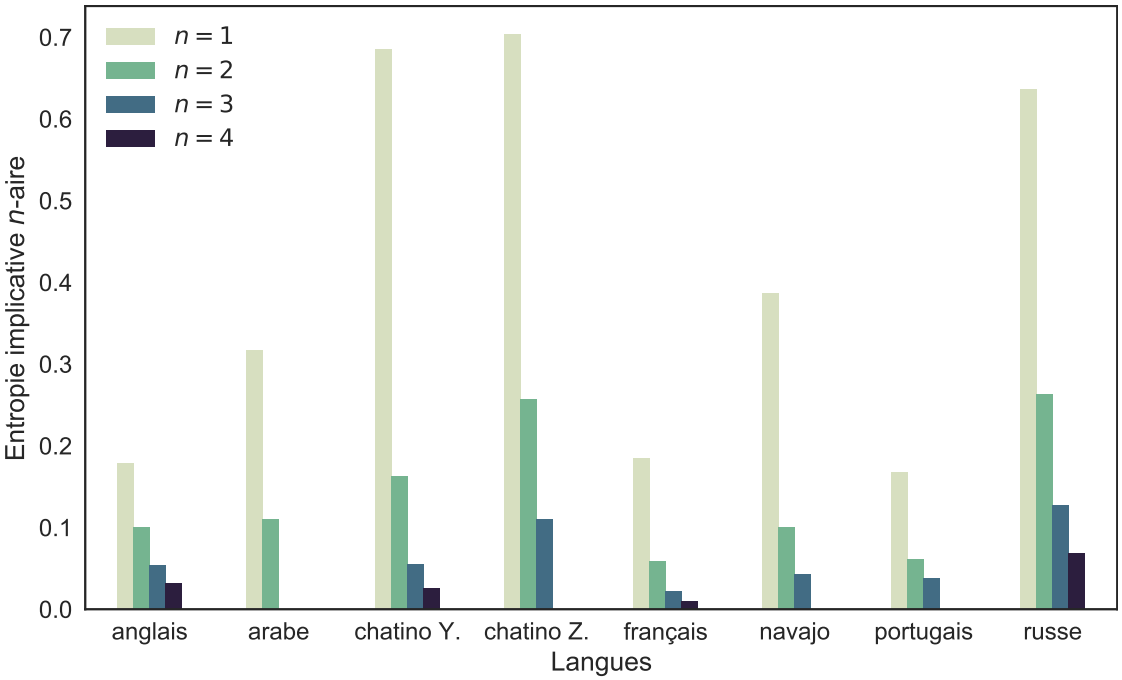


FIGURE 3.17 – Évolution des entropie implicatire n-aire en fonction de n.

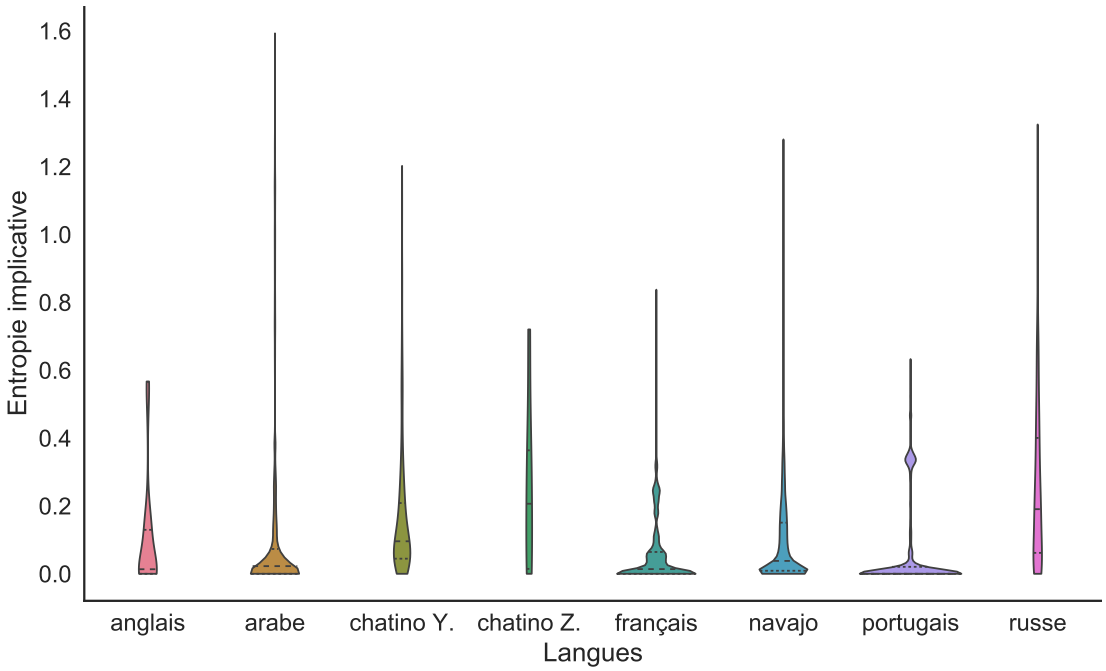


FIGURE 3.18 – Distributions des entropies implicatives binaires pour chaque système flexionnel étudié.

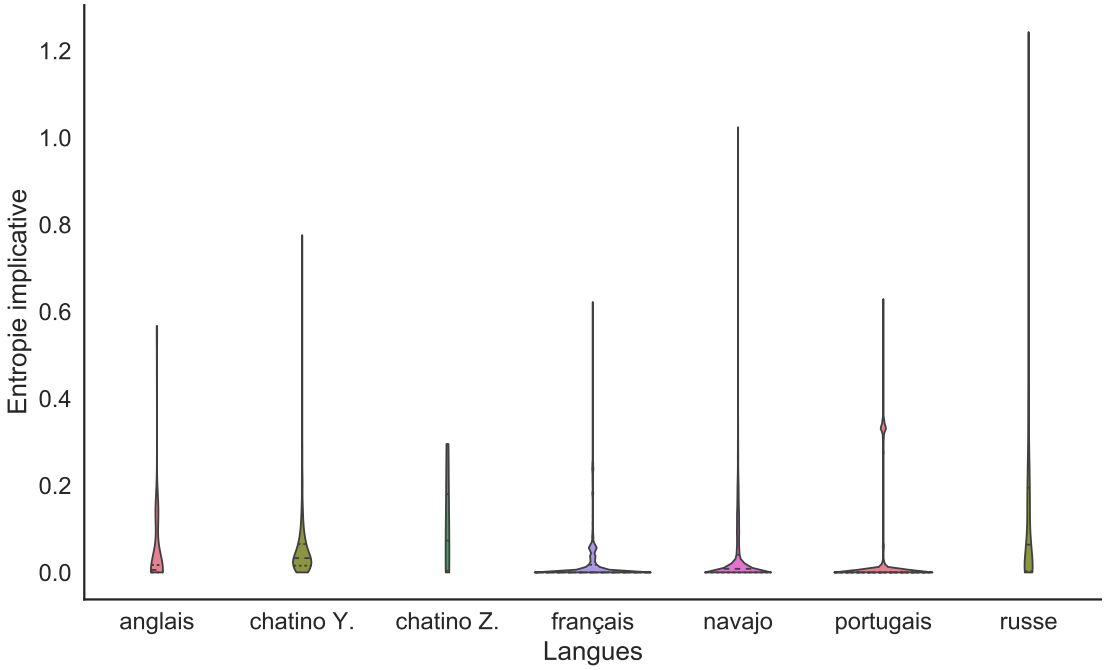


FIGURE 3.19 – Distributions des entropies implicatives ternaires calculées.

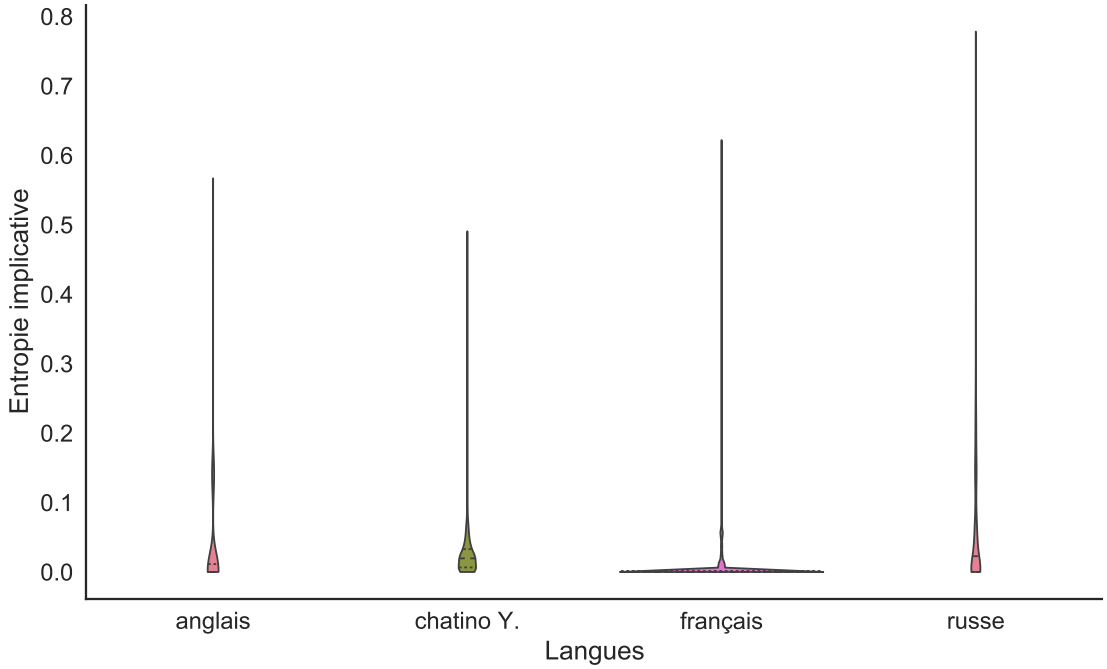


FIGURE 3.20 – Distributions des entropies implicatives à quatre prédicteurs calculées.

néralement basses à travers des systèmes variés, mais elles décroissent très fortement avec l'augmentation du nombre de prédicteurs. La prédiction jointe est donc généralement utile aux locuteurs pour résoudre le PCFP.

3.6 Des implications *n*-aires aux parties principales

Dans la perspective constructive au sens de Blevins (2006), le PCFP soulève la question de la quantité d'information minimale à partir de laquelle il est possible de générer l'ensemble des formes d'un paradigme. Dans cet esprit, Raphael Finkel et Greg Stump ont formalisé et déployé la notion de SYSTÈME DE PARTIES PRINCIPALES (Finkel et Stump 2007, 2009 ; Stump et Finkel 2013). Les systèmes de parties principales sont une notion héritée des grammaires pédagogiques. En termes classiques, il s'agit d'un ensemble de cases de paradigme suffisant à prédire toutes les autres cases de paradigme pour tous les lexèmes. Opérer cette prédiction revient alors à sélectionner la classe flexionnelle appropriée, puis à en appliquer les règles de construction pour générer toutes les autres formes. Les parties principales sont donc des ensembles de cases de

paradigme qui permettent d'identifier avec certitude la classe flexionnelle de n'importe quel lexème. Il existe une relation entre parties principales et implications paradigmatiques : pour tout système flexionnel ayant les cases $S = \{c_1, \dots c_n\}$, il existe un système de parties principales $C \subset S$ si et seulement si il existe une collection d'implications catégoriques dont les antécédents portent sur toutes les cases de C et aucune autre, et dont les conséquents déterminent l'ensemble des cases de $S \setminus C$. D'après Stump et Finkel, dans une visée pédagogique, les systèmes de parties principales doivent être de cardinalité minimale et fournir une prédiction catégorique. Ainsi elles peuvent fournir à l'apprenant le minimum d'information à mémoriser afin de fléchir avec certitude toutes les formes des lexèmes.

Les entropies implicatives n-aires permettent d'étudier ces systèmes de parties principales (Hockett 1967 ; Matthews 1972 ; Finkel et Stump 2007 ; Stump et Finkel 2013). En effet, tout ensemble de cases de paradigme fournissant une prédiction catégorique sur le reste du paradigme constitue un ensemble de parties principales.

Bonami et Beniamine (2016) nomment ENTROPIE RÉSIDUELLE la moyenne des entropies pour un ensemble de prédicteurs donné :

> Étant donné un système flexionnel ayant un ensemble de cases de paradigme $\mathcal{C}$, nous définissons l'ENTROPIE RÉSIDUELLE d'un ensemble $\mathcal{A} = \{A^1, \dots, A^n\} \subset \mathcal{C}$ comme la moyenne de tous les $H(A^1, \dots, A^n \Rightarrow B)$ pour $B \in \mathcal{C} \setminus \mathcal{A}$ [16].

Les ensembles de parties principales peuvent donc être déduits de l'entropie implicative n-aire : tous les ensembles de prédicteurs ayant une entropie résiduelle nulle forment des systèmes de parties principales. Le tableau 3.13 indique le nombre de systèmes de parties principales de cardinalité n trouvés dans chaque langue.

Globalement, on trouve assez peu de systèmes de parties principales à ces cardinalités [17]. Ces ensembles de parties principales permettent une prédiction des autres formes avec une certitude absolue. Cependant, les locuteurs ne se préoccupent vraisemblablement pas de prédiction

16. [En anglais dans le texte] « *Given an inflectional system with paradigm cells $\mathcal{C}$, we define the* RESIDUAL UNCERTAINTY *of any subset $\mathcal{A} = \{A^1, \dots, A^n\}$ of $\mathcal{C}$ as the average of all $H(A^1, \dots, A^n \Rightarrow B)$ for $B \in \mathcal{C} \setminus \mathcal{A}$* ».

17. Dans le cas du portugais, le problème relevé plus haut concernant les patrons identité de petite couverture mène à sous-estimer les parties principales catégoriques. Là où nous trouvons 25 ensembles de parties principales, Bonami et Beniamine (2016) en indiquent 184 en raison d'une erreur de frappe. Ils en trouvent en fait 177.

	$n=1$	$n=2$	$n=3$	$n=4$
Anglais	0	0	0	6
Arabe	0	0	—	—
Chatino Y.	0	0	0	0
Chatino Z.	0	0	1	
Français	0	0	0	12 000
Navajo	0	0	7	—
Portugais	0	25	1 148	—
Russe	0	0	0	0

TABLEAU 3.13 – Nombre de parties principales catégoriques de cardinalité n.

catégorique. D'un point de vue psycholinguistique, il est clair que les locuteurs ne manifestent pas une connaissance parfaite des systèmes, puisque même les locuteurs natifs compétents font occasionnellement des erreurs de flexion, et que celles-ci sont du même type que les erreurs des apprenants L1 et L2. Cela suggère qu'ils s'appuient sur des implications fiables mais non nécessairement catégoriques pour inférer les formes inconnues. En conséquence, y compris dans une perspective pédagogique, il n'est pas utile de se restreindre aux parties principales fournissant des prédictions catégoriques. Nous proposons d'examiner les parties principales permettant de produire des inférences très fiables, mais non nécessairement parfaites.

Le tableau 3.14 indique les systèmes de parties principales menant à une entropie résiduelle (moyenne par prédicteur) inférieure ou égale à 0.005 [18].

Dans tous les systèmes présentant un grand nombre de cases de paradigme, on trouve de nombreux systèmes de *presque* parties principales binaires. En anglais et en chatino de Zenzontepec, il n'y en a qu'à partir de trois prédicteurs. Dans tous les cas, quatre prédicteurs sont suf-

18. Une entropie de 0.005 correspond à la certitude d'un événement de probabilité 0.99 dans une situation de prédiction binaire. Suite à une erreur de calcul, Bonami et Beniamine (2016) indiquent les ensembles de cases qui prédisent chaque autre cases de paradigme avec au moins une entropie de 0.005, plutôt que tous les ensembles de cases qui prédisent en moyenne avec une entropie maximale de 0.005. Cette erreur explique la différence entre nos résultats.

	$n=1$	$n=2$	$n=3$	$n=4$
Anglais	0	0	11	30
Arabe	0	57	—	—
Chatino Y.	0	0	0	36
Chatino Z.	0	0	2	
Français	0	221	10 411	181 372
Navajo	0	16	10 520	—
Portugais	0	552	21 528	—
Russe	0	0	0	72

TABLEAU 3.14 – Nombre de *presque* parties principales de cardinalité n.

fisants pour tirer des inférences très fiables. Ces résultats vont dans le sens de notre hypothèse selon laquelle les systèmes présentant plus de cases de paradigme présentent généralement des structures implicatives plus fortement liées.

3.7 Conclusion

Nous avons employé les patrons d'alternances inférés automatiquement sur huit systèmes flexionnels afin d'y calculer la difficulté moyenne à résoudre le PCFP. Nous utilisons la formulation de l'entropie implicative selon Bonami et Boyé (2014) et son extension à la prédiction au sein des systèmes bipartites (Bonami et Beniamine 2016). Nous avons montré que la conjecture de basse entropie formulée par Ackerman, Blevins et Malouf (2009) se confirme sur ces systèmes, y compris ceux que nous analysons comme bipartites. Dans ces systèmes, nous avons montré que les deux sous-systèmes ne sont jamais indépendants, et que la prédiction du PCFP doit se faire conjointement sur chaque sous-système. Les résultats de la méthodologie de Bonami et Beniamine (2016) appliquée à l'ensemble de nos données confirment que le leur que dans le cadre du PCFP, la prédiction jointe est extrêmement utile au PCFP.

Nous avons remarqué une tendance générale selon laquelle les systèmes flexionnels présentant plus de cases de paradigme tendent à présenter une prédictibilité implicative plus forte.

Ces systèmes ont généralement des entropies plus basses, et présentent de grands groupes de prédicteurs d'entropie basse. On y trouve systématiquement des *presque* parties principales de cardinalité 2, ce qui n'est pas le cas dans les systèmes présentant moins de cases de paradigme. La figure 3.21 présente la relation linéaire observée entre la taille des paradigmes et les entropies implicatives dans chaque langue, de 1 à 3 prédicteurs. Chaque point y représente une mesure d'entropie implicative moyenne. Avec un prédicteur, on observe clairement une corrélation négative entre taille du paradigme et entropie implicative moyenne. La tendance observée s'affaiblit avec le nombre de prédicteurs, car les entropies s'homogénéisent alors vers le bas.

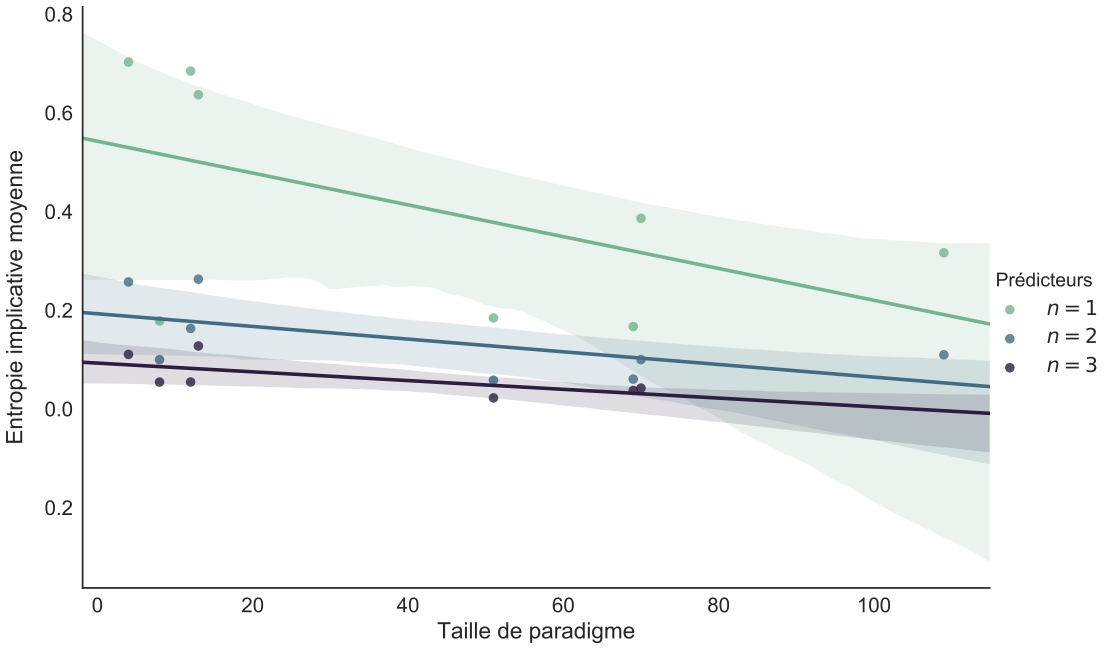


FIGURE 3.21 – Relation entre taille du paradigme et entropie implicative moyenne.

D'où provient cette relation entre taille de paradigme et entropie implicative ? Les tailles de paradigmes dépendent de choix de modélisation : en général, les morphologues choisissent de distinguer deux cases lorsqu'il existe deux contextes syntaxiques pour lesquels au moins certains lexèmes présentent un contraste formel. Ainsi, en anglais, nous distinguons l'infinitif de la première personne du présent en raison du verbe BE, qui présente un contraste entre I am et to be. Que l'on choisisse ou non de distinguer des cases de paradigme, il existe en anglais des

contextes d'emploi portant une valeur d'infinitif et des contextes d'emploi portant une valeur de présent. Les paradigmes ne forment donc qu'une partition des contextes d'emploi des lexèmes. Le PCFP, en raisonnant sur des cases de paradigme et non sur des contextes syntaxiques, ne capte donc pas exactement la tâche de prédiction que doit résoudre le locuteur. Le problème d'un locuteur du français ayant entendu la phrase (47) et voulant produire la phrase (48) est le même que celui d'un locuteur de l'anglais ayant entendu la phrase (49) et voulant produire la phrase (50).

(47) Elles mangent des falafels.

(48) Nous mangeons des falafels.

(49) They eat falafels.

(50) We eat falafels.

Lorsque nous évaluons le PCFP, nous quantifions bien la difficulté de ce problème en français, mais en anglais, nous ne prenons pas en compte la difficulté (nulle) à produire (50) à partir de (50). Ce problème, comme nous l'avons noté précédemment, pourrait être pallié en pondérant la moyenne des entropies par les fréquences de type des cases prédictrices et prédites. L'absence de données fiables ne nous a pas permis de le faire dans le cadre de cette thèse. Cependant, l'observation de la relation entre taille de paradigme et entropie implicative nous mène à formuler la CONJECTURE DE STABILITÉ DE L'ENTROPIE IMPLICATIVE : si nous considérions non pas les cases de paradigmes mais les contextes d'emplois syntaxiques, l'entropie implicative moyenne pondérée à travers les langues devrait être relativement stable. Cette conjecture expliquerait l'observation ci-dessus en suggérant que dans les petits systèmes, l'entropie haute n'est que le reflet de l'omission des prédictions syncrétiques. Pour confirmer cette conjecture, il nous faudrait d'une part un échantillon de langues beaucoup plus grand, permettant de confirmer la relation esquissée en figure 3.21, et d'autre part des données concernant les fréquences de cases de paradigmes dans chaque langue.

Deuxième partie

La structure de similarité des systèmes flexionnels

Chapitre 4

Classification des lexèmes en microclasses

Ce chapitre propose une classification flexionnelle en microclasses au sein desquelles tous les paradigmes sont caractérisés par des patrons identiques. Les microclasses peuvent se déduire à partir des patrons d'alternance. Elles sont fondées sur l'identité du comportement flexionnel : chaque microclasse est caractérisée par un vecteur de patrons, où chaque coordonnée du vecteur indique le patron instancié pour cette microclasse pour une paire de cases de paradigme. La classification des paradigmes en microclasses constitue le point de départ des entreprises de classification plus ambitieuses présentées dans les chapitres 5 et 6. Les systèmes de microclasses méritent cependant d'être étudiés en tant que tels, ne serait-ce que parce que les critères permettant de les inférer sont relativement simples et consensuels.

Nous décrivons tout d'abord dans la section 4.1 les distributions des fréquences de types de microclasse. Dans la section 4.2, nous décrivons les réseaux de similarités entre microclasses en suivant la méthodologie de Sims et Parker (2016). Enfin la section 4.3 explore l'utilité de l'inférence de structure hiérarchiques arborescentes pour observer les relations de distances et de similarité entre microclasses.

4.1 Distributions des microclasses

Comme nous l'avons discuté au chapitre 1, section 1.3.3, les microclasses forment une classification très fine, distinguant la moindre variation dans le comportement flexionnel. Pour cette raison, le nombre de microclasses trouvées est généralement considérablement plus haut que le nombre de macroclasses présentées dans les grammaires descriptives. Le tableau 4.1

indique le nombre de microclasses obtenu, pour chaque système flexionnel étudié dans cette thèse, à l'issue du chapitre 2. Nous comparons le nombre de microclasses au nombre de paradigmes de surface pour chaque lexique. Dans le cas du français, de l'anglais, du portugais européen et de l'arabe, le nombre de paradigmes coïncide avec le nombre de lexème. Dans les systèmes bipartites, c'est à dire en russe, en chatino et en navajo, nous comptons le nombre de paradigmes distincts pour chaque dimension. Par exemple, en navajo, notre lexique comporte 2157 lexèmes, qui s'analysent en 805 paradigmes de base_1 et 1497 paradigmes de base_2.

On distingue trois situations concernant la proportion de paradigmes par microclasses : en anglais, français, portugais les classes sont peu nombreuses comparé au nombre de paradigmes, résultant en des microclasses comportant en moyenne quelques dizaines de formes. Cette situation correspond à l'intuition selon laquelle les microclasses constituent des sous-généralisations plus précises que des macroclasses. Le système segmental du russe, le système tonal du chatino de Zenzontepec et le système verbal de l'arabe présentent des ratios intermédiaires, manifestant une généralisation moins forte des paradigmes aux microclasses.

	Paradigmes	Microclasses	Proportion
Français	5249	97	54.11
Anglais	6064	118	51.39
Portugais	1996	60	33.27
Arabe	1018	367	2.77
Russe (segments)	1530	159	9.62
Russe (accents)	159	87	1.83
Chatino de Zenzontepec (segments)	370	55	6.73
Chatino de Zenzontepec (tons)	75	52	1.44
Chatino de Yaitepec (segments)	284	203	1.40
Chatino de Yaitepec (tons)	130	129	1.01
Navajo (base_1)	805	805	1.00
Navajo (base_2)	1497	1268	1.18

Tableau 4.1 – Taille des microclasses.

En navajo, nous trouvons autant de microclasses de bases$_1$ que de bases$_1$ distinctes en surface, et presque autant de microclasses de base$_2$ que de base$_2$ distinctes en surface. On trouve un ratio autour de 1 également pour le système accentuel du russe, le chatino de Yaitepec, et le système segmental du chatino de Zenzontepec. Ces ratios sont très bas, soulignant le fait qu'une partition de microclasses constitue un mauvais modèle de la structure de ces systèmes flexionnels. S'il existe dans ces systèmes des similarités et des relations implicatives entre formes, comme nous l'avons vu dans le chapitre 3, elles ne s'organisent pas en microclasses. Autrement dit, les relations d'interprédictibilité locales à certaines paires de formes ne se combinent pas en faisceaux de relations parallèles. Notons que les systèmes du chatino et le système accentuel du russe présentent seulement un petit nombre de paradigmes de surface uniques. Pour le chatino, il se peut que nos données présentent principalement des paradigmes exemplaires de microclasses, et qu'en conséquence, les fréquences de type des microclasses ne constituent pas un échantillon réaliste. En chatino comme pour le système accentuel du russe, il y a suffisamment peu de formes distinctes pour qu'il soit imaginable qu'un locuteur puisse les mémoriser.

La proportion de paradigmes par microclasse nous renseigne cependant très peu sur les distributions de leur taille. Les deux figures 4.1 présentent les distributions de fréquence de type de microclasses par rang pour chaque système. Nous pourrions nous attendre à ce que les systèmes présentant un très faible nombre de paradigmes par microclasse en moyenne suivent une distribution distincte. Pour cette raison, nous produisons deux figures distinctes selon que le ratio est supérieur ou inférieur à 5 paradigmes par microclasses. Dans les deux cas, les distributions des fréquences de type des microclasses suivent une loi de puissance (les deux axes des figures sont présentés à l'échelle logarithmique). Il existe toujours peu de microclasses très fréquentes, et beaucoup plus de microclasses très rares (1 lexème). Cette observation rejoint celles de Blevins, Milin et Ramscar (2017) discutées au chapitre 3 concernant l'organisation zipfienne du lexique.

Comme nous l'avons dit, les microclasses constituent une classification très fine, où la moindre propriété non partagée donne lieu à une classe distincte. Dès lors, nous pouvons nous demander quelles relations entretiennent les microclasses au sein de chaque système. Nous proposons d'explorer la structure de similarité des systèmes de microclasses en fonction des

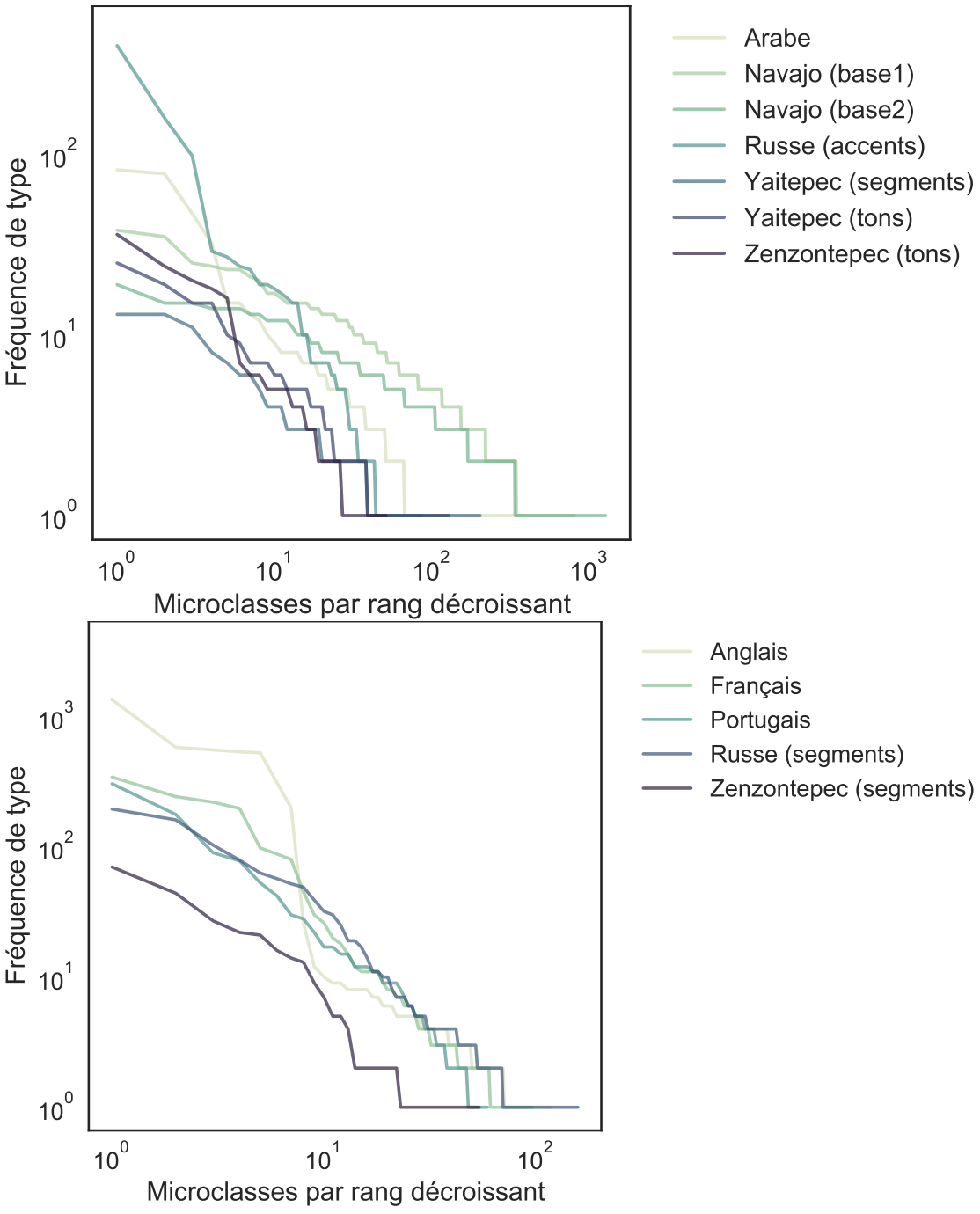


FIGURE 4.1 – Fréquence de type des microclasses.

patrons d'alternance qu'elles ont en commun.

4.2 Structure des réseaux de microclasses

Sims et Parker (2016) proposent de représenter la similarité entre systèmes de classes flexionnelles comme des graphes dans lesquels « [l]es nœuds sont des classes flexionnelles, leur taille reflétant le logarithme de la fréquence de type de la classe. Un arc entre deux classes indique que les classes partagent des exposants. Les arcs noirs relient des classes qui partagent au moins la moitié de leurs exposants [...]. Les arcs gris relient les classes qui partagent un exposant de moins que la moitié de leurs exposants [...] [1] ». Les autres arcs ne sont pas représentés. Ils obtiennent ainsi des représentations du système du français, de l'islandais, du grec, du russe, du kadiweu, du chinantec, du seri, du voro et du nuer fondés sur le partage d'exposants. Ils distinguent trois types de graphes : ceux du français, de l'islandais et du grec apparaissent très peu connectés. Ceux du russe, du kadiweu et du chinantec ressemblent à des réseaux en petits mondes, avec des sous-réseaux très connectés et quelques liens entre ceux-ci. Enfin, les trois systèmes du seri, du voro et du nuer présentent énormément d'arcs et sont proches de graphes aléatoires.

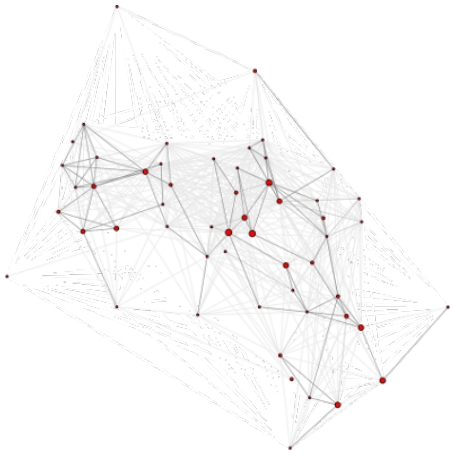
Nous pouvons caractériser ces trois types de graphes comme trois niveaux de distinctivité entre les classes au sens du principe 1 de Corbett (2009) déjà discuté au chapitre 1, et que nous rappelons ci-dessous :

> (principe I) Les classes sont **distinctives**
>
> (critère 1) à travers les classes, elles présentent des paradigmes concrets distincts.

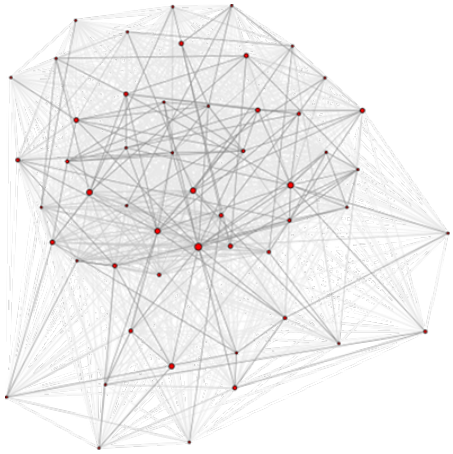
Nous proposons de mener la même analyse sur nos données en nous fondant non plus sur les exposants, mais sur les patrons. Il existe beaucoup plus de patrons d'alternance que de cases de paradigmes pour chaque système, si bien que les coupes à la moitié des traits et à la moitié moins un sont beaucoup moins utiles. Nous proposons de ne pas couper d'arcs, mais de les représenter sur une échelle de gris proportionnelle à la quantité de patrons partagés.

Ce faisant, nous retrouvons une classification assez similaire à celle de Sims et Parker (2016).

1. [En anglais dans le texte] « *The nodes are inflection classes, with the size of the node reflecting log type frequency of the class. An edge connecting two classes indicates that the classes share exponents. Black edges connect class nodes that share at least half of their exponents [...]. Gray edges are classes that share half-minus-one cells [...]* ».



Zenzontepec (segments)

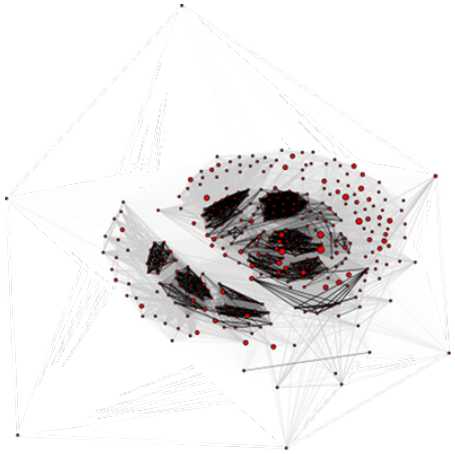


Zenzontepec (tons)

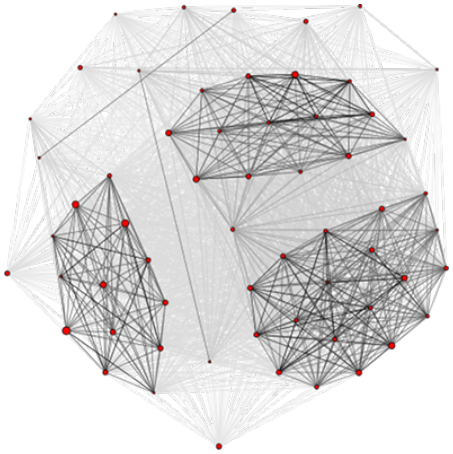
FIGURE 4.2 – Réseaux de microclasses de type peu connecté.

La catégorie des systèmes peu connectés est moins fournie, puisque les patrons captent plus de points de similarité que les exposants. On y trouve seulement les deux systèmes des verbes du chatino de Zenzontepec (figure 4.2).

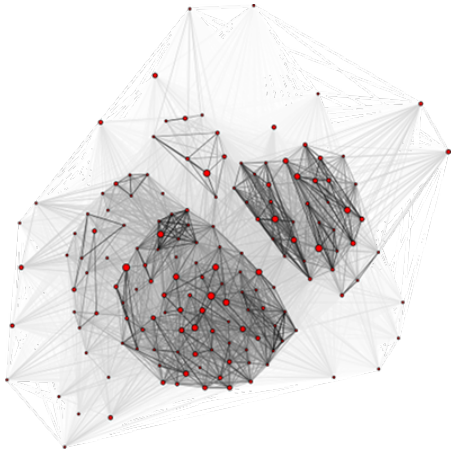
Plutôt que des systèmes ressemblant à de petits mondes, notre catégorie intermédiaire regroupe les systèmes dans lesquels la représentation en réseau laisse deviner des sous-réseaux bien connectés. Ces systèmes, présentés dans la figure 4.3, pourraient se prêter particulièrement bien à une description en termes de macroclasses. Le système flexionnel des verbes de l'arabe semble présenter une partition majeure en deux sous-réseaux, chacun peut-être à nouveau divisible en voisinages étroits. Par ailleurs, un grand nombre de classes à la périphérie ne sont pas connectées avec ces ensembles. Le système du portugais présente trois classes très nettes, de tailles équilibrées, ainsi également qu'un ensemble de classes moins connectées. Les segments du russe semblent s'organiser en macroclasses également, mais il est difficile de déterminer combien exactement, et la limite de leurs voisinages. Nous comptons également dans cette catégorie les accents du russe, qui se rapprochent également de la catégorie suivante. On peut cependant distinguer une séparation verticale au sein du réseau. Enfin, les tons du chatino de Yaitepec semblent également présenter des sous-réseaux plus denses, quoique ceux-ci soient moins faciles à discerner.



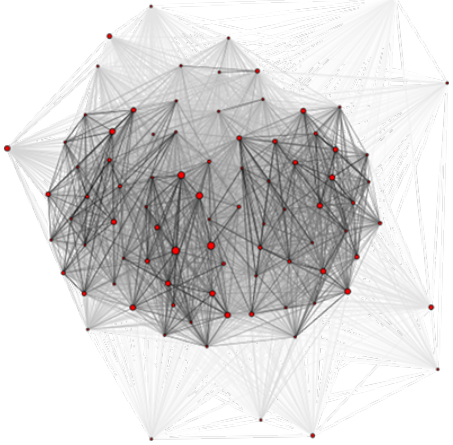
Arabe



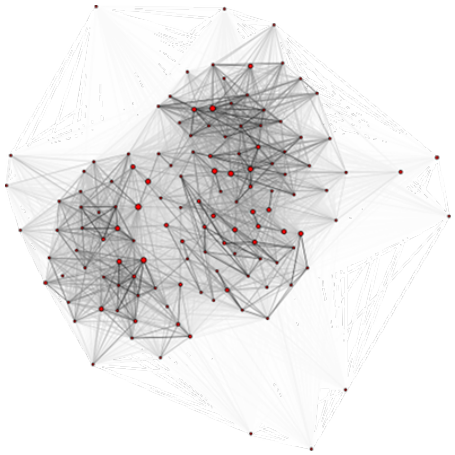
Portugais



Segments du russe



Accents du russe

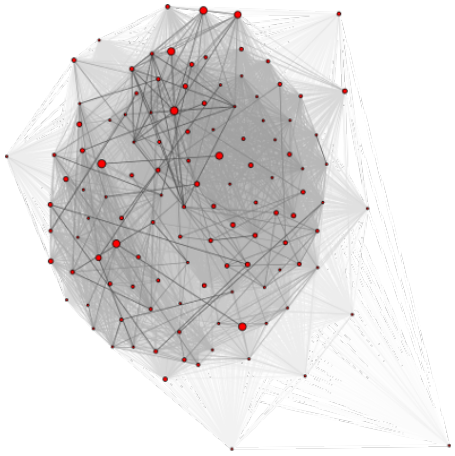


Tons du chatino de Yaitepec

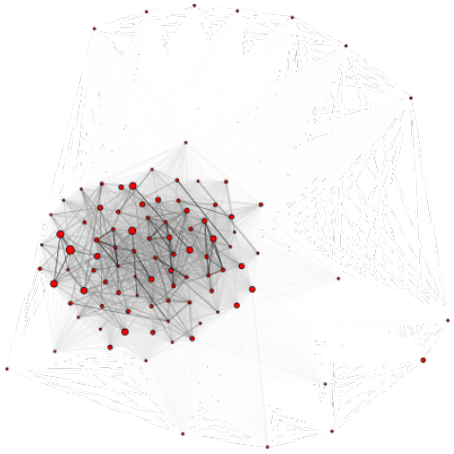
FIGURE 4.3 – Réseaux de microclasses de type « macroclasses ».

Enfin, les systèmes de l'anglais, du français, et du navajo (figure 4.4) présentent des connexions denses qui ne semblent pas, dans ces graphes, structurer l'espace en sous-classes.

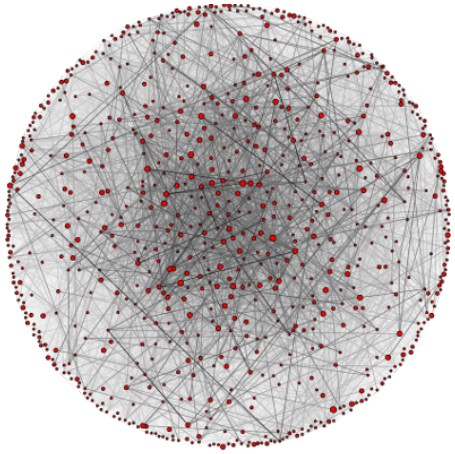
Ces représentations, quoique calculées sur des unités différentes, confirment les conclusions de Sims et Parker (2016) : les systèmes de classes flexionnelles varient non seulement par la connectivité de leur graphe (la quantité de partage entre microclasses), mais aussi par la structure même du réseau.



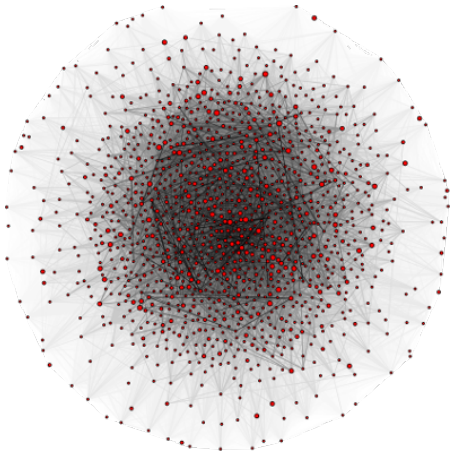
Anglais



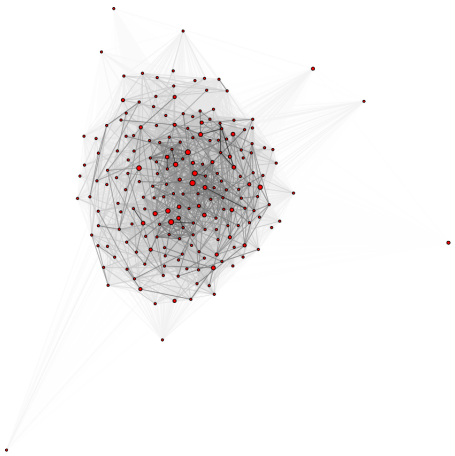
Français



Navajo base$_1$



Navajo base$_2$



Segments du chatino de Yaitepec

FIGURE 4.4 – Réseaux de microclasses de type dense.

4.3 Hiérarchiser les microclasses

De nombreux travaux, de la morphologie naturelle (Dressler, Mayerthaler et al. 1987 ; Dressler et Thornton 1996 ; Kilani-Schoch et Dressler 2005 ; Dressler, Kilani-Schoch et al. 2008 ; Dressler 2012) à la morphologie en réseau (Brown et Hippisley 2012), organisent les classes flexionnelles en structures hiérarchiques. Dans les deux cas, ces hiérarchies sont conçues selon un principe d'héritage par défaut. Kilani-Schoch et Dressler (2005) proposent par exemple que

> « les classes flexionnelles sont représentées sous la forme d'une structure hiérarchique en arbres et nœuds. Les nœuds les plus bas représentent les microclasses. La cime représente la macroclasse. Et ensuite, dans un ordre hiérarchique descendant, se succèdent les classes, les sous-classes, et si nécessaire les sous-sous-classes, etc., jusqu'aux microclasses. Dans cette hiérarchie arborescente, le principe de l'héritage du défaut (Corbett & Fraser 1993) […] signifie qu'un nœud peut être caractérisé par une propriété obligatoire ou par une propriété par défaut, et que cette propriété est héritée par les nœuds immédiatement dépendants de ce nœud supérieur. Les propriétés par défaut peuvent […] être annulées. »

Par ailleurs, les grammaires descriptives présentent souvent les classes flexionnelles sous la forme de classes et de sous-classes, employant ainsi implicitement une structure arborescente. C'est le cas de Campbell (2011) lorsqu'il décrit les classes flexionnelles du chatino de Zenzontepec. Il en va de même pour la grammaire Bescherelle des verbes français (Arrivé 2012), qui présente des classes, des sous-classes formant chacune des pages indépendantes, et en notes de bas de page des sous-sous classes. La figure 4.5 représente sous la forme d'un arbre la classification implicite du Bescherelle.

Cette section propose deux façons d'inférer des arbres flexionnels reflétant la distribution des patrons dans les microclasses. Nous présentons tout d'abord un algorithme simple permettant d'inférer un arbre par défaut. Nous proposons ensuite, suivant Bonami (2014), d'employer une méthode de *clustering* ascendant pour inférer des hiérarchies de similarité entre classes.

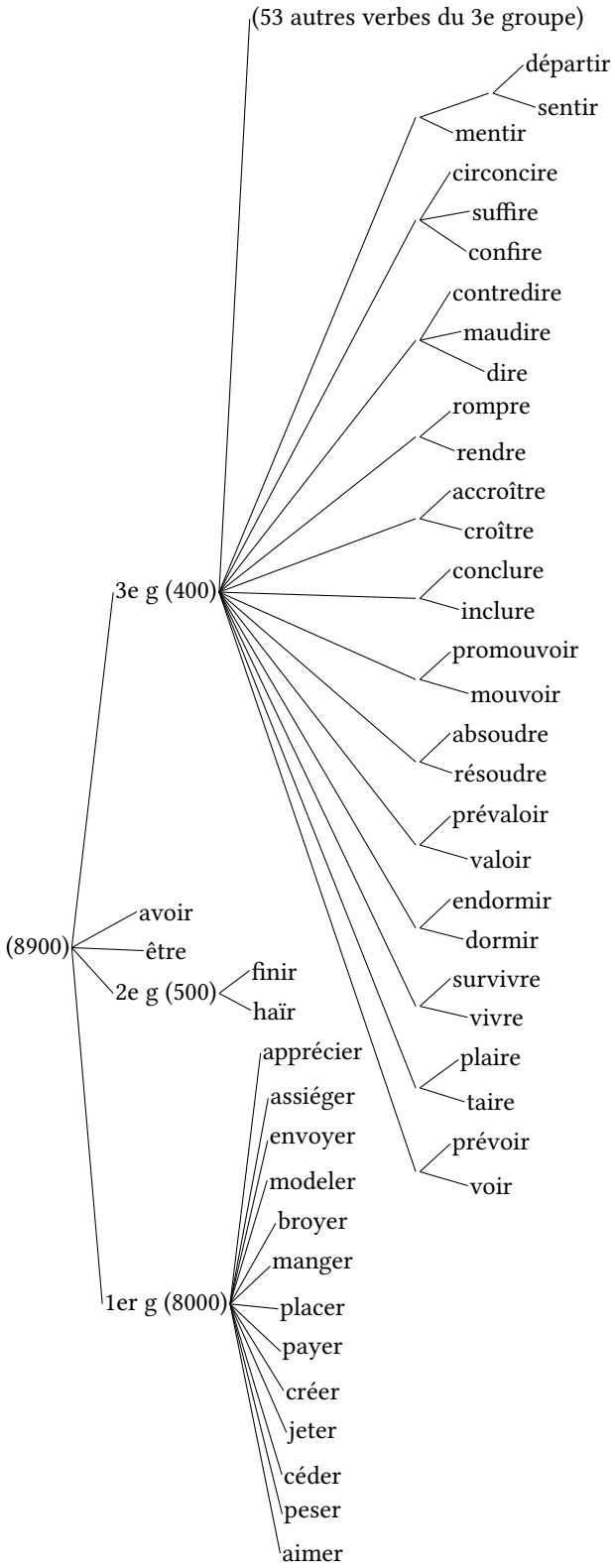


FIGURE 4.5 – Hiérarchie implicite de la grammaire Bescherelle (Arrivé 2012).

4.3.1 Structures hiérarchiques par défaut

Dans une grammaire constructive au sens de Blevins (2006), les défauts offrent une économie descriptive. Étant donnée une hiérarchie de classes flexionnelles dont les nœuds introduisent des règles ou des propriétés morphologiques, et dont les feuilles correspondent à des items lexicaux, Brown et Hippisley (2012) fournissent un critère afin de s'assurer qu'il n'existe pas de nœud inutile. Ils proposent que dans une telle hiérarchie, un nœud unaire (ayant un seul enfant) peut toujours être réduit en le fusionnant avec son enfant. En conséquence, un nœud n'est utile dans une hiérarchie de classes que s'il indique des valeurs partagées entre au moins deux sous-classes. Afin de déterminer les règles qui constituent des défauts, Brown et Hippisley (2012) proposent la règle de défaut de la majorité présentée en (51) :

(51) « Heuristique : Défaut de la majorité : La règle [...] qui est partagée par le plus de classes flexionnelles est traitée comme le défaut [2] ».

Brown et Hippisley (2012) identifient quatre principales classes flexionnelles nominales en russe, que nous présentons dans le tableau 4.2. Dans ce tableau, nous marquons en ligne par des cases grisées les partages entre classes que Brown et Hippisley (2012) identifient comme importants.

Ils proposent pour rendre compte de cette structure la description arborescente en (52). Dans celle-ci, les feuilles N_I à N_IV représentent chaque classe flexionnelle. Un premier nœud NOMINAL, qui recouvre les adjectifs (que nous ignorons ici) et les noms, définit les propriétés communes aux noms et adjectifs, ainsi que deux valeurs par défaut, une pour le nominatif singulier (exposant zéro) et une pour le nominatif pluriel en -i. Par ailleurs, le terme *evaluation* qui remplace parfois un exposant dénote une fonction de réalisation qui prend en entrée les traits morphologiques et peut attribuer des valeurs distinctes à des lexèmes d'une même classe. Le nœud NOM spécifie une voyelle thématique propre aux noms (pour trois cas du pluriel), un prépositionnel singulier par défaut en -e, et un syncrétisme par défaut entre le datif singulier et prépositionnel singulier. Il existe un nœud intermédiaire, N_O, qui manifeste les propriétés

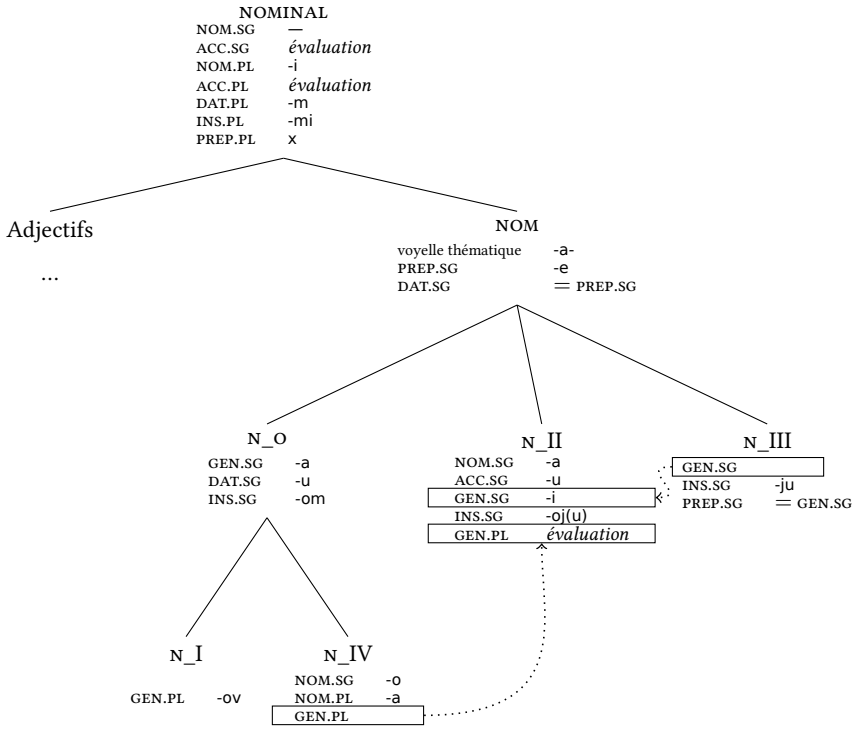
2. [En anglais dans le texte] « ***Heuristic : Majority default**: The rule [...] which is shared by most inflectional classes is treated as the default ».*

	I	II	III	IV
NOM.SG	zakon	kart-a	rukop'is'	bolot-o
ACC.SG	zakon	kart-u	rukop'is'	bolot-o
GEN.SG	zakon-a	kart-i	rukop'is'-i	bolot-a
DAT.SG	zakon-u	kart-e	rukop'is'-i	bolot-u
INS.SG	zakon-om	kart-oj	rukop'is'-ju	bolot-om
PREP.SG	zakon-e	kart-e	rukop'is'-i	bolot-e
NOM.PL	zakon-i	kart-i	rukop'is'-i	bolot-a
ACC.PL	zakon-i	kart-i	rukop'is'-i	bolot-a
GEN.PL	zakon-ov	kart	rukop'is'-ej	bolot
DAT.PL	zakon-am	kart-am	rukop'is'-am	bolot-am
INS.PL	zakon-ami	kart-ami	rukop'is'-ami	bolot-ami
PREP.PL	zakon-ax	kart-ax	rukop'is'-ax	bolot-ax

TABLEAU 4.2 – Principales classes flexionnelles du russe selon Brown et Hippisley (2012, p. 48).

partagées entre les classes I et IV. Ce nœud est sous-spécifié pour les autres propriétés. Enfin, certains points communs entre classes ne se déduisent pas de la hiérarchie des classes, mais sont manifestés par des références directes entre les cases de classes. Nous notons ces références par des arcs en pointillés. Par exemple, le génitif pluriel de la classe IV se forme en utilisant la fonction d'évaluation du génitif pluriel de la classe II. Il existe donc deux mécanismes d'héritage d'exposants dans cette structure : par référence directe entre les propriétés, ou par la structure hiérarchique.

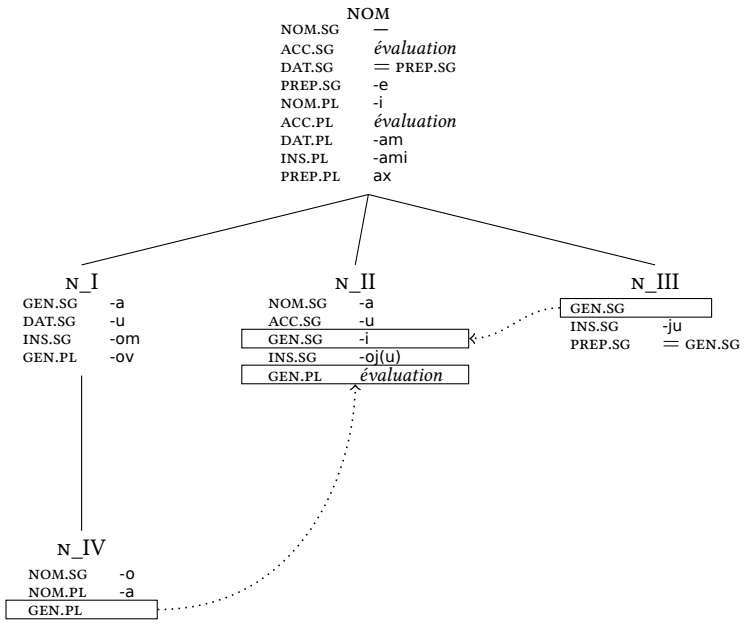
(52)



Dans cette organisation, le choix de nœuds intermédiaires permet des définitions plus succinctes des caractéristiques associés à chaque classe. Si l'on admet le principe de l'héritage par défaut, et que l'on s'interdit tout autre mécanisme d'héritage, on peut mener plus loin l'argument de l'économie descriptive. Si un nœud intermédiaire a n enfants, qui héritent de façon monotone de l'ensemble des propriétés qu'il définit, il est toujours économique de faire remonter les valeurs de l'un des enfants. Si certaines de ses propriétés ne sont pas partagées par les autres enfants, ils pourront les traiter comme des défauts et les réécrire. On peut donc faire remonter l'un de ses enfants sur le nœud N_O pour obtenir une structure par défaut plus com-

pacte. Si l'on considère que plus une classe est grande, plus il est coûteux de lui faire réécrire un défaut, il nous faut faire dépendre le choix de l'enfant à promouvoir de sa fréquence de type. De cette façon, un minimum de lexèmes appartient à des classes qui redéfinissent un défaut. Dans nos données, 577 lexèmes se terminent en /-ov/ au génitif pluriel, terminaison caractéristique de la classe I, tandis que seuls 100 lexèmes présentent un /-o/ au nominatif singulier comme les lexèmes de la classe IV. Nous en déduisons la hiérarchie en (53), qui est plus économique que celle en (52)[3].

(53)



L'argument peut-être poussé plus loin : dans une hiérarchie de classes flexionnelles avec héritage par défaut, si l'on souhaite que la structure arborescente mène à la description la plus économique, alors tous les nœuds internes de la hiérarchie doivent correspondre à des microclasses. De plus, si l'on considère que plus la classe est grosse, plus il est coûteux de réécrire un défaut, alors la hauteur d'une microclasse dans l'arbre doit refléter sa fréquence de type. Enfin, afin de ne réécrire que le nombre minimal de défauts, toute classe doit avoir pour parent la classe plus fréquente à laquelle elle est la plus similaire.

Nous proposons un algorithme très simple permettant de produire une telle hiérarchie à

3. Nous fusionnons ici les niveaux NOMINAL et NOM, car nous ne nous occupons que des noms. Le même raisonnement tiendrait pour une structure qui inclurait les adjectifs, quoi que le résultat puisse être différent.

partir des microclasses définies sur les patrons d'alternances.

1. Calculer la distance entre chaque paire de microclasses.
2. Initialiser l'arbre avec la microclasse la plus fréquente pour racine.
3. Pour chaque autre microclasse par fréquence décroissante, la rattacher dans l'arbre à la microclasse dont elle est la plus proche.
   En cas d'égalité, préférer le nœud le plus haut.

Cet algorithme repose sur une notion de distance entre classes. Nous proposons de mesurer la proximité entre microclasses par la distance de Hamming sur les vecteurs de patrons : la distance entre deux classes est le nombre de paires de cases pour lesquelles ces classes instancient des patrons distincts. Par ailleurs, cet algorithme n'est pas déterministe (il est possible de rencontrer des classes de fréquences identiques, ainsi que des classes également équidistantes de plusieurs autres classes). En conséquence, il existe non pas un unique arbre mais une famille d'arbre par défaut qui répondent à ces critères.

Nous appliquons cette procédure aux noms du russe (segments seuls). Nous ignorons les microclasses de taille 1 afin de faciliter la lisibilité de l'arbre, mais elles pourraient être rattachées aux microclasses les plus proches selon le même algorithme. La figure 4.6 présente la hiérarchie obtenue.

Comparons ces classes avec celles des structures (52) et (53). Dans cette hiérarchie, la classe la plus fréquente, qui constitue la racine et qui est exemplifiée par le nom UM correspond aux lexèmes de la classe I qui présentent une alternance palatale de la consonne finale. Ils instancient par exemple le patron C ⇌ Cʲe / X*[-labi]_ entre le nominatif singulier et le prépositionnel singulier. Au deuxième niveau, BAZA correspond à ceux des noms de la classe II qui présentent également la palatalisation, et instancient par exemple le patron Ca ⇌ Cʲe / X+_ entre nominatif et datif singuliers. Le nœud MNEN'IJO correspond à une variante en /e/ de la classe IV, et instancie par exemple le patron em ⇌ a / X+[-tri -lat][jɕːtɕlʲdʲxʲgʲkʲrʲsʲzʲtʲ]_ entre l'instrumental singulier et le nominatif pluriel. Le nœud NOZH correspond aux lexèmes de la classe I qui ne présentent pas de palatalisation. Le patron qu'ils instancient entre le nominatif singulier et le prépositionnel singulier est ⇌ e / X+[-nas -labi -lat][-syl -lab -rou -labi]_. Le nœud

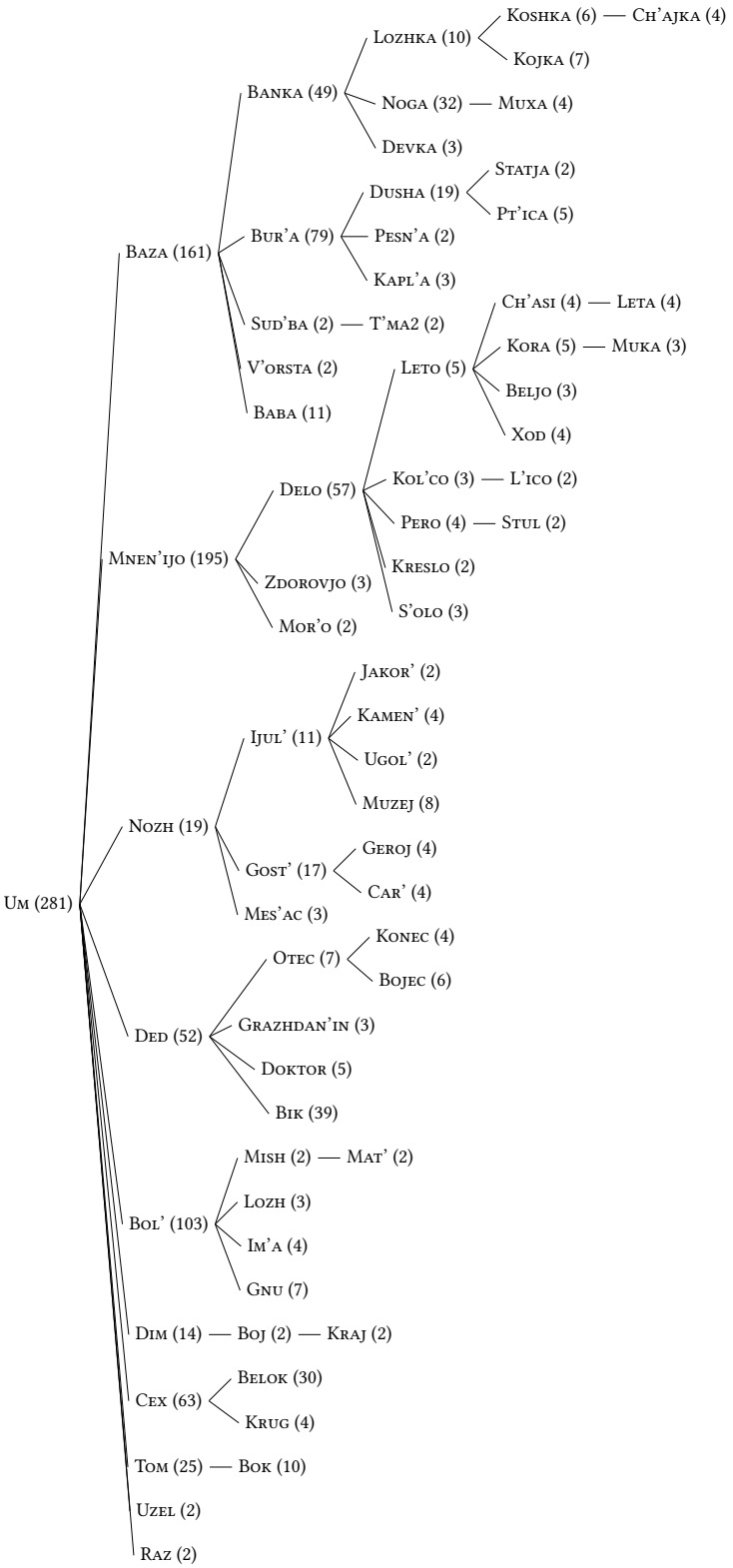


Figure 4.6 – Hiérarchie des classes flexionnelles nominales du russe par défaut (microclasses de plus d'un membre).

DED correspond à des lexèmes de la classe I qui présentent à la fois la palatalisation de la consonne finale et qui présentent un accusatif en /-a/. Ils partagent le même patron que les lexèmes de la microclasse UM pour l'alternance entre nominatif singulier et prépositionnel singulier, mais diffèrent par exemple par le patron qui relie le nominatif et l'accusatif singulier : ⇌ a / X+[-labi][-syl -rou -labi]_. Le nœud BOL' correspond aux lexèmes de la classe III. Entre le nominatif et le génitif singulier, ils instancient le patron ϵ ⇌ i / X*[-lat][-nas -lat]C_.

Il serait trop long de poursuivre cette analyse pour chaque nœud, mais il apparaît clairement que dans une telle hiérarchie, les nœuds qui sont sœurs peuvent avoir des relations de nature variées avec leur mère. Ainsi UM et DED partagent la plupart de leurs patrons d'alternance, tandis que UM et BOL' en partagent beaucoup moins. Cette hiérarchie, quoiqu'elle soit compacte, ne reflète pas très fidèlement la structure de similarité des classes. Elle ne correspond pas à l'intention de la structure en (52), dans laquelle le nœud intermédiaire N_O met en avant un ensemble de similarités saillantes entre deux classes. Quoique les hiérarchies de la Morphologie en Réseau fonctionnent sur le principe de l'héritage par défaut, cet héritage concerne non pas directement les classes flexionnelles, mais l'ordonnancement des règles constructives à travers les classes. Nous souhaitons au contraire produire une organisation arborescente des classes qui explicite leur similarités. La section qui suit explore l'inférence de telles hiérarchies.

4.3.2 Structures hiérarchiques monotones

Bonami (2014) propose de modéliser la structure de similarité entre microclasses en utilisant l'algorithme de classification hiérarchique UPGMA (Sokal et Michener 1958). L'algorithme UPGMA, pour *Unweighted Pair Group Method with Arithmetic Mean*, procède à partir d'une matrice de distances entre les éléments classifiés. Les feuilles de l'arbre représentent les éléments à classer (en l'occurrence, les microclasses). L'algorithme procède par fusions successives de deux nœuds existants, formant un nouveau nœud. La distance entre deux nœuds A et B de tailles respectives $|A|$ et $|B|$ est la moyenne des distances $d(x, y)$ pour toutes les paires d'objets (x, y) tel que $x \in A$ et $y \in B$:

$$\frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{y \in B} d(x, y)$$

Le choix d'une fonction de distance est libre. Bonami (2014) utilise l'algorithme UPGMA pour comparer une classification des microclasses du français fondée sur des exposants (ceux-ci sont trouvés par maximisation des radicaux, sans allomorphie radicale) et une classification fondée sur des patrons d'alternance strictement non préfixaux. Il constate que la seconde classification ressemble de près à la classification traditionnelle du français, tandis que la première produit une classification sans grande ressemblance avec la partition en trois classes décrite par les grammaires descriptives.

La figure 4.7 présente la classification résultante pour les microclasses du français de nos données. Nous ignorons les classes de verbes défectifs, car celles-ci sont très clairement séparées des données en raison des nombreuses alternances pour lesquelles elles sont sous-spécifiées.

Dans cette figure, comme dans les suivantes, les distances entre chaque microclasse et l'ensemble des autres classes sont présentées sous la forme d'une carte thermique fondée sur la matrice des distances. Chaque case de coordonnées i, j de la matrice de distance, représenté par un pixel coloré, indique la distance entre les lexèmes i et j. Puisque $d(i, j) = d(j, i)$, la carte thermique présente une symétrie diagonale. Les pixels les plus clairs indiquent des distances faibles entre lexèmes. La diagonale est blanche car $d(i, i) = 0$. Les pixels les plus foncés indiquent des distances élevées entre lexèmes. Nous indiquons en marge gauche et haute les classifications traditionnelles des lexèmes au moyen de labels colorés dont la signification est spécifiée en légende. La hiérarchie inférée par l'algorithme UPGMA est indiqué en marge haute et gauche de la matrice (les deux arbres dessinés sont l'image miroir l'un de l'autre). L'ordre des lexèmes en ligne et en colonne est identique, et déterminé par l'arbre inféré. Puisque cet ordre minimise les distances entre lexèmes contigus, on peut lire les similarités privilégiées par l'algorithme UPGMA sous la forme de rectangles de couleur claire autour de la diagonale. Enfin, pour des raisons de place, les légendes à droite et en bas n'indiquent pas la totalité des lexèmes exemplaires de microclasses, mais toutes les microclasses sont bien représentés par une cellule colorée.

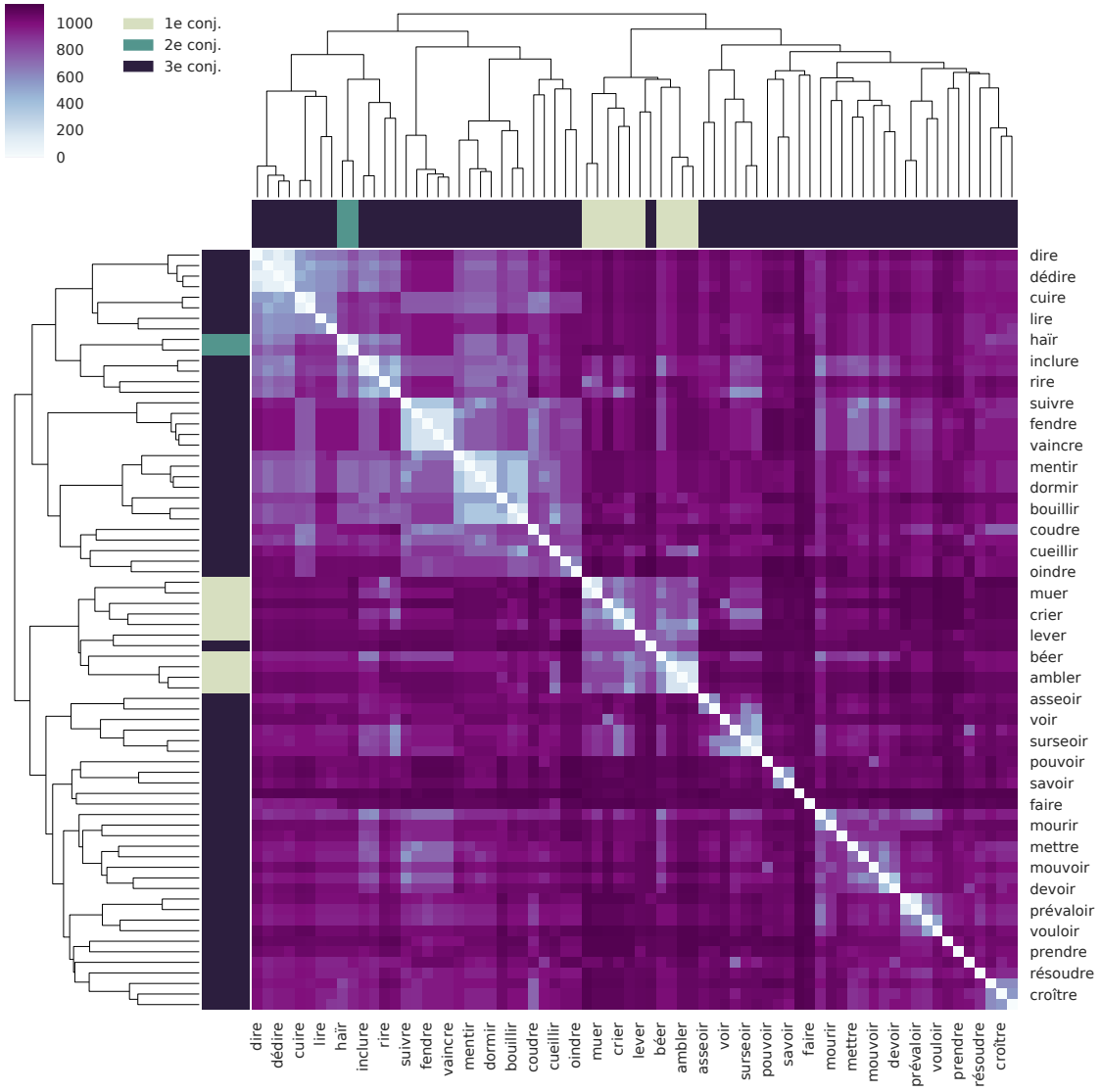


FIGURE 4.7 – Classification hiérarchique des verbes du français par l'algorithme UPGMA.

Comparons la structure obtenue avec la classification traditionnelle des verbes du français. Quoiqu'il soit usuellement considéré comme irrégulier, le verbe ALLER est classé avec les verbes du premier groupe. À cette exception près, tous les verbes du premier groupes sont réunis dans un sous-arbre commun. De même, les verbes du deuxième groupe sont conservés ensemble par la classification. Les verbes du troisième groupe, au comportement très hétérogène, n'apparaissent pas plus proches entre eux que des autres classes dans la classification, mais présentent des similarités plus locales. Dans le tableau, les carrés de couleur claire autour de la diagonale permettent de distinguer des groupes de verbes au sein desquels la distance est basse, par exemple DIRE, LUIRE et REDIRE. On distingue par ailleurs les verbes les moins réguliers par des lignes et colonnes très foncées dans la matrice. C'est le cas de ÊTRE et AVOIR, et dans une moindre mesure de SAVOIR et POUVOIR (similaires entre eux, mais peu similaires aux autres verbes).

D'une langue à l'autre, la structure de similarité entre les classes se traduit par des cartes thermiques très différentes. La figure 4.8 présente la classification obtenue pour les verbes du portugais. La classification traditionnelle, indiquée en marge gauche et haute par des labels colorés, distingue, du plus foncé au plus clair, les verbes dont l'infinitif se termine en -ar (première classe), -er (deuxième classe) et en -ir (troisième classe). Dans l'ensemble, la classification obtenue automatiquement est concordante. Nous avions vu dans la section 4.2 que le réseau des classes du portugais présentait trois clusters très distincts. La matrice de distance révèle ce contraste très fort. On remarque que certains verbes des trois classes traditionnelles, en majorité correspondant à la deuxième classe, se caractérisent par une similarité faible avec l'ensemble des autres verbes. Ces verbes sont groupés ensemble par l'analyse UPGMA, et se rattachent ensuite aux autres verbes de la seconde classe.

Contrastons ce résultat avec celui que l'on obtient sur le chatino de Zenzontepec, que nous avions classé parmi les réseaux peu connectés (figure 4.9). En marge de la matrice, nous indiquons les macroclasses assignées par Campbell (2011), nommées classes A, B et C. Notons que tandis que les systèmes verbaux du français et du portugais permettaient des distances entre 0 et 1000 ou 2000, les quatre cases de paradigme du chatino de Zenzontepec ne donnent lieu qu'à 6 patrons d'alternance. Cette propriété explique l'aspect pixellisé de la matrice de simi-

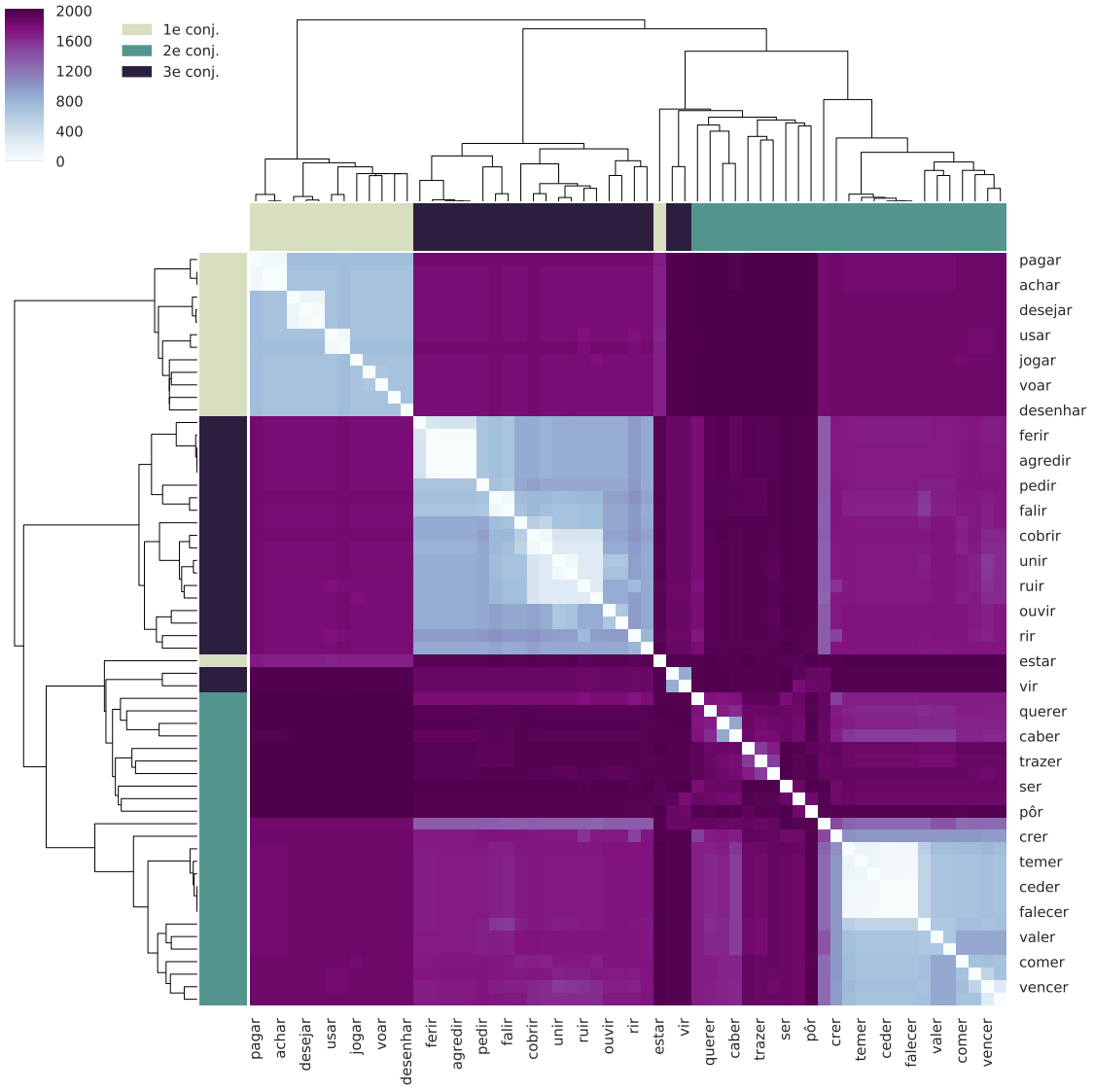


FIGURE 4.8 – Classification hiérarchique des verbes du portugais par l'algorithme UPGMA.

larité du chatino de Zenzontepec. La hiérarchie ne semble pas refléter les classes de Campbell (2011) : la classe C est divisée en deux, les classes A et B sont dispersées dans divers sous-arbres. Par ailleurs, l'observation de la diagonale ne distingue que deux ensembles de lexèmes très cohésifs, le sous-ensemble de la classe C situé en haut à gauche de la hiérarchie et un sous-ensemble de verbes de la classe A situé vers le centre de la matrice. L'orthogonalité du système tonal apparaît clairement dans la hiérarchie de la figure 4.10 : celle-ci est sans relation directe avec celle des exposants segmentaux.

En russe non plus (figure 4.11), la classification obtenue par l'algorithme UPGMA ne correspond pas étroitement aux classes attendues (nous comparons ici aux classes de Corbett (1982), plus précises que celles de Brown (1998)). On observe cependant de nombreux carrés de clarté variable organisés autour de la diagonale. La classe de ŠKOLA (classe II de Brown (1998)), située en haut à gauche de la matrice est remarquablement homogène, de même que celle de KOST' (au centre). Les autres classes sont divisées et plus dispersées. La figure 4.12 compare la classification des propriétés accentuelles des noms du russe avec la classification de Corbett (1982) fondée sur les exposants affixaux. Il apparaît, comme pour le chatino du Zenzontepec, que les deux classifications sont orthogonales. La classification accentuelle semble se diviser plus ou moins en trois grands groupes.

Nous n'avons commenté ici que les systèmes pour lesquels nous disposons d'une classification traditionnelle à laquelle comparer la hiérarchie obtenue. Nous présentons en annexe B l'ensemble des autres classifications hiérarchiques (anglais, arabe, navajo, chatino du Yaitepec).

L'observation des labels colorés en marge des matrices de distances permet de comparer les hiérarchies obtenues avec les classifications communément admises pour ces systèmes. Il est cependant impossible d'interpréter le dendrogramme produit par UPGMA comme une partition en macroclasses. En français, on remarque par exemple des sous-arbres au sein desquels la distance est basse. On pourrait vouloir distinguer une macroclasse englobant {SUIVRE, BATTRE, FENDRE, ROMPRE, VAINCRE}, ainsi qu'une macroclasse {VÊTIR, MENTIR, DORMIR, DESSERVIR, OFFRIR, BOUILLIR, SAILLIR}. Tous ces verbes sont également très proches, et classés sous un même nœud que {ÉCRIRE, COUDRE, CUEILLIR, FEINDRE, OINDRE}. De la même façon, en russe, tous les lexèmes de la classe ŠKOLA sont groupés ensemble (dans le coin haut gauche de la matrice).

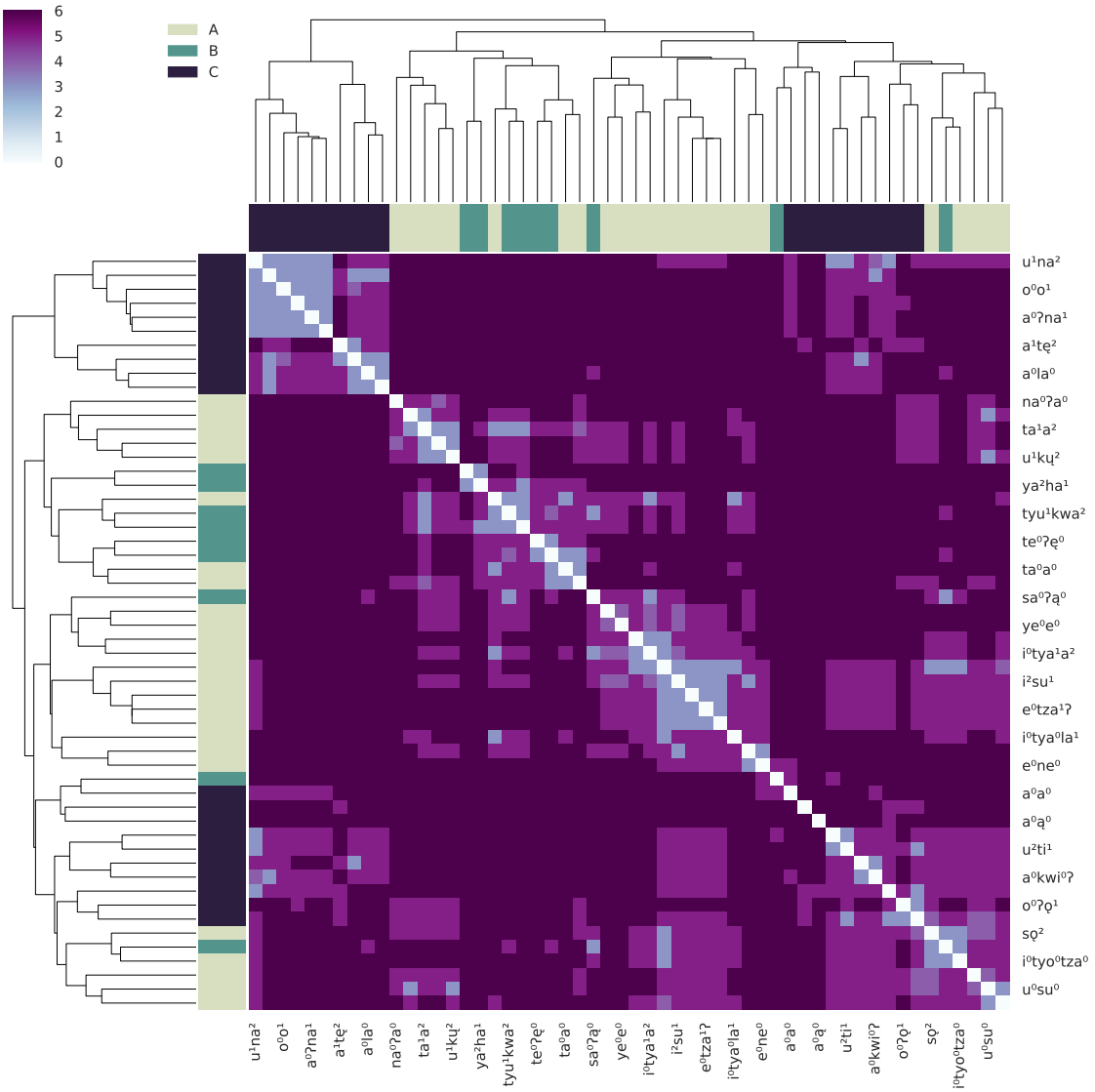


FIGURE 4.9 – Classification hiérarchique des verbes du chatino de Zenzontepec par l'algorithme UPGMA (segments).

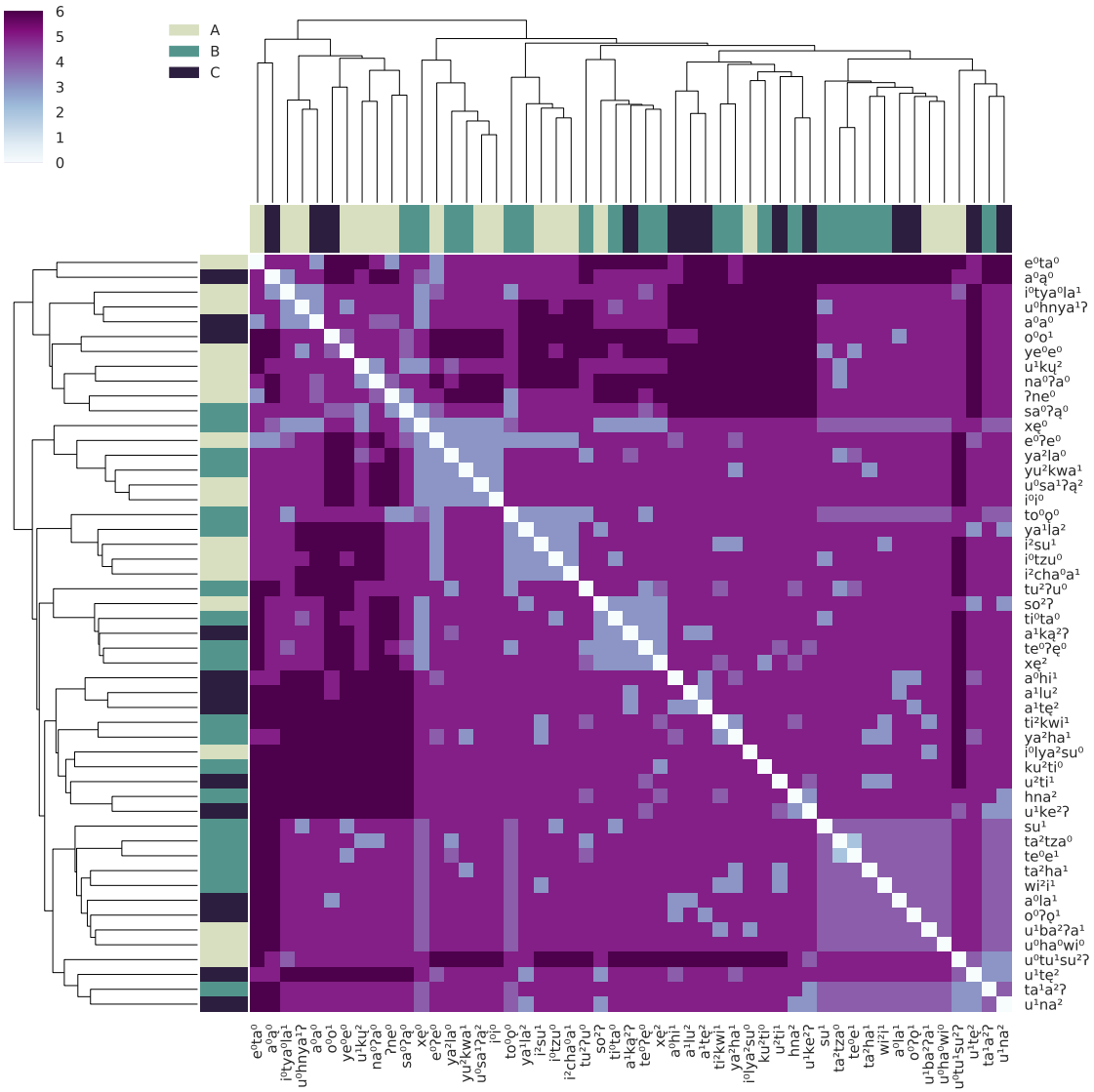


FIGURE 4.10 – Classification hiérarchique des verbes du chatino de Zenzontepec par l'algorithme UPGMA (tons).

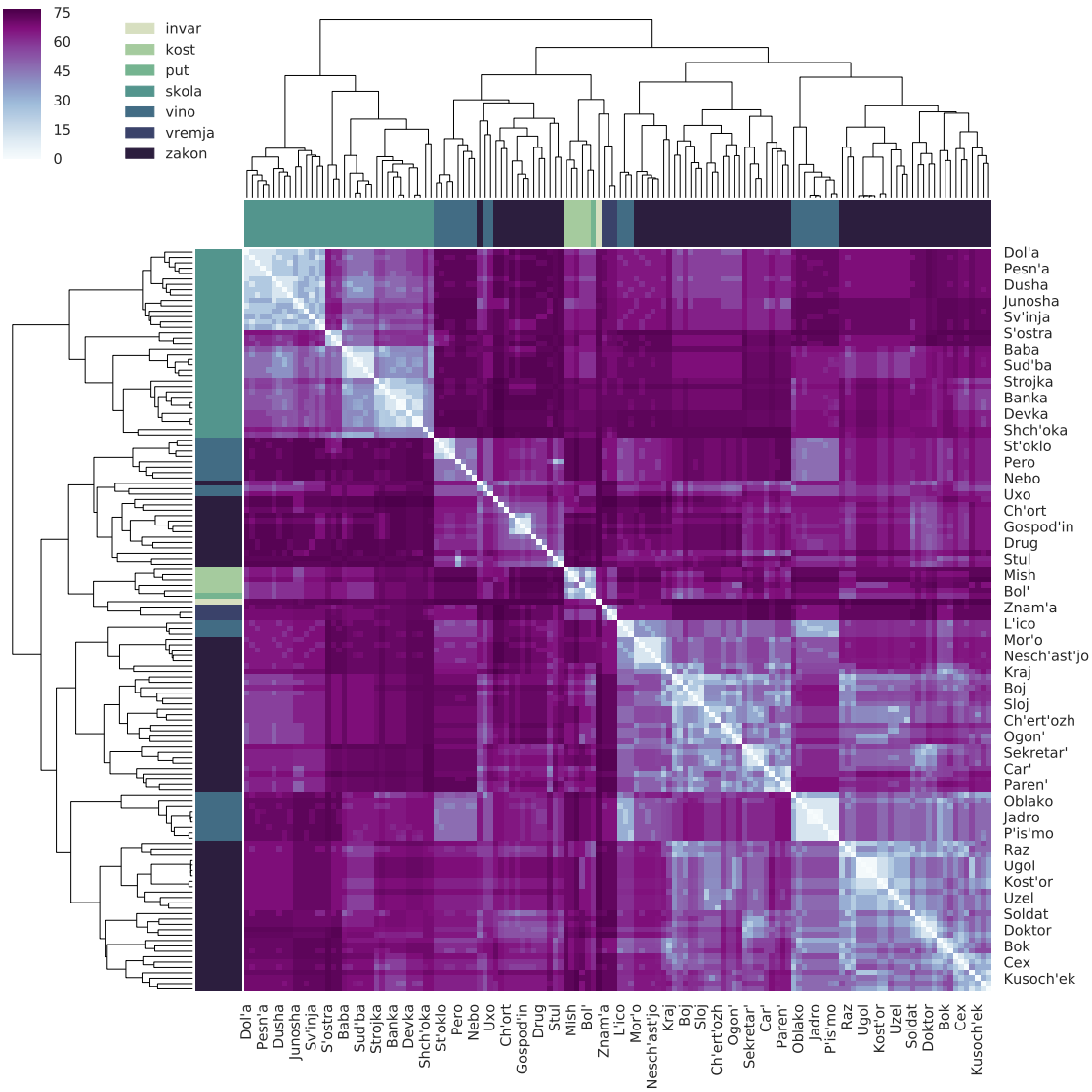


FIGURE 4.11 – Classification hiérarchique des noms du russe par l'algorithme UPGMA (segments).

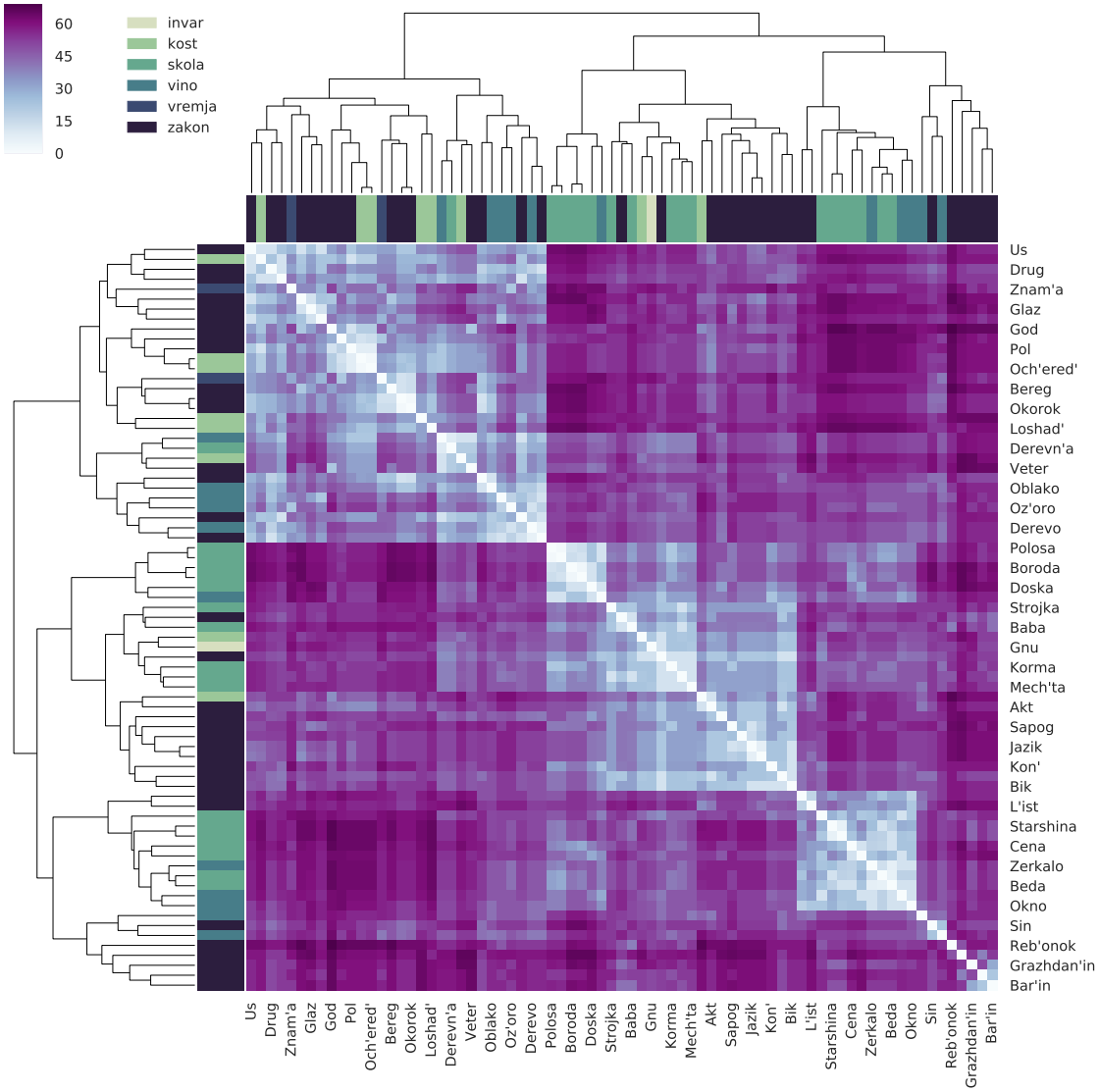


FIGURE 4.12 – Classification hiérarchique des noms du russe par l'algorithme UPGMA (accents).

Cependant, il n'est pas possible de décider s'ils doivent former une unique classe ou deux ou trois classes correspondant aux sous-arbres distincts. L'algorithme UPGMA assigne une distance moyenne à chaque nœud du dendrogramme, mais ne permet pas de décider de nœuds privilégiés que l'on pourrait considérer comme des macroclasses. En effet, la partition qui minimise ces distances est la partition en microclasses. Il nous faudrait un autre critère afin de décider d'une partition en macroclasses.

4.4 Conclusion

Nous avons défini les microclasses comme des ensembles de paradigmes qui sont caractérisés par des vecteurs de patrons identiques. Nous avons vu que les microclasses sont bien plus nombreuses que les classes flexionnelles décrites par les grammaires descriptives. En effet, la moindre variation de surface donne lieu à une microclasse séparée. Nous avons proposé trois façons d'explorer les similarités entre microclasses. Suivant Sims et Parker (2016), elles peuvent être représentées comme des graphes où les nœuds sont des microclasses et où les arcs expriment des similarités entre classes. Trois types de systèmes s'en dégagent : des systèmes peu connectés (chatino de Zenzontepec), des systèmes où apparaissent clairement des sous-graphes mieux connectés (arabe, portugais, russe, tons du chatino de Yaitepec), et des systèmes très connectés, mais qui ne laissent pas deviner de structuration nette (anglais, français, navajo, segments du chatino de Yaitepec).

Nous avons proposé un algorithme permettant de construire des arbres de classes flexionnelles avec héritage par défaut, de façon à minimiser d'une part la taille de l'arbre, et d'autre part le nombre de lexèmes appartenant à des classes qui réécrivent les défauts. Ces arbres permettent de lire, de la racine aux feuilles, un gradient de régularité (au sens de la fréquence de type des microclasses), tout en exposant des relations de similarité saillantes. Cependant, ils ne manifestent pas très précisément les similarités entre classes. Enfin, la distance entre deux microclasses peut être exprimée par la distance de Hamming sur leurs vecteurs de patrons. Nous avons utilisé un algorithme de classification hiérarchique (UPGMA), qui, en combinaison avec l'affichage de la matrice de distances, permet d'offrir une vue détaillée des structures de

similarité des microclasses.

Nous avons vu qu'en français et en portugais, les hiérarchies obtenues correspondent d'assez près aux classes décrites par les grammaires traditionnelles, ce qui n'est pas le cas en chatino de Zenzontepec et en russe. Par ailleurs, il apparaît clairement que les accents et les segments du russe répondent à une classification distincte. Les hiérarchies obtenues ne permettent cependant pas de distinguer une partition en macroclasses, ce qui limite les comparaisons avec les descriptions usuelles de ces systèmes. Le chapitre qui suit poursuit l'analyse des systèmes flexionnels en cherchant la meilleure façon de regrouper les microclasses en macroclasses.

Chapitre 5

Classification des lexèmes en macroclasses

Les grammaires descriptives présentent généralement les classes flexionnelles sous la forme de macroclasses au sein desquelles les lexèmes sont apparentés par la similarité, plutôt que l'identité, de leur comportement flexionnel. Nous avons vu que les MICROCLASSES définies par l'identité des patrons d'alternance sont liées entre elles par des réseaux de similarité, qui peuvent être organisés en hiérarchies. Cependant celles-ci ne peuvent pas être utilisées pour déduire directement des partitions de macroclasses. Ce chapitre présente une méthode pour inférer une partition en macroclasses à partir d'une partition de microclasses définies par des vecteurs de patrons d'alternance [1].

En cherchant automatiquement ces macroclasses, nous souhaitons d'une part évaluer à quel point une telle partition constitue un bon modèle des classes flexionnelles, et d'autre part comparer les classifications obtenues avec les classifications habituellement admises pour les mêmes langues. Peut-on déterminer, sur la seule base des patrons d'alternance, un niveau intermédiaire entre les microclasses et l'ensemble du système, qui présente les propriétés des macroclasses traditionnelles ? Ou ces macroclasses sont-elles un outil de description commode mais simplificateur ? Peuvent elles se déduire de l'observation des propriétés formelles des paradigmes ? Et si l'on peut déterminer une telle partition, quel est son pouvoir descriptif ?

La similarité entre les microclasses étant une propriété graduelle, il nous faut, pour établir une classification catégorique sur cette base nous reposer sur un autre critère. Nous nous

1. Ce chapitre se fonde sur un travail publié dans le *Journal of Language modelling* (Beniamine, Bonami et Sagot 2017). Nous élaborons d'une part en étendant l'analyse à d'autres langues en nous appuyant sur les patrons inférés au chapitre 2, et d'autre part en proposant une évaluation quantitative des macroclasses obtenues.

appuyons ici à cet effet sur la longueur de description, qui implémente le principe du rasoir d'Occam. Intuitivement, le modèle qui optimise au mieux la longueur de description du système flexionnel est par le même biais le modèle qui généralise le mieux sur ces données.

Dans la section 5.1, nous discutons de quelques travaux existants qui se penchent sur l'inférence automatique de classes flexionnelles, et dont nous nous inspirons. Nous décrivons dans la section 5.2 un modèle probabiliste permettant d'évaluer la longueur de description d'une partition en macroclasses. La section suivante (5.3) présente l'algorithme de recherche ascendant que nous employons pour trouver une partition en macroclasse optimale du point de vue de la longueur de description. Dans la section 5.4, nous décrivons les partitions obtenues pour les verbes du français, les verbes du portugais, les noms du russe (segments uniquement), et les verbes du chatino de Zenzontepec. Enfin la section 5.5 présente une évaluation quantitative des macroclasses inférées, comparées aux classes traditionnelles.

5.1 L'Inférence de classes flexionnelles

La tâche d'inférer automatiquement des classes flexionnelles a récemment connu un intérêt croissant.

Une première tentative en ce sens par Goldsmith et O'Brien (2006) utilisait un réseau de neurones pour prédire des exposants à partir de traits. L'espoir des auteurs était que la couche cachée du réseau s'organise en classes flexionnelles. Cependant, les expériences menées sur l'espagnol et l'allemand n'ont pas produit un tel résultat. Beaucoup plus récemment, Malouf (2017) a développé un usage plus prometteur des réseaux de neurones pour modéliser le comportement flexionnel. Les résultats obtenus ne s'interprètent pas non plus directement en termes d'une partition en macroclasses.

Brown et Evans (2012) proposent d'inférer des macroclasses pour le système des noms du russe. Ils évaluent la redondance à travers les paradigmes au moyen d'une distance de compression. Ils opèrent une classification automatique [2] sur cette base en utilisant l'outil CompLearn

2. La tâche computationnelle consistant à inférer des classifications non supervées, soit sous une forme arborescente, soit sous la forme d'une partition, s'appelle « *clustering* » en anglais. En l'absence d'une traduction satisfaisante, nous emprunterons par la suite ce terme à l'anglais.

(Cilibrasi et Vitanyi 2005). Le résultat de CompLearn est un arbre binaire sans racine. Puisque cet arbre est difficilement interprétable comme une partition en macroclasses, Brown et Evans utilisent une série d'heuristiques pour sélectionner un ensemble de nœuds préférés dans l'arbre. Leur but est de valider les hypothèses proposées par Brown (1998), Brown et Hippisley (2012), entre autre le fait que « [l]es noms du russe se répartissent en quatre classes générales [3] ». Leur expériences valident cette hypothèse. Leur approche mesure cependant la similarité entre les formes entières, et non entre les comportements flexionnels à proprement parler. En effet, les distances de compression étant fondées sur les formes, elles captent autant, sinon plus, de la similarité entre les radicaux qu'entre exposants. Il est donc incertain que l'arbre résultant, ou les partitions qui en sont déduites, encode strictement de la structure flexionnelle.

Bonami (2014) propose d'élaborer la stratégie de Brown et Evans (2012) en inférant les réalisations séparément de l'inférence des classes. Il produit des classifications flexionnelles fondées d'une part sur une segmentation affixale des verbes français, d'autre part sur des patrons d'alternance. Il construit des dendrogrammes suivant la méthodologie UPGMA (Sokal et Michener 1958) que nous décrivons au chapitre 4. Notre généralisation de l'algorithme d'inférence des patrons nous a permis d'utiliser cette méthodologie pour un plus grand ensemble de langues. Cependant les distances évaluent la qualité d'une classe, et non d'une partition. Ce faisant, elles ne se prêtent pas directement au choix d'une partition en macroclasses dans l'arbre.

Une autre approche s'appuie sur l'idée que, en théorie, l'ensemble de macroclasses optimal doit fournir la description la plus économique du système flexionnel entier. Cette idée a été explorée par Sagot et Walther (Sagot et Walther 2011 ; Walther et Sagot 2011 ; Walther 2013 ; Sagot et Walther 2013 ; Walther 2016), qui comparent automatiquement des descriptions conçues manuellement. Leurs descriptions sont de type constructives et formées d'une grammaire et d'un lexique morphologique, écrites dans le formalisme Alexina. Elles sont comparées au moyen d'une mesure quantitative de leur économie descriptive, fondée sur la notion de théorie de l'information de LONGUEUR DE DESCRIPTION (Rissanen 1978). Cette approche leur permet de comparer des descriptions concurrentes du français, du maltais, du khaling et du latin. Ils nomment « patrons flexionnels » les différentes macroclasses supposées par les analyses

3. [En anglais dans le texte] « *Russian nouns fall broadly into four paradigm classes* ».

qu'ils évaluent. Par exemple, Sagot et Walther (2011) comparent la longueur de description de quatre descriptions du système flexionnel verbal du français qui contrastent par leur nombre de macroclasses (entre une et 139). Leur travail est une inspiration importante pour notre modèle d'inférence des macroclasses. Cependant, nous voyons deux principales limitations à leur travail. D'une part, le formalisme spécifique sur lequel ils s'appuient pour décrire les différentes théories impose de ne comparer que des grammaires constructives des données, et ne se prête pas à la comparaison d'analyses abstractives. D'autre part, et de façon plus fondamentale, ils ne comparent chaque fois qu'une poignée de grammaires conçues manuellement, et ne peuvent donc explorer qu'une petite fraction de l'ensemble des descriptions possibles.

Notre approche s'inspire également de celle de Lee et Goldsmith (2013). Ces auteurs partent d'une représentation des paradigmes et définissent un algorithme de *clustering* glouton qui utilise le principe de longueur de description minimale (Rissanen 1978) afin de décider quels paradigmes grouper ensemble. L'usage de la longueur de description minimale comme critère pour la tâche de *clustering* constitue une amélioration par rapport aux travaux de Sagot et Walther (2011), Walther et Sagot (2011) et Walther (2013), qui dépend de l'écriture manuelle des descriptions. Cependant, l'approche de Lee et Goldsmith (2013) est marquée par un choix de représentations des paradigmes que nous pensons malencontreux. Ils représentent les formes de paradigme comme des « sacs de lettres », c'est à dire des ensembles de caractère orthographiques. Par exemple, les mots *delay* et *delayed* sont chacun représentés par le même ensemble : {a, d, e, l, y}, ignorant l'existence d'une alternance entre ces formes. Ces représentations ne rendent donc pas toujours compte des variations de surface et ne constituent pas des modèles plausibles des connaissances des locuteurs. Ils présentent également le même problème que la stratégie de Brown et Evans (2012) concernant la distinction entre matériel flexionnel et matériel lexical. Par exemple, l'ensemble de caractères pour le mot *daring* ({a, d, g, i, n, r}) est plus proche de celui de *denigrate* ({a, d, e, g, i, n, r, t}) que de la forme du même lexème *dare* ({a, d, e, r}).

L'approche que nous présentons ci-dessous combine les idées de Bonami concernant l'utilisation de patrons d'alternance pour caractériser le comportement flexionnel des lexèmes, et de Sagot, Walther, Lee et Goldsmith concernant l'usage du critère de longueur de description

minimale pour la décision des *clusters*.

5.2 Évaluer les macroclasses avec la longueur de description

Notre but est d'inférer une partition en macroclasses directement à partir des microclasses. Nous proposons la définition suivante des macroclasses (54) :

(54) Un système de MACROCLASSES est un système optimal d'ensembles de microclasses mutuellement exclusifs.

Cette définition requiert, pour être opérationnalisée, un critère d'évaluation des partitions de microclasses en macroclasses. Nous souhaitons trouver le système de macroclasses qui rend compte au mieux des régularités dans les données. Le processus de classification en macroclasses procède de bas en haut : nous commençons par postuler un système de microclasses, et envisageons à chaque étape de fusionner deux classes existantes. Dans le système initial, chaque microclasse doit être décrite séparément comme ayant un ensemble de patrons. Chaque fusion permet une économie pour tous les patrons que les classes ont en communs, car nous n'avons plus besoin de les spécifier qu'une fois au lieu de deux. Cependant, chaque fois que deux classes à fusionner présentent des patrons distincts, il nous faudra désambiguïser quelles microclasses choisissent quel patron. Suivant le rasoir d'Occam, fusionner des classes peut donc être vu comme bénéfique à la concision tant que le gain obtenu par les patrons communs dépasse la perte due à la désambiguïsationn.

Nous proposons donc qu'une partition d'un ensemble de lexèmes en macroclasses est préférable à une autre si elle mène à une description plus concise du système entier. La raison pour laquelle nous choisissons une description concise n'est pas que la concision est une qualité en elle-même, mais plutôt qu'elle reflète la capacité de la description à capter les régularités dans les données.

Dans les deux sections qui suivent, nous présentons le modèle probabiliste qui nous permet de déterminer la longueur de description, puis l'algorithme de recherche glouton qui utilise ce critère pour trouver le meilleur ensemble de macroclasses pour un ensemble de microclasses donné.

5.2.1 Le principe de longueur de description minimale

La longueur de description minimale (LDM) est un principe permettant de sélectionner un modèle d'un ensemble de données étant donné un espace de modèles déterminé (Rissanen 1984 ; Grünwald 2007). Lorsqu'un ensemble de donnée présente de la structure, celle-ci peut être utilisée pour fournir une description plus concise des données. Différents modèles peuvent capter cette structure avec plus ou moins de succès. La qualité du modèle peut donc être déterminée en regardant la longueur de description des données à l'aise du modèle. Celle-ci comprend d'une part la longueur requise pour décrire le modèle lui-même, et d'autre part la longueur de la description des données par le modèle. Le principe de LDM affirme que le meilleur modèle est celui qui mène à la plus petite description. Il permet de comparer des modèles strictement commensurables et écrits dans le même formalisme (et non n'importe quels modèles arbitraires).

Contrairement à Walther et Sagot (2011), notre perspective n'est pas constructive : nous ne cherchons pas la plus courte grammaire qui génère les données. Nous comparons des descriptions hautement redondantes (chaque case de paradigme figure dans autant de paires qu'il y a d'autres cases de paradigme). Ces descriptions nous sont utiles car elles exhibent tous les comportements flexionnels sur la base desquelles deux lexèmes peuvent être jugés comme appartenant ou non à la même macroclasse. Les descriptions que nous mesurons sont distinctes des modèles que nous jugeons, qui sont simplement des partitions de macroclasses, c'est à dire des ensembles d'ensembles de lexèmes. Cette utilisation de la longueur de description, quoique peut-être moins intuitive, est en fait typique de son usage en inférence statistique, où les descriptions sont construites pour la comparaison des modèles mais n'y sont pas identiques, ni n'ont de valeur inhérente en tant que telles.

5.2.2 Spécification des descriptions

Il nous faut donc définir une description des systèmes de macroclasses en termes de patrons d'alternance, ainsi qu'une façon de déterminer la longueur de ces descriptions. Les données sont constituées d'un ensemble de lexèmes chacun situé dans un espace de traits formé d'un

vecteur de patrons (indexé par paire de cases). La description d'un système de macroclasses flexionnelles doit contenir les quatre composants suivants :

1. Une spécification M qui assigne les lexèmes à des microclasses.
2. Une spécification C qui assigne les microclasses *clusters*, qui sont candidats à constituer des macroclasses.
3. Une spécification $\mathcal{P}$ qui assigne, pour chaque paire de cases, les patrons instanciés dans chaque macroclasse. Pour chaque macroclasse comptant plus d'une microclasse, il existera au moins une paire de cases pour laquelle la macroclasse présente plusieurs patrons.
4. Une spécification de l'information résiduelle R qui ne peut être déduite de l'assignation des microclasses aux macroclasses. Cela revient à spécifier, chaque fois que plusieurs patrons concurrents sont disponibles dans une même macroclasse, quelle microclasse instancie quel patron.

Les deux sous-sections suivantes présentent tout d'abord un exemple détaillé de ces descriptions et de leur longueur, puis une définition formelle du modèle.

5.2.2.1 Idée générale

Afin de mieux comprendre comment ces descriptions peuvent être utilisées pour comparer des systèmes de macroclasse concurrents, nous proposons de considérer un système jouet comprenant les trois verbes français AMENER, BOIRE et DIRE au présent de l'indicatif pluriel. Le tableau 5.1 indique tout à la fois les formes et les patrons qui en sont déduits. Nous omettons les contextes des patrons pour faciliter la lecture.

Chacun de ces trois verbes appartiennent à des microclasses différentes, puisqu'ils ne partagent pas tous les mêmes patrons. Considérons maintenant trois façons de les grouper en macroclasses. Le tableau 5.2 présente une spécification informelle mais détaillée des quatre composants de la description pour trois classifications de ces trois verbes. Dans chaque cas, deux des trois verbes sont groupés en une macroclasse, et le troisième verbe forme seul une macroclasse.

	Formes			Patrons (alternances seules)		
	1PL	2PL	3PL	1PL⇌2PL	1PL⇌3PL	2PL⇌3PL
AMENER	/amənɔ̃/	/aməne/	/amɛn/	ɔ̃ ⇌ e (p_1)	ə_ɔ̃ ⇌ ɛ_ (p_3)	ə_e ⇌ ɛ_ (p_6)
BOIRE	/byvɔ̃/	/byve/	/bwav/	ɔ̃ ⇌ e (p_1)	y_ɔ̃ ⇌ wa_ (p_4)	y_e ⇌ wa_ (p_7)
DIRE	/dizɔ̃/	/dit/	/diz/	zɔ̃ ⇌ t (p_2)	ɔ̃ ⇌ ϵ (p_5)	t ⇌ z (p_8)

TABLEAU 5.1 – Sous paradigmes et patrons pour trois verbes français au pluriel du présent de l'indicatif.

Le tableau met en évidence le fait que la description M, qui assigne les lexèmes aux microclasses, ne varie jamais d'une partition en macroclasse à l'autre. Par ailleurs, dans ce cas précis, l'assignation des microclasses aux macroclasses (C) ne varie pas non plus en longueur d'une partition à l'autre. Cependant, les trois partitions envisagées diffèrent en termes d'assignation des patrons aux macroclasses et en termes d'information résiduelle. La seconde classification groupe ensemble deux microclasses qui partagent un patron, l'assignation des patrons aux macroclasses y est donc plus brève (le patron p_1 y est mentionné une fois plutôt que deux). De même, l'information résiduelle est plus courte, car le cluster fournit une information non ambiguë concernant l'alternance 1PL $\sim$ 2PL, contrairement aux deux autres clusters qui doivent désambiguïser entre p_1 et p_2. Ainsi, la seconde classification semble à vue d'œil présenter une description plus courte, préférable aux deux autres.

Il existe deux autres partitions de macroclasses possibles pour ces données : l'une ne comporte qu'une macroclasse unique, et l'autre coïncide avec la classification en microclasses. Ces deux classifications sont illustrées par le tableau 5.3. Dans le premier cas, toute la désambiguïsation est faite dans le résidu, tandis que P est très court. Dans le second cas, toute la désambiguïsation se fait dans l'assignation des patrons, et aucune dans le résidu. La quantité de caractères écrits dans le tableau donne l'impression que la seconde description est plus courte. Cependant, la première description capte en fait une généralisation que la première ignore. En termes de théorie de l'information, elle est en fait plus courte.

Afin de mesurer précisément la longueur de ces descriptions, il nous faut un schème ex-

Partition	{{AMENER}, {BOIRE, DIRE}}	{{AMENER, BOIRE}, {DIRE}}	{{AMENER, DIRE}, {BOIRE}}
M	AMENER $\mapsto m_1$	AMENER $\mapsto m_1$	AMENER $\mapsto m_1$
	BOIRE $\mapsto m_2$	BOIRE $\mapsto m_2$	BOIRE $\mapsto m_2$
	DIRE $\mapsto m_3$	DIRE $\mapsto m_3$	DIRE $\mapsto m_3$
C	$m_1 \mapsto c_1$	$m_1 \mapsto c_1$	$m_1 \mapsto c_1$
	$m_2 \mapsto c_2$	$m_2 \mapsto c_1$	$m_2 \mapsto c_2$
	$m_3 \mapsto c_2$	$m_3 \mapsto c_2$	$m_2 \mapsto c_1$
$\mathcal{P}$	c_1 : 1PL $\sim$ 2PL : $\{p_1\}$	c_1 : 1PL $\sim$ 2PL : $\{p_1\}$	c_1 : 1PL $\sim$ 2PL : $\{p_1, p_2\}$
	1PL $\sim$ 3PL : $\{p_3\}$	1PL $\sim$ 3PL : $\{p_3, p_4\}$	1PL $\sim$ 3PL : $\{p_3, p_5\}$
	2PL $\sim$ 3PL : $\{p_6\}$	2PL $\sim$ 3PL : $\{p_6, p_7\}$	2PL $\sim$ 3PL : $\{p_6, p_8\}$
	c_2 : 1PL $\sim$ 2PL : $\{p_1, p_2\}$	c_2 : 1PL $\sim$ 2PL : $\{p_2\}$	c_2 : 1PL $\sim$ 2PL : $\{p_1\}$
	1PL $\sim$ 3PL : $\{p_4, p_5\}$	1PL $\sim$ 3PL : $\{p_5\}$	1PL $\sim$ 3PL : $\{p_4\}$
	2PL $\sim$ 3PL : $\{p_7, p_8\}$	2PL $\sim$ 3PL : $\{p_8\}$	2PL $\sim$ 3PL : $\{p_7\}$
R	$m_2 : p_1$	$m_1 : p_3$	$m_1 : p_1$
	$m_3 : p_2$	$m_2 : p_4$	$m_3 : p_2$
	$m_2 : p_4$	$m_1 : p_6$	$m_1 : p_3$
	$m_3 : p_5$	$m_2 : p_7$	$m_3 : p_5$
	$m_2 : p_7$		$m_1 : p_6$
	$m_3 : p_8$		$m_3 : p_8$

TABLEAU 5.2 – Description détaillée de trois classifications des paradigmes du tableau 5.1.

Partition	{{AMENER, BOIRE, DIRE}}	{{AMENER}, {DIRE}, {BOIRE}}
M	AMENER $\mapsto m_1$	AMENER $\mapsto m_1$
	BOIRE $\mapsto m_2$	BOIRE $\mapsto m_2$
	DIRE $\mapsto m_3$	DIRE $\mapsto m_3$
C	$m_1 \mapsto c_1$	$m_1 \mapsto c_1$
	$m_2 \mapsto c_1$	$m_2 \mapsto c_2$
	$m_3 \mapsto c_1$	$m_3 \mapsto c_3$
$\mathcal{P}$	c_1 : 1PL $\sim$ 2PL : $\{p_1, p_2\}$	c_1 : 1PL $\sim$ 2PL : $\{p_1\}$
	1PL $\sim$ 3PL : $\{p_3, p_4, p_5\}$	1PL $\sim$ 3PL : $\{p_3\}$
	2PL $\sim$ 3PL : $\{p_6, p_7, p_8\}$	2PL $\sim$ 3PL : $\{p_6\}$
		c_2 : 1PL $\sim$ 2PL : $\{p_1\}$
		1PL $\sim$ 3PL : $\{p_4\}$
		2PL $\sim$ 3PL : $\{p_7\}$
		c_3 : 1PL $\sim$ 2PL : $\{p_2\}$
		1PL $\sim$ 3PL : $\{p_5\}$
		2PL $\sim$ 3PL : $\{p_8\}$
R	$m_1 : p_1$	
	$m_2 : p_1$	
	$m_3 : p_2$	
	$m_1 : p_3$	
	$m_2 : p_4$	
	$m_3 : p_5$	
	$m_1 : p_6$	
	$m_2 : p_7$	
	$m_3 : p_8$	

TABLEAU 5.3 – Description détaillée de deux classifications opposées pour les paradigmes du tableau 5.1.

plicite pour la description de M, C, $\mathcal{P}$ et R comme des séquences de symboles. Tout message qui prend la forme d'une séquence de symboles présente une distribution de la probabilité de ces symboles via leur fréquence relative dans le message. La théorie de l'information (Shannon 1948) fournit une façon de déterminer la taille en bits du plus court encodage de ce message.

Intuitivement, cet encodage dépend de la longueur du message (toutes choses égales par ailleurs, les messages plus longs sont plus longs à encoder), et de la fréquence des symboles dans le message (les symboles répétés sont moins surprenants et coûtent donc moins cher). Plus précisément, la longueur de la plus courte description possible d'un message m est le produit entre la longueur de ce message et l'entropie de la distribution des symboles S du message.

$$\begin{aligned}\mathrm{DL}(m) &= |m| \cdot H(m) \\ &= -|m| \cdot \sum_{x \in S} P(x) \cdot \log_2 P(x) \\ &= -\sum_{x \in S} \mathrm{count}(x) \cdot \log_2 \frac{\mathrm{count}(x)}{|m|}\end{aligned}$$

La section qui suit présente en détails le schème utilisé pour encoder chaque composante du modèle.

5.2.2.2 Description formelle

Nous détaillons ici la classe de modèles proposés pour modéliser les macroclasses, ainsi que le calcul de leur longueur de description. Comme mentionné précédemment, nous ne cherchons pas à trouver la description qui soit la plus courte possible, mais plutôt à trouver le partitionnement des microclasses en macroclasses qui produise la plus grande réduction de la longueur de description. En conséquence, il nous suffit de mesurer les composants de la description qui sont susceptibles de changer d'un système de macroclasses à un autre.

La description des microclasses est constante à travers toutes les partitions des microclasses en macroclasses pour un système donné. Il n'est donc pas nécessaire de l'inclure, mais nous l'avons fait pour permettre des comparaisons entre des analyses fondées sur des patrons de nature différentes (Beniamine, Bonami et Sagot 2017). Par ailleurs, la longueur de description

définie ci-dessous ne prend pas en compte le nombre de bits nécessaires pour déclarer chaque patron, chaque lexème, le nom des cases et de leurs paires, la description de la procédure pour décoder les données. En effet, aucune de ces informations ne varie d'une partition en macroclasse à l'autre, donc aucune n'est utile pour sélectionner une partition.

À la suite de Sagot et Walther (2011), nous décomposons la longueur de description en plusieurs termes, chacun encodant une partie de la description. Nous définissons la longueur d'une description D d'un système flexionnel comme la somme des longueurs de descriptions des trois composants définis ci-dessous : microclasses, clusters, patrons et résidu.

$$\mathrm{DL}(D) = \mathrm{DL}_M(D) + \mathrm{DL}_C(D) + \mathrm{DL}_P(D) + \mathrm{DL}_R(D).$$

Nous fondons nos exemples sur le système présenté dans les tableaux 5.1, 5.2 et 5.3. Les diagrammes et descriptions explicites, sauf indication contraire, correspondent au partitionnement {{AMENER, BOIRE}, {DIRE}}.

Des lexèmes aux microclasses Nous définissons $\mathrm{DL}_M(D)$ comme le nombre de bits minimum nécessaire pour décrire la correspondance entre l'ensemble de lexèmes et l'ensemble des macroclasses dans la description D. Si nous supposons que l'ensemble des lexèmes $\mathcal{L}$ est ordonné d'une façon pré-définie, la correspondance peut être exprimée simplement par la liste d'identifiants de microclasses parallèles à une liste ordonnée des lexèmes de $\mathcal{L}$.

Nommons $\mathcal{M}$ l'ensemble des identifiants de macroclasses. Si nous définissions $\mathrm{occ}(m)$ comme le nombre d'occurrences d'un identifiant de microclasse $m \in \mathcal{M}$, alors la longueur de description $\mathrm{DL}_M(D)$ peut être définie comme suit :

$$\begin{aligned}\mathrm{DL}_M(D) &= -|\mathcal{L}| \cdot \sum_{m \in \mathcal{M}} \frac{\mathrm{occ}(m)}{|\mathcal{L}|} \cdot \log_2 \frac{\mathrm{occ}(m)}{|\mathcal{L}|} \\ &= -\sum_{m \in \mathcal{M}} \mathrm{occ}(m) \cdot \log_2 \frac{\mathrm{occ}(m)}{|\mathcal{L}|}.\end{aligned}$$

Si l'on applique cette définition à notre exemple, qui contient trois microclasses apparaissant chacune une fois, nous obtenons :

$$\begin{aligned}\mathrm{DL}_M\left(D_{\{\{\text{AMENER, BOIRE}\},\{\text{DIRE}\}\}}\right) &= -3\,\log_2\frac{1}{3}\\ &\approx 4.75\end{aligned}$$

Des microclasses aux macroclasses Nous pouvons considérer de la même façon que l'ensemble des microclasses $\mathcal{M}$ est associé à un ordre pré-défini. Nous pouvons également exprimer la correspondance entre microclasses et macroclasses en listant des identifiants de macroclasses de telle façon à ce que le i^e identifiant de cluster indique le cluster auquel appartient la i^e microclasse. Parallèlement, pour un ensemble de clusters $\mathcal{C}$, nous pouvons écrire :

$$\mathrm{DL}_C\left(D\right) = -\sum_{c\in\mathcal{C}}\mathrm{occ}(c)\cdot\log_2\frac{\mathrm{occ}(c)}{|\mathcal{M}|}.$$

Le nombre d'occurrences $\mathrm{occ}(c)$ d'un cluster $c \in \mathcal{C}$ dans cette partie de la description correspond à sa taille, c'est à dire au nombre de microclasses qu'il contient.

Appliquer cette définition à notre exemple, dans lequel un cluster couvre deux microclasses et le second une seule microclasse, mène à la longueur de description suivante :

$$\begin{aligned}\mathrm{DL}_C\left(D_{\{\{\text{AMENER, BOIRE}\},\{\text{DIRE}\}\}}\right) &= -2\,\log_2\frac{2}{3}\;-\;\log_2\frac{1}{3}\\ &\approx 2.75\end{aligned}$$

Comme mentionné précédemment, cette mesure est identique pour les deux autres partitions du tableau 5.2, car elles présentent la même distribution de clusters :

$$\begin{aligned}\mathrm{DL}_C\left(D_{\{\{\text{AMENER}\},\{\text{BOIRE, DIRE}\}\}}\right) &= \mathrm{DL}_C\left(D_{\{\{\text{AMENER, DIRE}\},\{\text{BOIRE}\}\}}\right)\\ &= \mathrm{DL}_C\left(D_{\{\{\text{AMENER, BOIRE}\},\{\text{DIRE}\}\}}\right)\end{aligned}$$

La longueur de description DL_C est plus petite dans les descriptions qui présentent des clusters grands et peu nombreux, car moins d'informations est nécessaire pour sélectionner le bon cluster pour chaque microclasse. Le cas extrême est la présence d'un unique cluster. En ce cas, la probabilité de ce cluster est 1 et la valeur correspondante pour DL_C est 0. Inversement, DL_C est plus haute lorsqu'il y a de nombreux petits clusters. Ci-dessous, nous fournissons le calcul de DL_C pour les partitions du tableau 5.3 :

$$\mathrm{DL}_C\left(D_{\{\{\text{AMENER, BOIRE, DIRE}\}\}}\right) = -3\ \log_2 \frac{3}{3} = 0$$
$$\mathrm{DL}_C\left(D_{\{\{\text{AMENER}\},\ \{\text{DIRE}\},\ \{\text{BOIRE}\}\}}\right) = -3\ \log_2 \frac{1}{3} \approx 4.75$$

Assignation des patrons aux macroclasses Pour chaque paire de cases dans le paradigme, la description associe les clusters (macroclasses potentielles) aux patrons d'alternance utilisés par les lexèmes de ce cluster. Cette relation n'est pas une fonction : plusieurs patrons peuvent être utilisés par différents lexèmes appartenant au même cluster, et plusieurs clusters peuvent présenter un patron identique.

Soit $\mathcal{K}$ l'ensemble des cases de paradigme. $\mathcal{K}^2$ est alors l'ensemble des n paires de cases, dont nous pouvons considérer qu'elles sont associées avec un ordre pré-défini $\mathbf{k_1} \prec \mathbf{k_2} \prec \ldots \prec \mathbf{k_n}$. Nous référerons à l'ensemble des identifiants de patrons d'alternances par $\mathcal{P}$. La relation entre patrons et clusters peut donc être encodée sous la forme d'une séquence de paires de la forme (c, p), où $c \in \mathcal{C}$ est un identifiant de cluster et $p \in \mathcal{P}$ est un identifiant de patron d'alternance. Plus précisément, puisque $\mathcal{C}$ est aussi associé avec un ordre total, la relation entre patrons et clusters peut être fournie comme suit : tout d'abord, chaque paire (c, p) pour la première paire de cases $\mathbf{k_1}$ peut être fournie, ordonnée selon le cluster ; puis, chaque paire pour $\mathbf{k_2}$ peut être fournie à la suite. Le passage des paires de $\mathbf{k_1}$ à $\mathbf{k_2}$ est visible par le passage du dernier cluster de $\mathcal{C}$ au premier. On peut ainsi fournir à la suite les paires d'identifiants correspondant à toutes les paires de cases.

Décomposons $\mathrm{DL}_P(D)$ en deux éléments : la contribution $\mathrm{DL}_{Pc}(D)$ des identifiants de

clusters, et la contribution $\mathrm{DL}_{Pp}(D)$ des identifiants de patrons. Appelons $\mathrm{occ}_{\mathbf{k}}(c)$ (resp. $\mathrm{occ}_{\mathbf{k}}(p)$) le nombre d'occurrences d'un cluster donné c (resp. d'un patron donné p) dans les paires de la forme (c, p) associées avec une paire de cases $\mathbf{k} \in \mathcal{K}$. Soit N le nombre total de paires de la forme (c, p), i.e. $N = \sum_{c' \in \mathcal{C}} \mathrm{occ}_{\mathbf{k}}(c') = \sum_{p' \in \mathcal{P}} .\mathrm{occ}_{\mathbf{k}}(p')$. La probabilité de l'occurrence d'un cluster donné $c \in \mathcal{C}$ est donc :

$$\begin{aligned} P(c) &= \sum_{\mathbf{k} \in \mathcal{K}^2} \frac{\mathrm{occ}_{\mathbf{k}}(c)}{\sum_{c' \in \mathcal{C}} \mathrm{occ}_{\mathbf{k}}(c')} \\ &= \frac{1}{N} \sum_{\mathbf{k} \in \mathcal{K}^2} \mathrm{occ}_{\mathbf{k}}(c). \end{aligned}$$

En conséquence :

$$\begin{aligned} \mathrm{DL}_{Pc}(D) &= -N \sum_{c \in \mathcal{C}} P(c) \cdot \log_2 P(c) \\ &= -\sum_{c \in \mathcal{C}} \sum_{\mathbf{k} \in \mathcal{K}^2} \mathrm{occ}_{\mathbf{k}}(c) \cdot \log_2 \frac{\mathrm{occ}_{\mathbf{k}}(c)}{N} \end{aligned}$$

De même, la probabilité de l'occurrence d'un identifiant de patron $p \in \mathcal{P}$ est :

$$\begin{aligned} P(p) &= \sum_{\mathbf{k} \in \mathcal{K}^2} \frac{\mathrm{occ}_{\mathbf{k}}(p)}{\sum_{p' \in \mathcal{P}} \mathrm{occ}_{\mathbf{k}}(p')} \\ &= \frac{1}{N} \sum_{\mathbf{k} \in \mathcal{K}^2} \mathrm{occ}_{\mathbf{k}}(p). \end{aligned}$$

En conséquence :

$$\begin{aligned} \mathrm{DL}_{Pp}(D) &= -N \sum_{p \in \mathcal{P}} P(p) \cdot \log_2 P(p) \\ &= -\sum_{p \in \mathcal{P}} \sum_{\mathbf{k} \in \mathcal{K}^2} \mathrm{occ}_{\mathbf{k}}(p) \cdot \log_2 \frac{\mathrm{occ}_{\mathbf{k}}(p)}{N} \end{aligned}$$

La longueur de description $DL_P(D) = DL_{Pc}(D) + DL_{Pp}(D)$ de la section « patrons » de la description peut donc être calculée comme suit :

$$DL_P(D) = -\sum_{\mathbf{k} \in \mathcal{K}^2} \left(\sum_{c \in \mathcal{C}} \mathrm{occ}_{\mathbf{k}}(c) \cdot \log_2 \frac{\mathrm{occ}_{\mathbf{k}}(c)}{N} + \sum_{p \in \mathcal{P}} \mathrm{occ}_{\mathbf{k}}(p) \cdot \log_2 \frac{\mathrm{occ}_{\mathbf{k}}(p)}{N} \right)$$

Si l'on applique cette définition à notre exemple, nous obtenons :

$$\begin{aligned}\mathrm{DL}_P\left(D_{\{\{\text{AMENER, BOIRE}\},\{\text{DIRE}\}\}}\right) = & -2\log_2\frac{1}{2}\\ & -2\log_2\frac{1}{2}\\ & -3\log_2\frac{1}{3}\\ & -\log_2\frac{1}{3}-2\log_2\frac{2}{3}\\ & -3\log_2\frac{1}{3}\\ & -\log_2\frac{1}{3}-2\log_2\frac{2}{3}\\ \approx & \ 14.26\end{aligned}$$

De la même façon, nous avons :

$$\begin{aligned}\mathrm{DL}_P\left(D_{\{\{\text{AMENER}\},\{\text{BOIRE, DIRE}\}\}}\right) &= \mathrm{DL}_P\left(D_{\{\{\text{AMENER, DIRE}\},\{\text{BOIRE}\}\}}\right)\\ &= -10\log_2\frac{1}{3}-8\log_2\frac{2}{3}\\ &\approx 20.52\end{aligned}$$

$$\begin{aligned}\mathrm{DL}_P\left(D_{\{\{\text{AMENER, BOIRE, DIRE}\}\}}\right) = & -2\log_2\frac{1}{2}-6\log_2\frac{1}{3}\\ \approx & \ 11.5\end{aligned}$$

$$\begin{aligned}\mathrm{DL}_P\left(D_{\{\{\text{AMENER}\},\{\text{BOIRE}\},\{\text{DIRE}\}\}}\right) = & -16\log_2\frac{1}{3}-2\log_2\frac{2}{3}\\ \approx & \ 26.52\end{aligned}$$

Comme attendu, la façon la plus efficace d'assigner les patrons à des clusters est de n'avoir qu'un seul cluster, et la pire façon est d'avoir autant de clusters que de microclasses.

Ambiguïté résiduelle Puisqu'un cluster peut être associé avec plusieurs patrons pour une même paire de cases, la réunion de microclasses en clusters peut produire de l'ambiguïté. Une description complète doit rendre compte de l'information nécessaire pour les désambiguïser. L'information nécessaire est distribuée à la fois par paire de cases et par cluster.

Soit un identifiant de cluster $c \in \mathcal{C}$ et une paire de cases $\mathbf{k} \in \mathcal{K}^2$, l'information résiduelle correspondante est fournie sous la forme d'un ensemble de paires de la forme (m, p), où $m \in \mathcal{M}$. Une telle paire signifie que la microclasse m suit le patron p pour la paire de cases $\mathbf{k}$. Bien sûr, seules les microclasses qui appartiennent au cluster identifié par c peuvent, et doivent, être incluses. Puisque cette liste est une information déjà donnée, et puisque les microclasses sont ordonnées, l'information résiduelle pour un cluster c et une paire $\mathbf{k} \in \mathcal{K}^2$ donnée peut être formulée comme une liste de patrons, un pour chaque microclasse de c, dans l'ordre approprié. Dans cette liste, chaque patron p adviendra avec une probabilité $\text{occ}^c_{\mathbf{k}}(p)/\text{occ}(c)$, où $\text{occ}^c_{\mathbf{k}}(p)$ est le nombre de microclasses dans c qui utilisent le patron p pour la paire de cases $\mathbf{k}$. Nous appelons $\mathcal{P}_{\mathbf{k}}(c)$ l'ensemble des patrons qui sont utilisés par au moins une microclasse dans un cluster c pour la paire $\mathbf{k}$. Nous appelons $DL_R(D)$ la longueur de description de l'ambiguïté résiduelle d'une description D.

$$DL_R(D) = \sum_{c \in \mathcal{C}} \sum_{\mathbf{k} \in \mathcal{K}^2} \sum_{p \in \mathcal{P}_{\mathbf{k}}(c)} \text{occ}^c_{\mathbf{k}}(p) \cdot \log_2 \frac{\text{occ}^c_{\mathbf{k}}(p)}{\text{occ}(c)}.$$

Dans notre exemple, pour {{AMENER, BOIRE}, {DIRE}} (5.2), dans le cluster c_1, pour chaque paire de cases présentant des ambiguïtés (1PL $\sim$ 3PL et 2PL $\sim$ 3PL), chacun des deux patrons correspond à une microclasse, résultant en quatre patrons à désambiguïser, de fréquence identique (p_3, p_4, p_6, p_7). On a donc :

$$\text{DL}_R(D_{\{\{\text{AMENER, BOIRE}\},\{\text{DIRE}\}\}}) = -4 \log_2 \frac{1}{2} = 4$$

Nous avons également :

$$\begin{aligned}\mathrm{DL}_R(D_{\{\{\text{AMENER}\},\ \{\text{BOIRE, DIRE}\}\}}) &= \mathrm{DL}_R(D_{\{\{\text{AMENER, DIRE}\},\ \{\text{BOIRE}\}\}}) \\ &= -6\ \log_2 \frac{1}{2} = 6\end{aligned}$$

$$\mathrm{DL}_R(D_{\{\{\text{AMENER, BOIRE, DIRE}\}\}}) = -2\ \log_2 \frac{2}{3} - 7\ \log_2 \frac{1}{3} \approx 12.26$$

$$\mathrm{DL}_R(D_{\{\{\text{AMENER}\},\ \{\text{BOIRE}\},\ \{\text{DIRE}\}\}}) = 0$$

De façon attendue à nouveau, tandis que le regroupement de microclasses tend à faire diminuer DL_P, il tend à augmenter l'ambiguïté, et donc DL_R, tandis que de plus petits clusters mènent à moins d'ambiguïté, et donc une plus petite DL_R. En minimisant la longueur de description totale, nous cherchons un équilibre entre ces mesures. Ainsi, on veut grouper les microclasses autant que possible lorsque le gain lié au regroupement de patrons identiques est préférable à la perte liée à l'ambiguïté produite.

Partition D	$DL_M(D)$	$DL_C(D)$	$DL_{\mathcal{P}}(D)$	$DL_R(D)$	Total $DL(D)$
{{AMENER}, {BOIRE, DIRE}}	4.75	2.75	20.52	6	34.01
{{AMENER, BOIRE}, {DIRE}}	4.75	2.75	14.26	4	25.75
{{AMENER, DIRE}, {BOIRE}}	4.75	2.75	20.52	6	34.01
{{AMENER, BOIRE, DIRE}}	4.75	0	11.50	12.26	28.5
{{AMENER}, {DIRE}, {BOIRE}}	4.75	4.75	26.52	0	36.01

TABLEAU 5.4 – Longueurs de description pour toutes les classifications en macroclasses du tableau 5.1.

Nous regroupons maintenant toutes les longueurs de description partielles pour les systèmes concurrents du tableau 5.1 dans le tableau 5.4 afin de les comparer. Il apparaît que la partition {{AMENER, BOIRE}, {DIRE}} minimise la longueur de description du système.

5.3 Algorithme de recherche

Notre critère pour l'évaluation de la qualité d'une partition en macroclasses d'un système flexionnel est donc DL(D). Afin de trouver la partition optimale selon ce critère, il nous faut explorer l'espace des macroclasses possibles. Évaluer l'ensemble des partitions possibles n'est pas une stratégie réaliste en pratique. Pour un système de 15 microclasses, il existe plus d'un million de partitions différentes à considérer. Pour un système tel que celui du français, où nous trouvons 97 microclasses, le nombre de partitions à considérer approche le nombre d'atomes dans l'univers (10^{80})[4]. La taille de l'espace de recherche rend donc impossible une exploration complète de l'ensemble des possibilités. Ici, nous employons une recherche gloutonne ascendante, qui cherche des macroclasses à partir des microclasses en fusionnant incrémentalement deux clusters. L'algorithme est le suivant :

Algorithm 1 Recherche gloutonne des macroclasses.

1. Commencer avec une partition de clusters D contenant chacun exactement une microclasse.
2. Pour chaque paire de clusters $c_1, c_2 \in D^2$
   - soit $D' = (D \setminus \{c_1, c_2\}) \cup \{c_1 \cup c_2\}$ la partition de clusters où c_1 et c_2 sont fusionnés
   - évaluer la longueur de description $DL(D')$
3. Fusionner les deux clusters $c_1, c_2 \in D^2$ qui produisent la $DL(D')$ la plus basse.
4. Répéter les étapes 2-3 jusqu'à ce qu'aucune fusion binaire ne permette de diminuer la longueur de description.

Nous exemplifions la recherche avec un système imaginaire de cinq macroclasses nommées de A à E. La figure 5.1 illustre la façon dont l'algorithme procède. Les nombres utilisés comme longueur de description ici sont arbitraires et ne servent qu'à illustrer l'algorithme.

4. Le nombre de partitions possibles pour un ensemble de cardinalité n est le n^{ime} nombre de Bell, noté B_n, où $B_0 = 1$ et

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_n.$$

Le nombre de Bell a une croissante extrêmement rapide, bien plus rapide, par exemple, que la fonction exponentielle.

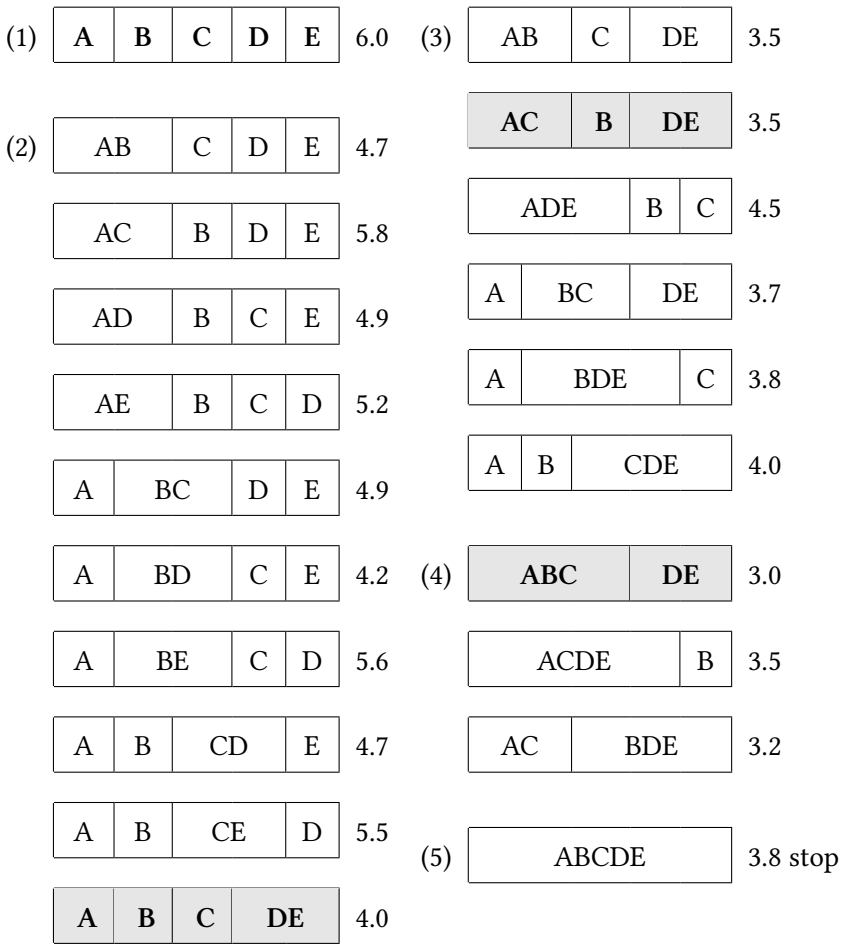


FIGURE 5.1 – Exemple d'une exécution de l'algorithme de recherche.

L'étape (1) correspond à l'état initial, où chaque microclasse forme son propre cluster. Nous supposons arbitrairement que la longueur de description associée est de 6 bits. Dans l'étape (2), nous sélectionnons la paire de microclasse qui produit la plus petite longueur de description, en examinant les dix modèles qui peuvent être obtenus par fusion d'une paire de clusters. La partition correspondante est indiquée en gris dans le tableau. En l'occurrence, la paire dont la fusion produit la plus petite longueur de description est DE ($DL = 4$). L'étape (3) consiste à déterminer la prochaine fusion optimale. Il existe cette fois deux fusions optimales possibles, soit AB, soit BC, les deux menant à un système dont la longueur de description est de 3.5. Dans ce type de situations, nous choisissons aléatoirement l'une des deux solutions, ici AC. L'étape (4) réitère encore ce processus, et trouve une seule fusion optimale, avec une longueur de description de 3. Enfin, l'étape (5) envisage la fusion en un unique cluster, mais la longueur de description obtenue serait de 3.8, ce qui est supérieur à celle du système obtenu à l'étape (4). En conséquence, il n'est plus bénéfique de fusionner des clusters, et nous concluons que la partition $\{A \cup B \cup C, D \cup E\}$ est optimale.

Remarquons qu'il n'existe aucune garantie *a priori* qu'il existera plusieurs macroclasses. Il est possible que la longueur de description continue de diminuer, et que nous obtenions un cluster unique. En conséquence, cet algorithme nous permet d'évaluer empiriquement si un système présente des macroclasses non triviales. Par ailleurs, cet algorithme n'est pas déterministe : à l'étape (3) ci-dessus, nous avons fait un choix aléatoire, ce qui implique qu'un choix différent aurait pû mener à un résultat différent. Nous avons testé empiriquement (Beniamine, Bonami et Sagot 2017) si la variation introduite de cette façon conduit à des variations dans les ensembles de macroclasses. Nous avons établi qu'en français comme en portugais, les variations se réduisent aux fusions initiales et n'ont pas d'impact sur l'ensemble de macroclasses que l'algorithme détermine comme optimal. Par ailleurs, comme pour la plupart des algorithmes gloutons, nous ne pouvons qu'espérer que l'optimum local trouvé par l'algorithme correspond effectivement à un optimum global, et que les macroclasses trouvées sont les meilleures possibles. Nous ne connaissons malheureusement pas de moyen de garantir la découverte de l'optimum global sans examiner l'ensemble des partitions possibles, et nous pensons que l'algorithme présenté ici procède à une exploration raisonnable de cet ensemble.

5.4 Résultats empiriques

Cette section présente le résultat de l'application de l'algorithme décrit dans la section 5.2 aux systèmes pour lesquels il est usuel de reconnaître des macroclasses : les verbes du français, les verbes du portugais, les noms du russe (segments uniquement), et les verbes du chatino de Zenzontepec. L'algorithme employé n'impose pas de trouver des macroclasses non triviales (c'est à dire distinctes des microclasses ou d'un unique cluster), il permet donc de tester si la description en macroclasses est intéressante du point de vue de la longueur de description. Notons que l'algorithme des macroclasses fonctionne exclusivement sur des paradigmes complets : en conséquence, tous les lexèmes défectifs ou surabondants sont ignorés.

La figure 5.4 décrit l'historique des fusions binaires qui ont permis d'inférer les macroclasses des verbes du portugais. les arcs noirs représentent les fusions qui font décroître la longueur de description. Les arcs gris représentent les fusions qui ne l'ont pas fait décroître. En conséquence, les macroclasses sont les nœuds qui dominent les arcs noirs, et au-dessus desquels les arcs sont gris. Ces nœuds sont étiquetés avec le nombre de lexèmes compris dans la macroclasse. Il est important de se souvenir que cet arbre ne peut pas se lire comme les dendrogrammes inférés au chapitre 4 au moyen de l'algorithme UPGMA. D'une part, les nœuds intermédiaires ne représentent pas des sous-classes mais des étapes intermédiaires dans la recherche d'une partition optimale. D'autre part, l'algorithme utilisé n'est pas déterministe, puisqu'il choisit au hasard lorsque deux fusions équivalentes sont possibles. En conséquence, plusieurs exécutions pourraient, en théorie, mener à des résultats différents. Dans Beniamine, Bonami et Sagot (2017), nous avons testé la stabilité des résultats sur 100 itérations pour les verbes du français et du portugais. Dans les deux cas, l'ordre des fusions variait, mais les variations se limitaient aux premières fusions (le bas de l'arbre de l'historique), et les macroclasses étaient stables.

Nous avons vu dans le chapitre 4 que les verbes du portugais semblent naturellement partitionnés en trois grandes classes, avec quelques verbes marginaux. Nous présentons les macroclasses inférées pour ces verbes dans le tableau 5.5. Pour chaque macroclasse, nous donnons le nombre de lexèmes concernés, et listons l'ensemble des microclasses qu'elle contient par fréquence de type décroissante. Nous indiquons également la macroclasse traditionnelle pour

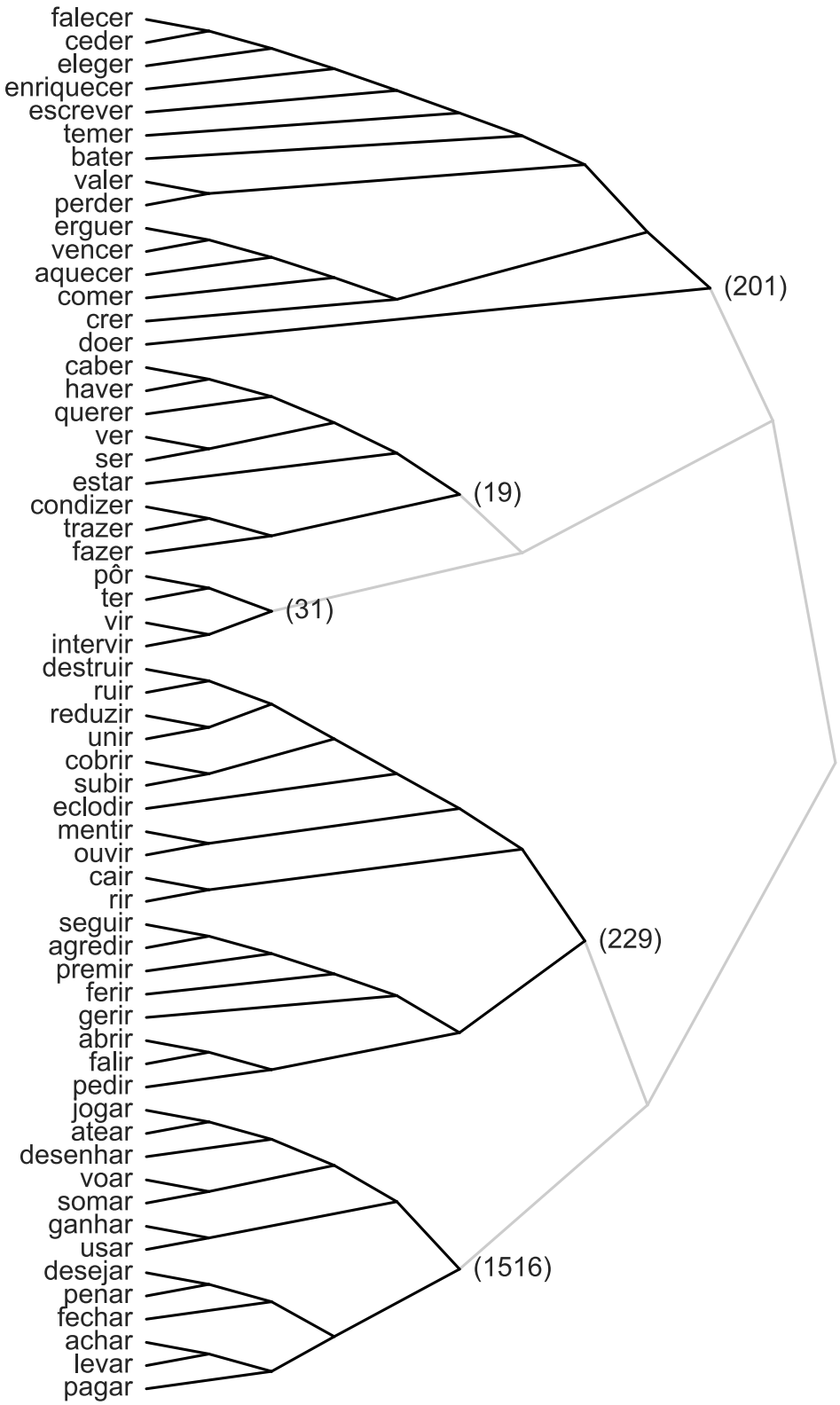


FIGURE 5.2 – Historique des fusions lors de la recherche des macroclasses en portugais.

chaque microclasse. Après chaque lexème exemplaire de microclasse, nous indiquons la taille de la microclasse entre parenthèses. Nous nommons chaque macroclasse par un numéro ainsi que le lexème exemplaire de la plus fréquente de ses microclasses. L'ordre dans lequel les macroclasses sont numérotées est arbitraire.

Dans le tableau 5.5, les trois premières macroclasses (UNIR, USAR, CEDER) correspondent chacune à une classe traditionnelle. Les deux dernières, FAZER et PÔR, plus petites, ne sont sont pas homogènes du point de vue des voyelles thématiques. La quatrième macroclasse (FAZER) groupe ensemble des verbes au comportement très varié, qui ont principalement en commun de ne pas présenter beaucoup de similarités avec le reste des lexèmes. La macroclasse cinq (PÔR) réunit des verbes qui présentent une alternance radicale en /-ɲ/ à l'indicatif passé imperfectif, au subjonctif, et au présent indicatif 1SG. Cette propriété mène à un ensemble d'alternances spécifiques qui les distingue des autres macroclasses, et a un plus grand effet sur la classification que les voyelles thématiques, qui peuvent être dans cette classe /-o/, /-e/ ou /-i/. L'accord entre la classification obtenue automatiquement et la classification usuelle est remarquable (elles sont identiques pour 98% des verbes).

Macroclasses	Taille	Trad.	Taille	Lexèmes
1 : UNIR	229	3e conj.	229	UNIR (90), RUIR (28), SEGUIR (22), FERIR (15), CAIR (11), REDUZIR (11), FALIR (9), SUBIR (9), DESTRUIR (6), MENTIR (5), GERIR (4), PEDIR (4), AGREDIR (3), COBRIR (3), ECLODIR (3), ABRIR (2), RIR (2), OUVIR (1), PREMIR (1)
2 : USAR	1516	1e conj.	1516	USAR (911), LEVAR (305), JOGAR (177), ATEAR (53), VOAR (17), ACHAR (15), PENAR (12), SOMAR (12), DESEJAR (7), DESENHAR (4), FECHAR (1), GANHAR (1), PAGAR (1)
3 : CEDER	201	2e conj.	201	CEDER (78), VENCER (42), COMER (30), BATER (12), ERGUER (9), ESCREVER (8), FALECER (6), DOER (3), AQUECER (2), CRER (2), ELEGER (2), ENRIQUECER (2), TEMER (2), VALER (2), PERDER (1)
4 : FAZER	19	2e conj.	18	FAZER (5), VER (5), CONDIZER (2), QUERER (2), CABER (1), HAVER (1), SER (1), TRAZER (1)
		1e conj.	1	ESTAR (1)
5 : PÔR	31	2e conj.	26	PÔR (17), TER (9)
		3e conj.	5	VIR (4), INTERVIR (1)

TABLEAU 5.5 – Macroclasses inférées pour les verbes du portugais, comparées à la classification traditionnelle.

Toujours au chapitre 4, nous avions vu que les microclasses des verbes du français formaient un réseau plus étroitement lié. Nous présentons les macroclasses inférées en les comparant également à la classification usuelle dans le tableau 5.6. La macroclasse 2 (AXER) correspond étroitement au premier groupe des verbes français. Cependant, elle inclue également le verbe ALLER. Les verbes traditionnellement considérés comme appartenant à la seconde conjugaison, au nombre de 343, appartiennent tous à la même microclasse HAÏR, à l'exception de la variante HAÏR2 (qui présente les formes de présent singulier en /-e/ au lieu de /-ai/). Celles-ci sont bien classées ensemble dans la macroclasse 5 (HAÏR), qui regroupe également des verbes du troisième groupe. Cette classe n'est pas fondamentalement surprenante, voir Plénat (1987) pour une proposition similaire. Enfin, il n'est pas surprenant que le troisième groupe soit réparti dans toutes les macroclasses, puisque son hétérogénéité est connue. Ce résultat est donc également assez conforme à la classification traditionnelle.

En ce qui concerne le russe, le tableau 5.7 compare nos résultats à la classification de Corbett (1982). Les classifications sont également similaires. La Macroclasse 1 (BAZA) comprend les verbes du type ŠKOLA dont l'instrumental singulier est en /-oj/ et contraste avec la macroclasse 3 (BUR'A), présentant les verbes du même type dont l'instrumental singulier est en /-ej/. Les macroclasses 2 (GRAZHDAN'IN) et 4 (UM) correspondent toutes deux au type ZAKON. La macroclasse 2 distingue les lexèmes de ce type qui présentent une alternance de la base entre le singulier et le pluriel, comme dans les exemples (55) à (57).

(55) BAR'IN : barʲin (NOM.SG) ⇌ barʲe (NOM.PL)

(56) DRUG : drug (NOM.SG) ⇌ druzja (NOM.PL)

(57) SIN : sin (NOM.SG) ⇌ sinovja (NOM.PL)

La macroclasse 5 (BOL') regroupe les noms indéclinables, ainsi que ceux des classes KOST' et PUT', qui s'en rapprochent en présentant de nombreux patrons identité. Les noms de la classe VINO se répartissent entre les macroclasses 6 (STUL), 7 (MNEN'IJO) et 8 (S'OLO). Enfin, la très petite macroclasse 9 (IM'A) correspond aux lexèmes du type VREMJA.

Nous présentons dans le tableau 5.8 les macroclasses obtenues pour les alternances segmentales des verbes du chatino de Zenzontepec. Nous les comparons à la classification de Campbell

Macroclasses	Taille	Trad.	Taille	Lexèmes
1 : TENIR	75	3e groupe	75	TENIR (26), PRENDRE (11), COURIR (8), DEVOIR (7), ACQUÉRIR (5), MOUDRE (3), MOUVOIR (3), VIVRE (3), PLEUVOIR (2), VALOIR (2), MOURIR (1), POUVOIR (1), PRÉVALOIR (1), RÉSOUDRE (1), VOULOIR (1)
2 : AXER	4547	1e groupe	4546	AXER (3587), AILLER (243), AMBLER (220), FIER (197), LEVER (98), CRIER (88), MUER (80), BÉER (30), ENVOYER (2), BAYER (1)
		3e groupe	1	ALLER (1)
3 : VOIR	11	3e groupe	11	VOIR (3), BOIRE (2), CROIRE (2), ASSEOIR (1), PRÉVOIR (1), RASSEOIR (1), SURSEOIR (1)
4 : AVOIR	3	3e groupe	3	AVOIR (1), SAVOIR (1), ÊTRE (1)
5 : HAÏR	526	2e groupe	344	HAÏR (343), HAÏR2 (1)
		3e groupe	182	FENDRE (45), FEINDRE (18), METTRE (15), PARAÎTRE (12), BATTRE (11), MENTIR (11), ÉCRIRE (11), OFFRIR (9), OINDRE (8), CROÎTRE (4), SAILLIR (4), COUDRE (3), CUEILLIR (3), DORMIR (3), ROMPRE (3), SUIVRE (3), VÊTIR (3), BOUILLIR (2), DESSERVIR (2), EXCLURE (2), FUIR (2), INCLURE (2), NAÎTRE (2), RIRE (2), VAINCRE (2)
6 : CUIRE	48	3e groupe	48	CUIRE (20), DÉDIRE (7), FAIRE (6), LIRE (4), TAIRE (4), DIRE (2), LUIRE (2), RELUIRE (2), CIRCONCIRE (1)

TABLEAU 5.6 – Macroclasses inférées pour les verbes du français, comparées à la classification traditionnelle.

Macroclasses	Taille	Corbett (1982)	Taille	Lexèmes
1 : Baza	300	skola	300	Baza (161), Banka (49), Noga (32), Baba (11), Lozhka (10), Kojka (7), Koshka (6), Ch'ajka (4), Muxa (4), Devka (3), Su-d'ba (2), T'ma2 (2), V'orsta (2), Kukla (1), Mech'ta (1), S'os-tra (1), Shch'oka (1), Sl'oza (1), Strojka (1), V'osna (1)
2 : Grazhdan'in	9	zakon	9	Grazhdan'in (3), Bar'in (1), Bolgar'in (1), Drug (1), Gos-pod'in (1), Sin (1), Xoz'ain (1)
3 : Bur'a	119	skola	119	Bur'a (79), Dusha (19), Pt'ica (5), Kapl'a (3), Pesn'a (2), Statja (2), Barishn'a (1), D'ad'a (1), Dol'a (1), Junosha (1), Kuxn'a (1), Sem'ja (1), Sv'inja (1), Svech'a (1), T'ot'a (1)
4 : Um	647	zakon	647	Um (281), Cex (63), Ded (52), Bik (39), Belok (30), Tom (25), Nozh (19), Gost' (17), Dim (14), Ijul' (11), Bok (10), Muzej (8), Otec (7), Bojec (6), Doktor (5), Car' (4), Geroj (4), Kamen' (4), Konec (4), Krug (4), Mes'ac (3), Boj (2), Jakor' (2), Kraj (2), Raz (2), Ugol' (2), Uzel (2), Ch'ert'ozh (1), Den' (1), Glaz (1), Kost'or (1), Koz'ol (1), Kusoch'ek (1), L'od (1), Lob (1), Ogon' (1), Pal'ec (1), Paren' (1), Poll'itra (1), Rot (1), Roz-hok (1), Sapog (1), Sekretar' (1), Sloj (1), Soldat (1), Son (1), Storozh (1), Uch'itel' (1), Uch'ot (1), Ugol (1), Ugolok (1), Veter (1)
5 : Bol'	119	kost	111	Bol' (103), Lozh (3), Mat' (2), Mish (2), Loshad' (1)
		invar	7	Gnu (7)
		put	1	Put' (1)
6 : Stul	14	zakon	8	Stul (2), Brat (1), Ch'elovek (1), Ch'ort (1), Muzh (1), Reb'onok (1), Sosed (1)
		vino	6	Pero (4), Koleno (1), Uxo (1)
7 : Mnen'ijo	274	zakon	202	Mnen'ijo (195), Zdorovjo (3), Mor'o (2), Nesch'ast'jo (1), Platjo (1)
		vino	72	Delo (57), Kol'co (3), Kreslo (2), L'ico (2), Jabloko (1), Jadro (1), Oblako (1), Okno (1), Okoshko (1), P'atno (1), P'is'mo (1), Serdco (1)
8 : S'olo	9	vino	9	S'olo (3), Dno (1), Nebo (1), St'oklo (1), Sudno (1), V'odro (1), Z'orno (1)
9 : Im'a	6	vremja	6	Im'a (4), Sem'a (1), Znam'a (1)

Tableau 5.7 – Macroclasses inférées pour les noms du russe, comparées à la classification de Corbett (1982).

(2011). Notre algorithme produit deux macroclasses. La première (U^1NA^2) comporte les classes C de Campbell (2011), ainsi qu'un irrégulier qui appartient selon lui à la classe By. Les classes A et B sont classées ensemble dans la macroclasse 2 (U^0SU^0). Ce dessin semble plus proche de ce que décrit Campbell (2011) que ce que laissait présager la classification hiérarchique fondée sur les distances au chapitre 4.

Pour chacun des systèmes flexionnels étudiés, on trouve donc des macroclasses non triviales. En portugais, en français, en russe et en chatino du Zenzontepec, celles-ci suivent de près les classes décrites dans la littérature.

Macroclasses	Taille	Campbell (2011)	Taille	Lexèmes
1 : U¹NA²	50	C	49	U¹NA² (13), A⁰LA⁰ (7), A¹LU² (5), A⁰TA⁰ (5), A⁰HI¹ (2), A⁰KA⁰ (2), U²TI¹ (2), A¹KĄ²ʔ (1), A¹TĘ² (1), A⁰A⁰ (1), A⁰KWI⁰ʔ (1), A⁰Ą⁰ (1), A⁰ʔNA¹ (1), O⁰HO⁰ʔ (1), O⁰O¹ (1), O⁰ʔO¹ (1), O⁰ʔǪ¹ (1), U¹TĘ² (1), U⁰ʔNE⁰ (1), U⁰ʔWE⁰ (1)
		B	1	YA⁰TĘ⁰ (1)
2 : U⁰SU⁰	341	A	193	U⁰SU⁰ (85), LYA¹ (44), U¹Kų² (16), SǪ² (14), I²SU¹ (9), E⁰NE⁰ (2), I⁰TYA⁰LA¹ (2), TU⁰U⁰ʔ (2), TYA²NA⁰ (2), U⁰TI¹Kų² (2), E⁰TA⁰ (1), E⁰TZA¹ʔ (1), E⁰ʔE⁰ (1), I⁰CHI¹ (1), I⁰I⁰ (1), I⁰TYA¹A² (1), I⁰TYO⁰TZA⁰ (1), NA⁰ʔA⁰ (1), NE⁰E⁰ (1), TA¹A² (1), TA⁰A⁰ (1), TYE¹ʔĘ² (1), U⁰TU¹SU²ʔ (1), YE⁰E⁰ (1), ʔNE⁰ (1)
		B	148	SU¹ (70), YA²HA¹ (27), TO⁰Ǫ⁰ (22), YA¹A² (21), TYU¹KWA² (4), SA⁰ʔĄ⁰ (2), TE⁰ʔĘ⁰ (1), YU⁰TE⁰ʔ (1)

Tableau 5.8 – Macroclasses inférées pour le chatino du Zenzontepec, comparées à la classification de Campbell (2011).

5.5 Évaluation

Cette section se propose d'évaluer les résultats des classifications obtenues dans la section précédente, et de comparer plus formellement les différentes classifications disponibles pour décrire les systèmes de classe étudiés. Notre but n'était pas de reproduire les classifications connues pour ces systèmes, mais de découvrir des classifications qui reflètent les régularités des données. Il n'est donc pas possible, comme c'est parfois le cas dans les tâches de *clustering*, d'évaluer les macroclasses obtenues en les comparant à une classification de référence idéale. Notre évaluation doit répondre aux questions suivantes, qui portent tant sur les classes inférées automatiquement que sur les classifications proposées dans la littérature :

1. À quel point l'ensemble du système reflète-t-il les distances entre microclasses que nous avons observées au chapitre 4 ?
2. Est-il raisonnable de penser que l'optimum local de longueur de description découvert par la stratégie de recherche est proche de l'optimum global ?
3. À quel point chaque classe est-elle cohérente et distincte des autres classes ?
4. À quel point les classifications obtenues diffèrent-elles des classifications traditionnelles ?
5. À quel point les macroclasses nous informent-elles sur les microclasses ?

5.5.1 Distances et macroclasses

Tout d'abord, nous pensons qu'un bon système de macroclasses devrait refléter la structure de similarité des microclasses. Afin de visualiser les microclasses, nous appliquons une méthode de positionnement multidimensionnel (MDS), qui nous permet de projeter la matrice de distances entre les microclasses sur un espace à deux dimensions. Pour chaque langue, les figure 5.3 à 5.6 représentent chaque microclasse par un point dont la taille est proportionnelle au nombre de lexèmes qu'elle contient. Les macroclasses sont indiquées par la coloration des points.

Comme dans les visualisations présentées au chapitre 4, les verbes du portugais apparaissent très nettement partagés en trois principales classes. La visualisation montre que ces

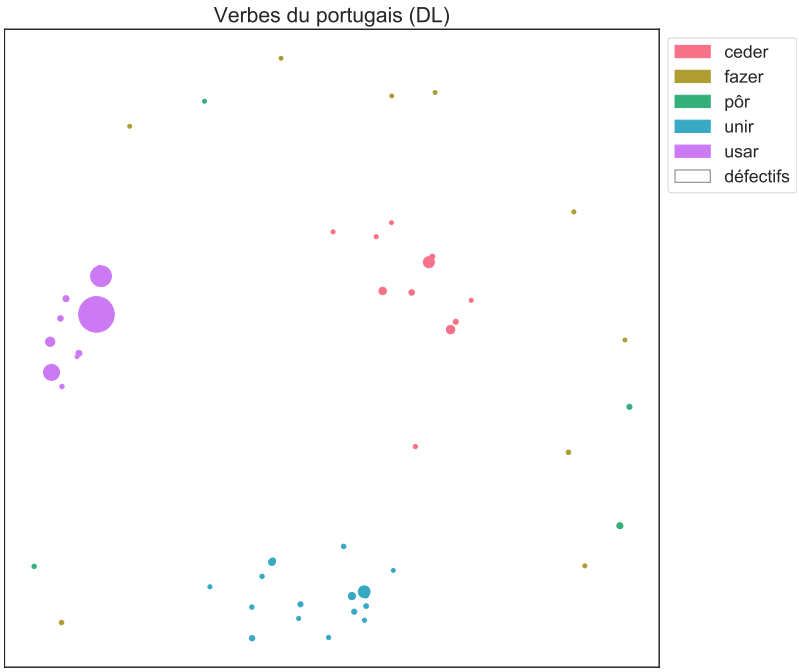


FIGURE 5.3 – Visualisation des classes inférées en portugais.

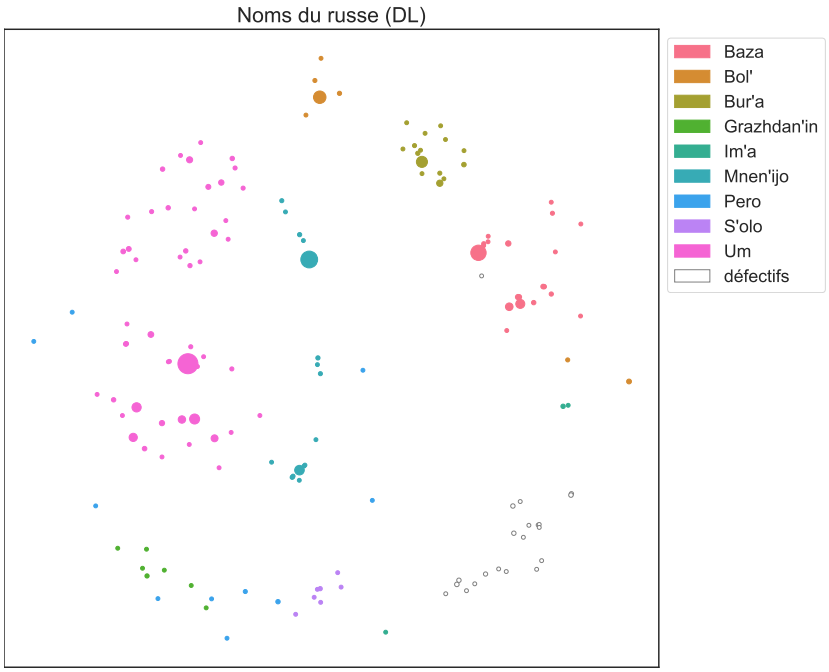


FIGURE 5.4 – Visualisation des classes inférées en russe.

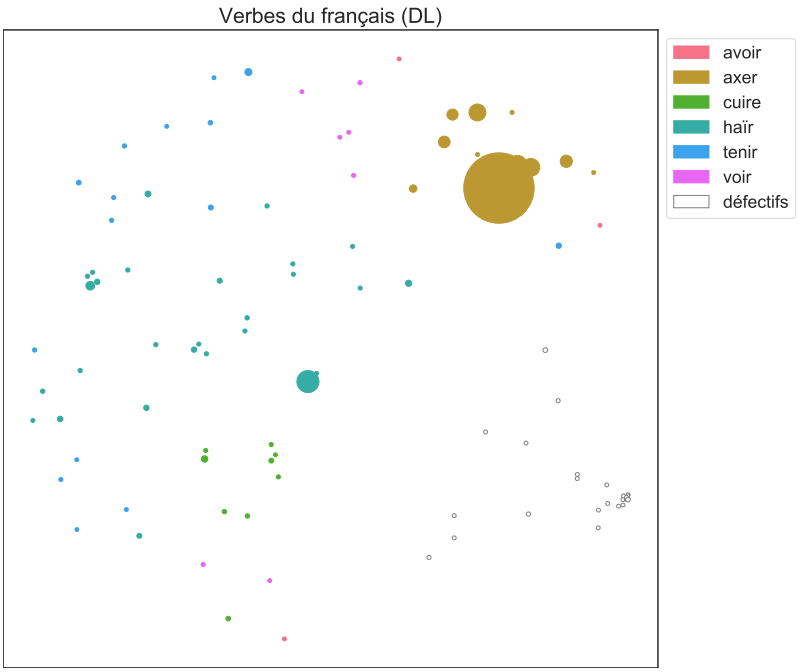


FIGURE 5.5 – Visualisation des classes inférées en français.

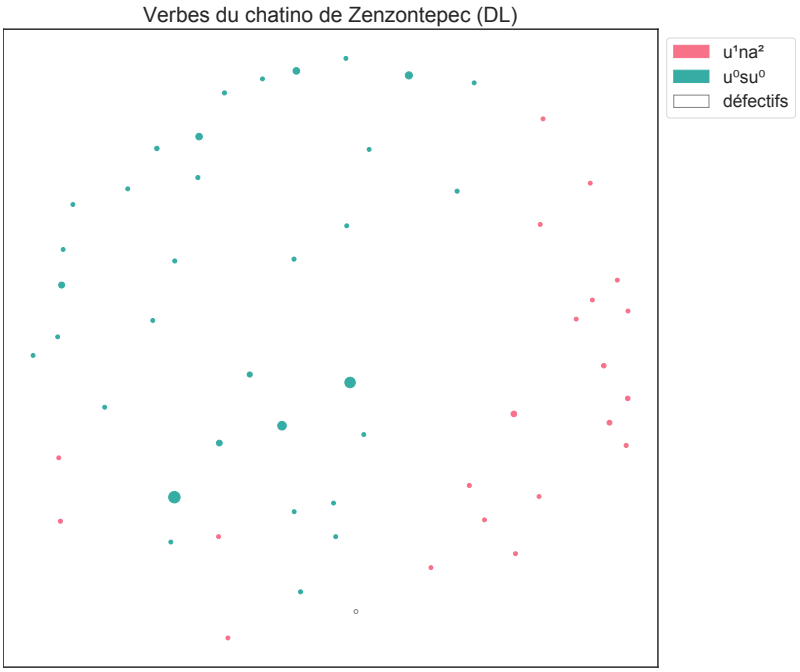


FIGURE 5.6 – Visualisation des classes inférées en chatino de Zenzontepec.

classes sont bien captées par la classification au moyen des longueurs de description. Nous trouvons également deux classes formées de points plus dispersés. En russe et en français, la classification en macroclasses obtenue avec la longueur de description fait bien apparaître ensemble des microclasses voisines, mais l'espace est moins clairement divisé en macroclasses. En russe et en français, l'ensemble des points défectifs, non classés par l'algorithme, apparaît très homogène, et semble former une classe à part. En effet, la défectivité d'un lexème pour une case de paradigme se répercute sur l'ensemble des alternances à laquelle cette case appartient, ce qui mène à de nombreux points de similarité entre lexèmes défectifs dans la mesure où les cases défectives tendent à être les mêmes d'un lexème à l'autre (Boyé 2000). En chatino de Zenzontepec, les classes apparaissent réparties de façon homogène dans l'espace, et la classification automatique a opéré une coupe presque linéaire dans le plan. Rappelons nous que la projection des distances sur un espace à deux dimensions ne peut pas faire ressortir toutes les relations multidimensionnelles entre les classes. En conséquence, la dispersion de certaines macroclasses n'indique pas nécessairement qu'elles soient mal conçues.

5.5.2 Cohésion et distinctivité des macroclasses

Afin d'observer plus précisément l'adéquation des macroclasses inférées dans ce chapitre, nous évaluons celles-ci au moyen des scores de silhouette (Rousseeuw 1987 ; Kaufman et Rousseeuw 1990). Le score de silhouette pour une microclasse m attribuée à une macroclasse $\mathcal{M}$ dépend de sa distance moyenne $a(m)$ aux autres microclasses de $\mathcal{M}$ [5], et de sa distance moyenne $b(m)$ aux observations de la classe $\mathcal{M}'$ dont m est le plus proche :

$$\text{sil}(i) = \frac{b(m) - a(m)}{max(\{b(m) - a(m)\})}$$

Dans tous les cas, $-1 \leq \text{sil}(m) \leq 1$. Le score $\text{sil}(m)$ évalue l'adéquation de l'observation m à sa classe $\mathcal{M}$. Un score bas indique une microclasse mal classée, et un score haut indique une microclasse bien classée. La largeur de silhouette d'une macroclasse $\mathcal{M}$ est la moyenne des scores de silhouette des microclasses qui la composent. De même, le score de silhouette d'une partition entière est la moyenne des largeurs de silhouettes de ses macroclasses.

5. Si $A = \{m\}$, alors $a(m)$ est défini à 0.

Cette mesure est intéressante car elle permet de quantifier les principes d'homogénéité interne et d'hétérogénéité externe des classes proposés par Corbett (2009), déjà évoqués au chapitre 1 et que nous rappelons ci-dessous :

> (principe I) [Les classes] sont **distinctives**, c'est à dire qu'elles présentent
> (critère 1) à travers les classes, des paradigmes concrets distincts,
> (critère 3) à l'intérieur d'une classe, des paradigmes concrets identiques.

Des macroclasses canoniques suivant ces critères auraient un score de silhouette de 1. En effet, dans un système canonique, en vertu du critère 3, pour toute microclasse m, on a $a(m) = 0$. Le score de silhouette de chaque macroclasse canonique est donc $\frac{b(m)-0}{max(\{b(m)-0\})} = \frac{b(m)}{max(\{b(m)\})} = 1$. Cette propriété est vraie également des microclasses. Le score de silhouette pénalise donc une microclasse qui est plus proche d'une autre macroclasse que de sa propre macroclasse : d'un point de vue des distances, on peut alors penser que cette microclasse est assignée à la mauvaise macroclasse. Le score de silhouette ne pénalise pas en revanche la violation du critère (1) dans un système de microclasses.

Les figures 5.7 à 5.10 comparent les silhouettes des systèmes de macroclasses inférés dans ce chapitre aux systèmes de macroclasses établis par des linguistes ou aux systèmes traditionnels pour les mêmes langues. Dans ces figures, chaque macroclasse est représentée dans une couleur distincte et étiquetée par un lexème exemplaire. Une ligne rouge indique le score de silhouette moyen pour l'ensemble du système. En russe (5.7), la classification proposée par Corbett (1982) ignore les noms indéclinables, ainsi qu'un petit nombre de lexèmes qui ne suivent aucune des déclinaisons décrites. Nous associons ces lexèmes respectivement aux macroclasses « irreg » et « invar ». Le système inféré automatiquement présente des silhouettes plus satisfaisantes que les types décrits par Corbett (1982), et le score moyen est supérieur. Cependant, les silhouettes sont imparfaites, les classes de BRAT et BELOK comprenant chacune des microclasses aux scores négatifs.

En portugais (5.8), la seconde conjugaison de la classification traditionnelle comporte des microclasses mal classées. La classification en macroclasses obtenue automatiquement est nettement plus adaptée aux données. Cependant la classe de FAZER présente des lexèmes hétérogènes.

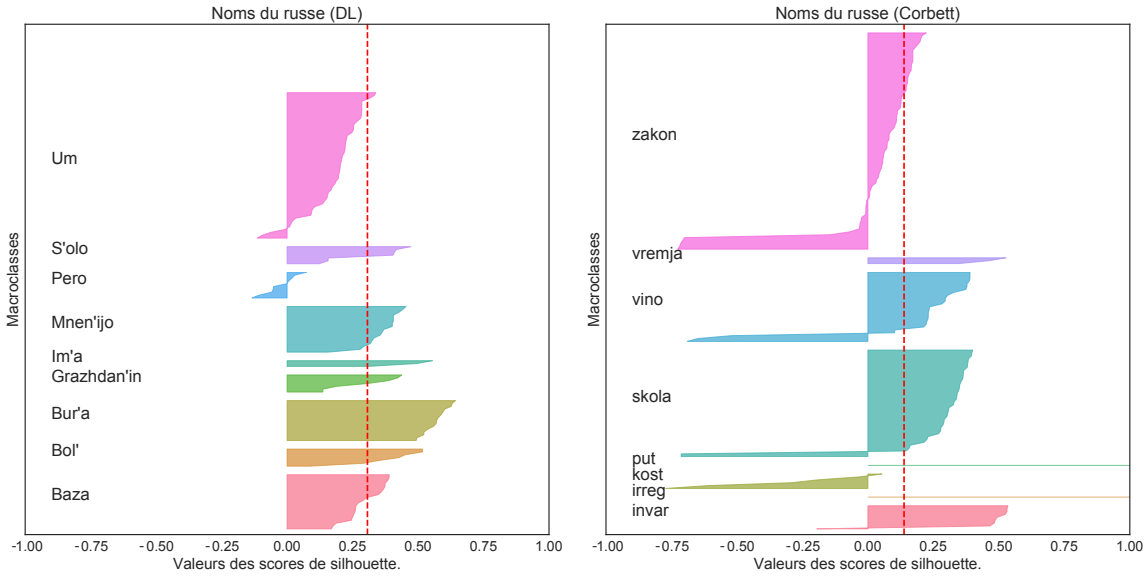


FIGURE 5.7 – Comparaison des scores de silhouette pour les macroclasses des noms du russe.

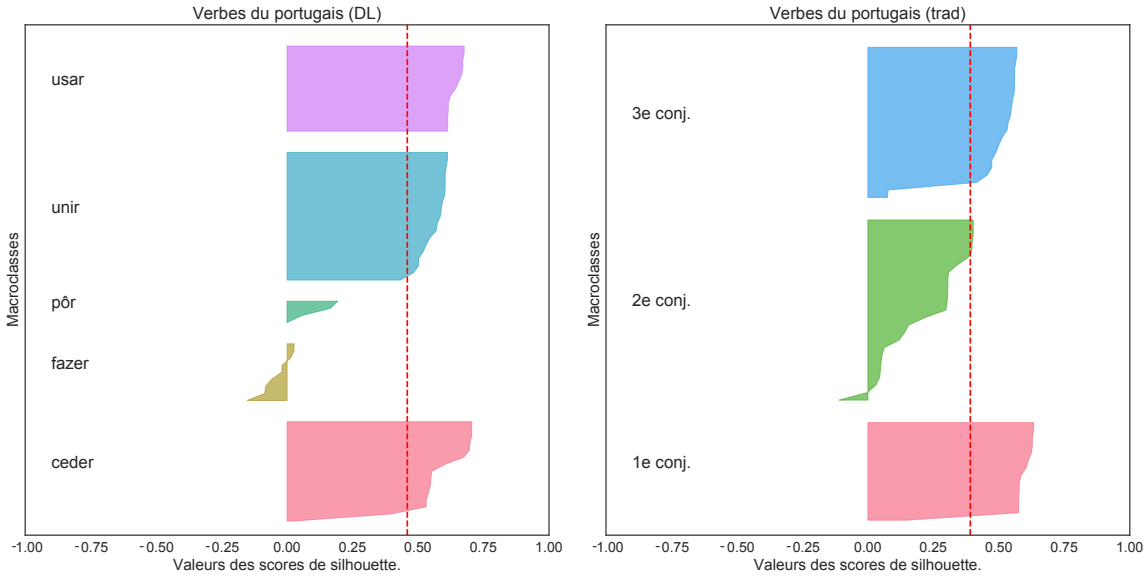


FIGURE 5.8 – Comparaison des scores de silhouette pour les macroclasses des verbes du portugais.

En français, la seconde conjugaison de la classification traditionnelle est tellement homogène qu'elle n'est constituée que de deux microclasses. Son score de silhouette est très haut. La première conjugaison présente un profil moins bon, et la troisième, formée de verbes de types très variables, présente un profil presque entièrement négatif. La conséquence en est un score négatif pour cette classification. Par contraste, la classification obtenue automatiquement présente des profils beaucoup plus équilibrés.

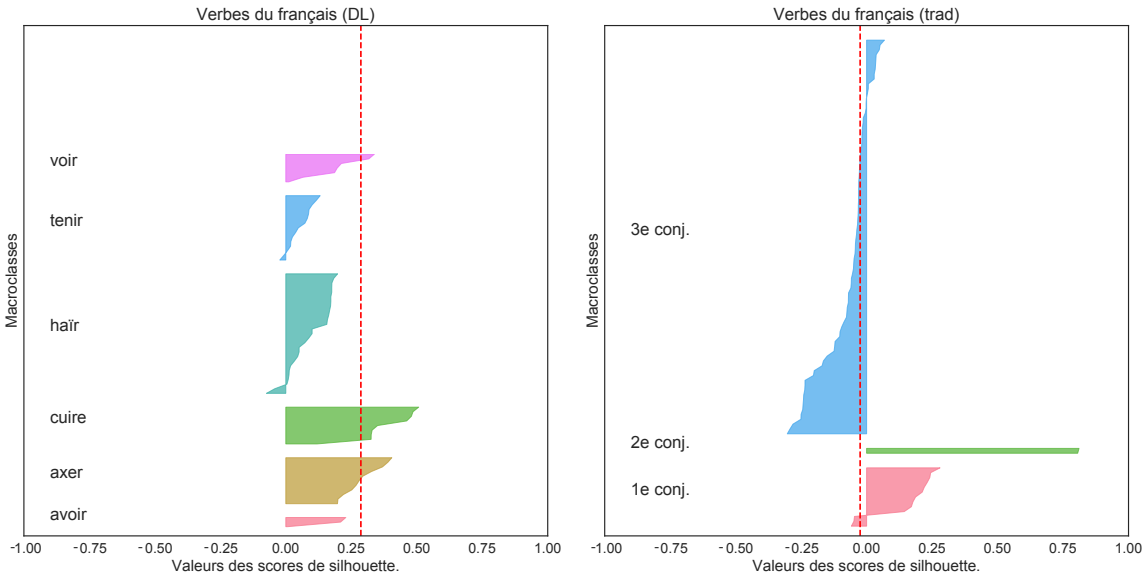


FIGURE 5.9 – Comparaison des scores de silhouette pour les macroclasses des verbes du français.

En chatino de Zenzontepec, nous comparons la classification obtenue aux deux niveaux de la classification de Campbell (2011). La classification des super-classes présente un meilleur profil que celle des sous-classes, mais également que la bipartition obtenue par notre algorithme. Cependant, dans les trois cas, la qualité des classes est assez basse.

Le tableau 5.9 présente les scores de silhouette pour chaque système de macroclasses. Pour les systèmes concurrents, les meilleurs évaluations sont indiquées en gris. Du point de vue des scores de silhouette, les systèmes obtenus en minimisant les longueurs de descriptions sont meilleurs, sauf pour le cas du chatino de Zenzontepec.

Les scores de silhouette sont parfois utilisés pour sélectionner un nombre de classes pour les algorithmes de classification qui dépendent d'un tel paramètre, ou pour sélectionner une

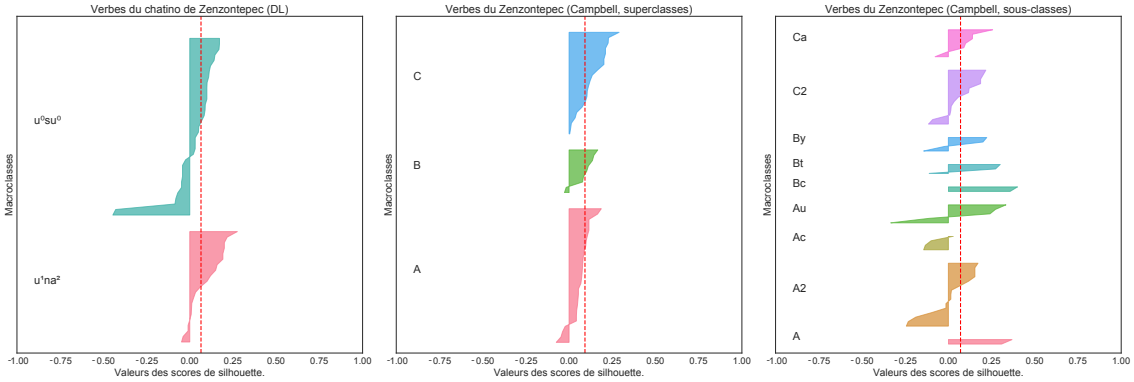


FIGURE 5.10 – Comparaison des scores de silhouette pour les macroclasses des noms du chatino de Zenzontepec.

C	sil(C)
Noms du russe (Corbett 1982)	0.138
Noms du russe (DL)	0.307
Verbes du Zenzontepec (Campbell 2011, sous-classes)	0.070
Verbes du Zenzontepec (Campbell 2011, superclasses)	0.092
Verbes du chatino de Zenzontepec (DL)	0.065
Verbes du français (traditionnel)	-0.025
Verbes du français (DL)	0.287
Verbes du portugais (traditionnel)	0.391
Verbes du portugais (DL)	0.459

TABLEAU 5.9 – Scores de silhouette pour les systèmes de macroclasses.

partition dans la hiérarchie obtenue par un algorithme de classification hiérarchique. Nous avons évalué l'évolution du score de silhouette au fur et à mesure des fusions binaires opérées dans les dendrogrammes obtenus avec la méthode UPGMA (Sokal et Michener 1958) au chapitre 4. La figure 5.11 montre comment ces scores évoluent en fonction du nombre de classes. Il apparaît clairement que ces scores augmentent presque linéairement avec le nombre de classes, si bien que les systèmes optimaux sont des systèmes qui ne fournissent pas ou peu de généralisation sur les microclasses. Les scores de silhouette ne sont donc pas adaptés à la sélection d'un nombre spécifique de macroclasses. Il nous faudrait pour ce faire une mesure qui pénalise un score $b(m)$ bas (c'est à dire la ressemblance entre deux macroclasses).

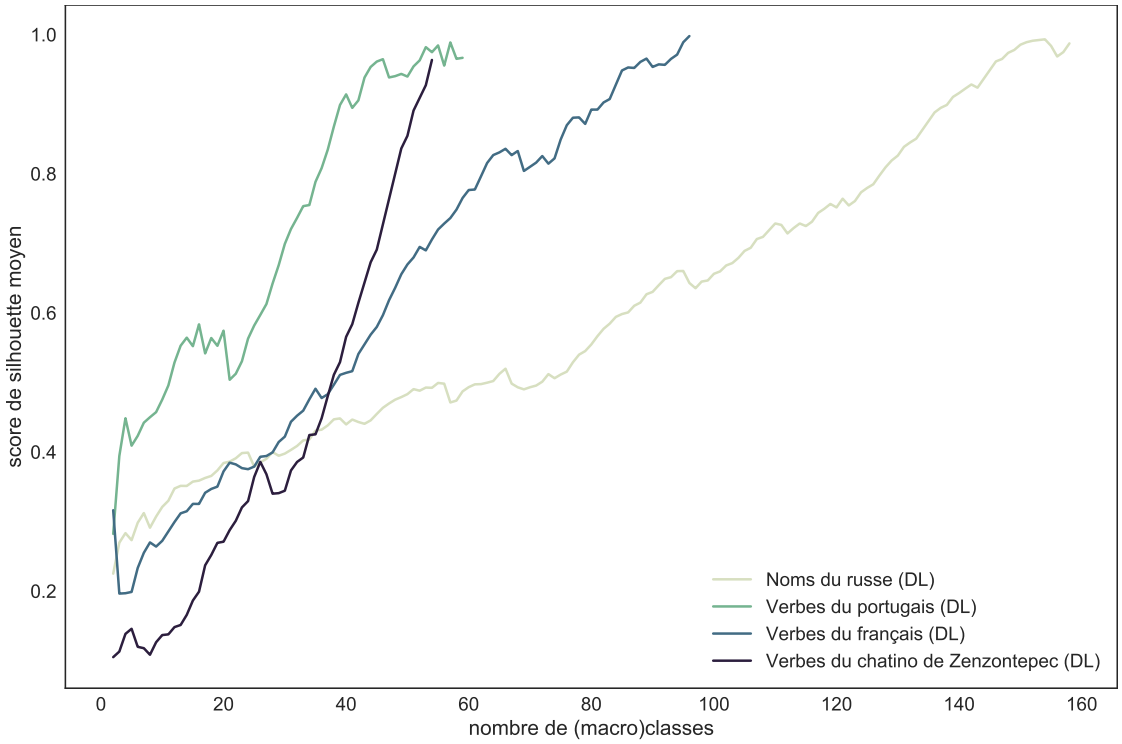


FIGURE 5.11 – Scores de silhouette pour toutes les partitions contenues dans les dendrogrammes du chapitre 4.

Cette première évaluation nous apprend donc que l'algorithme présenté dans ce chapitre permet d'inférer des classes généralement plus cohésives et distinctes que les classifications connues. Cependant, il apparaît clairement au vu de la figure 5.11 qu'il n'existe pas de systèmes

de macroclasses de petite cardinalité idéales de ce point de vue. Même dans le cas du portugais, pour lequel il existe des optimums locaux autour de 5 et 15 classes, la qualité des silhouettes du système augmente avec le nombre de macroclasses postulées.

Notons ici que le critère de minimisation des distances de l'algorithme UPGMA, le critère de longueur de description, et les scores de silhouette répondent chacun à une logique différente : le premier évalue la qualité de chaque classe par sa cohérence interne, le second évalue à quel point chaque microclasse est conforme à sa macroclasse, enfin, le dernier évalue l'économie descriptive générale du système.

5.5.3 Qualité de la recherche

En comparant la longueur de description des macroclasses inférées automatiquement avec celles des classifications connues, nous pouvons évaluer la qualité de l'algorithme de recherche ascendante gloutonne : en effet, s'il se perdait dans un mauvais optimum local, il se pourrait que les classifications obtenues présentent une longueur de description moins brève que les classifications connues. Le tableau 5.10 présente la longueur de description de chaque système de macroclasses (arrondie à l'unité).

C	$\mathrm{DL}(C)$
Noms du russe (Corbett 1982)	80767
Noms du russe (DL)	74997
Verbes du Zenzontepec (Campbell 2011, sous-classes)	6061
Verbes du Zenzontepec (Campbell 2011, superclasses)	5646
Verbes du chatino de Zenzontepec (DL)	5534
Verbes du français (traditionnel)	922722
Verbes du français (DL)	877323
Verbes du portugais (traditionnel)	978657
Verbes du portugais (DL)	930383

Tableau 5.10 – Longueurs de description des systèmes de macroclasses.

On observe que pour chaque système, les macroclasses inférées automatiquement présentent bien une plus petite longueur de description. La figure 5.12 compare pour notre algorithme et pour l'algorithme UPGMA l'évolution de la longueur de description en fonction du nombre de classes inférées au cours de la recherche ascendante. Pour les besoins de la comparaison, nous ignorons les entrées défectives pour la classification hiérarchique.

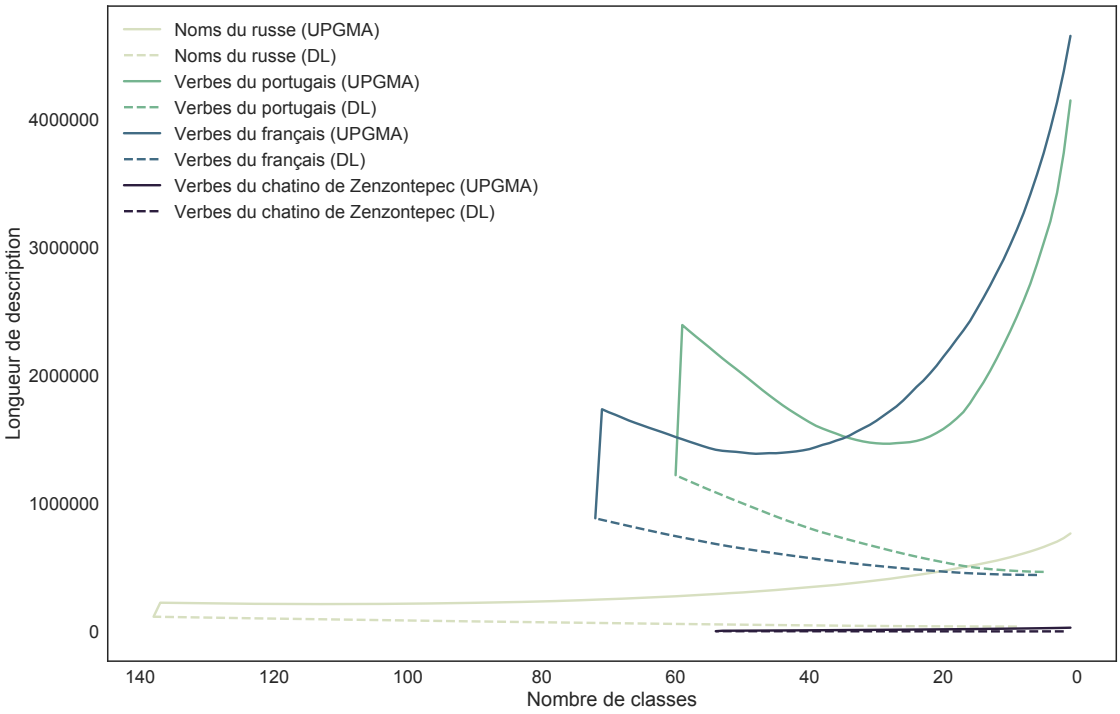


FIGURE 5.12 – Évolution de la longueur de description en fonction du nombre de classes.

Au fil des itérations, tant pour l'algorithme UPGMA que pour l'algorithme présenté à la section 5.2, le nombre de macroclasse décroît. Comme il est attendu, l'algorithme fondé sur la longueur de description (lignes pointillées dans la figure 5.12) conduit à une longueur de description décroissante au fil de la réduction du nombre de classes. En effet, deux classes ne sont fusionnées que si cela autorise une réduction de la longueur de description. En revanche, l'algorithme UPGMA produit des classes qui n'ont pas cette caractéristique, l'optimum en termes de longueurs de description se situant en milieu de courbe [6]., et toujours beaucoup plus haut

6. Contrairement à notre algorithme, l'algorithme UPGMA poursuit les fusions jusqu'à obtenir une unique classe.

que l'optimum atteint par notre algorithme.

Nous en concluons que l'algorithme de recherche présenté à la section 5.2 est raisonnablement efficace, puisqu'il trouve systématiquement une longueur de description plus basse d'une part que les systèmes de macroclasses connus pour chaque langue, et d'autre part que l'ensemble des partitions envisagées par l'algorithme UPGMA.

5.5.4 Similarité entre les classifications

Lorsque nous avons présenté les systèmes de macroclasses inféré automatiquement, nous avons dit qu'ils suivaient de près les classes décrites dans la littérature. Cette proximité peut elle-même être évaluée quantitativement. Nous proposons d'employer une mesure de théorie de l'information afin de mesurer la similarité entre les classes inférées avec la longueur de description et les classifications connues des mêmes systèmes.

Pour ce faire, il nous faut tout d'abord déterminer une distribution de probabilité sur les classes d'une partition. Soit un ensemble de classes C, constitué de k classes $\{C_1, ..., C_k\}$, chaque classe C_i contenant $|C_i|$ lexèmes, nous pouvons définir la probabilité d'une classe par sa fréquence de type relative (Wagner et Wagner 2007) :

$$P_C(i) = \frac{|C_i|}{\sum_{m=1}^{k} |C_m|}$$

Soient deux partitions C et C' comprenant respectivement k et l classes, et classant toutes deux le même ensemble de n lexèmes, la probabilité jointe $P(i, j)$ d'appartenir à la fois à la classe $i \in C$ et à la classe $j \in C'$ se définit :

$$P_{C,C'}(i, j) = \frac{|C_i \cap C'_j|}{\sum_{m=1}^{k} |C_m|}$$

Nous pouvons évaluer à quel point chaque classification nous informe sur l'autre au moyen de l'information mutuelle $I(C, C')$ définie comme suit :

C	$NMI(C^{DL}, C)$
Verbes du portugais	0.93
Verbes du français	0.87
Noms du russe (Corbett 1982)	0.75
chatino de Zenzontepec (Campbell 2011, sous-classes)	0.54
chatino de Zenzontepec (Campbell 2011, superclasses)	0.32

Tableau 5.11 – Information mutuelle normalisée entre systèmes de macroclasses.

$$I(C, C') = \sum_{i=1}^{k} \sum_{j=1}^{l} P_{C,C'}(i,j) \log 2 \frac{P_{C,C'}(i,j)}{P_C(i) P_{C'}(j)}$$

Suivant Fred et Jain (2003) et Wagner et Wagner (2007), et comme précédemment (chapitre 3), nous normalisons cette mesure de façon à ce qu'elle s'échelonne entre 0 et 1.

$$NMI(C, C') = \frac{2I(C, C')}{H(C) + H(C')}$$

Le tableau 5.11 présente cette mesure pour chaque système. Les partitions présentent un accord fort en portugais, français et en russe, mais non en chatino de Zenzontepec.

5.5.5 Informativité des macroclasses

Les systèmes de classes flexionnelles traditionnels peuvent être vus comme des simplifications des systèmes de microclasses. La classification moins fine mène nécessairement à la perte d'information, mais le plus petit nombre de classes peut la rendre plus facile à mémoriser, et fournir une vue d'ensemble des systèmes. Ces propriétés sont également utiles dans une perspective didactique. Les microclasses ont au contraire l'avantage de classer exactement le comportement de chaque lexème (connaître la microclasse d'un lexème, c'est savoir exactement comment le fléchir), mais leur nombre les rend difficiles à interpréter intuitivement. En conséquence, un bon système de macroclasses devrait être aussi informatif que possible sur les microclasses.

Afin de mesurer combien les différents systèmes de macroclasses envisagés (inférés automatiquement ou non) sont porteurs d'information sur les microclasses, nous employons à nouveau l'information mutuelle. Nous calculons cette fois l'information mutuelle entre chaque partition en macroclasses C et les partitions de microclasses M correspondantes. Celle-ci nous dit combien d'information les systèmes de macroclasses apportent sur les systèmes de microclasses dont ils sont une approximation (et réciproquement). Elle permet donc d'évaluer la qualité de cette approximation.

C	NMI$(C;M)$
Noms du russe (Corbett 1982)	0.502
Noms du russe (DL)	0.687
Verbes du Zenzontepec (Campbell 2011, sous-classes)	0.816
Verbes du Zenzontepec (Campbell 2011, superclasses)	0.517
Verbes du chatino de Zenzontepec (DL)	0.250
Verbes du français (traditionnel)	0.481
Verbes du français (DL)	0.531
Verbes du portugais (traditionnel)	0.491
Verbes du portugais (DL)	0.531

Tableau 5.12 – Information mutuelle normalisée entre systèmes de macroclasses et systèmes de microclasses.

Le tableau 5.12 indique l'information mutuelle normalisée entre chaque système de macroclasses et le système de microclasse correspondant. En russe, en français et en portugais, la classification inférée automatiquement est plus informative sur les microclasses que les autres classifications évaluées. En Zenzontepec, cependant, ce sont les classes les plus précises de Campbell (2011) qui sont les plus informatives.

5.6 Conclusion

Dans ce chapitre, nous avons présenté un algorithme de classification des microclasses en macroclasses fondé sur le principe du rasoir d'Occam (Beniamine, Bonami et Sagot 2017). Celui-ci procède de bas en haut. Il est initialisé avec une partition en microclasses et fusionne deux classes à chaque étape. Le choix des fusions minimise la longueur de description du système de classes flexionnelles. Celle-ci est définie de façon formelle en termes probabilistes. L'algorithme cesse de fusionner les classes lorsque la longueur de description du système ne décroît plus. Les classes restantes sont alors les macroclasses. Si la longueur de description pouvait toujours être améliorée par une fusion, alors l'algorithme pourrait choisir de représenter le système par une unique macroclasse. De même, s'il n'y avait aucun bénéfice à fusionner deux microclasses, il pourrait arrêter son choix sur le système de microclasses entier. L'algorithme est donc à même d'évaluer si l'une de ces solutions triviales est préférable à un système de macroclasses.

Empiriquement, sur les systèmes des noms du russe, des verbes du français, du portugais, et du chatino du Zenzontepec, le système produit chaque fois une classification de macroclasses non triviale. Ces classes captent bien, dans une certaine mesure, des relations de distances entre les microclasses. En portugais, en français et en russe, ces macroclasses sont meilleures que les classifications traditionnelles du point de vue des longueurs de description, du respect des relations de distances (et en particulier, de l'équilibre entre cohésion interne et distinctivité), et du point de vue de l'information qu'elles apportent sur les microclasses. Ces macroclasses sont également très similaires aux classes traditionnelles. En chatino du Zenzontepec, notre classification, qui ne comporte que deux classes, est préférable du point de vue des longueurs de description, mais les superclasses de Campbell (2011) présentent un meilleur score de silhouette moyen, et ses sous-classes une meilleure informativité sur les microclasses. Cependant, les scores de silhouette sont particulièrement bas pour les trois partitions envisagées, et la longueur de description décroît très peu par rapport au système de microclasses. On peut se demander si les macroclasses sont véritablement intéressantes pour cette langue. Cela est peut-être dû à la petite taille des paradigmes. En portugais, les graphes de silhouette montrent que les trois classes les plus grandes sont très cohérentes et distinctives. Dans les autres langues, aucun des

systèmes, ni inférés, ni connus, ne sont très bons de ce point de vue. Nous pensons que les macroclasses ne sont pas, en général, un très bon modèle de ces systèmes : d'une part, elles sont moins précises que des microclasses, et d'autre part, elles n'expriment que de façon très grossières leurs relations de similarité. Les résultats de nos évaluations constituent un argument fort en faveur de la position de Brown et Hippisley (2012) :

> [N]ous soutenons ici que la notion traditionnelle de classes flexionnelles comme entités monolithiques est trompeuse. Elles devraient plutôt être perçues comme des entités partielles qui peuvent être modélisées en termes d'une hiérarchie spécifique à la morphologie flexionnelle [7].

Le chapitre 6 propose de développer un modèle plus fidèle des classes flexionnelles sous la forme d'une hiérarchie à héritage multiples.

7. [En anglais dans le texte] « *[W]e argue here that the traditional notion of inflectional classes as monolithic entities is misleading. Instead, they should be perceived as partial entities which can be modelled in terms of a specific hierarchy for inflectional morphology* ».

Chapitre 6

Classification des lexèmes en hiérarchies à héritages multiples

En nous fondant sur les patrons d'alternance, nous avons défini deux types de partitions de classes flexionnelles : les microclasses regroupent les lexèmes ayant exactement le même comportement flexionnel ; les macroclasses regroupent des lexèmes au comportement similaire, et correspondent aux classes généralement distinguées par les grammaires descriptives. La modélisation d'un système flexionnel par des microclasses ne rend compte que d'un petit sous ensemble des nombreux points de similarité entre ces microclasses. Sa modélisation en macroclasses met en avant certaines de ces similarités, mais en ignore nécessairement d'autres. Ce chapitre propose que tout point de similarité entre microclasses peut être considéré comme une classe flexionnelle en soi. En conséquence, nous représentons les classes flexionnelles sous la forme d'une hiérarchie où l'héritage peut être multiple (une classe peut avoir plusieurs superclasses non ordonnées entre elles), et qui rend compte très exactement de tous les points communs entre microclasses.

Dans une première section (6.1), nous discutons les propriétés des classes flexionnelles qui nécessitent une modélisation par une hiérarchie à héritage multiple. Nous présentons dans la section 6.2 le formalisme mathématique nommé ANALYSE FORMELLE DE CONCEPTS, que nous utilisons pour produire la hiérarchie à partir des vecteurs de patrons. Dans la section 6.3, nous employons ce formalisme pour analyser quantitativement la canonicité des treillis de classes flexionnelles. Les hiérarchies obtenues directement à partir des patrons sont difficiles à inter-

préter, en raison de la grande redondance des patrons. La section 6.4 décrit une méthode pour obtenir des hiérarchies lisibles fondées sur une sélection de classes et de patrons d'alternances, et présente quelques exemples.

6.1 Canonicité et structure des classes

Revenons encore aux critères de canonicité proposés par Corbett (2009) pour les classes flexionnelles. Le principe I concerne la distinctivité des classes d'un système flexionnel, qui présentent :

(critère 1) à travers les classes, des paradigmes concrets distincts ;

(critère 2) mais des paradigmes abstraits identiques ;

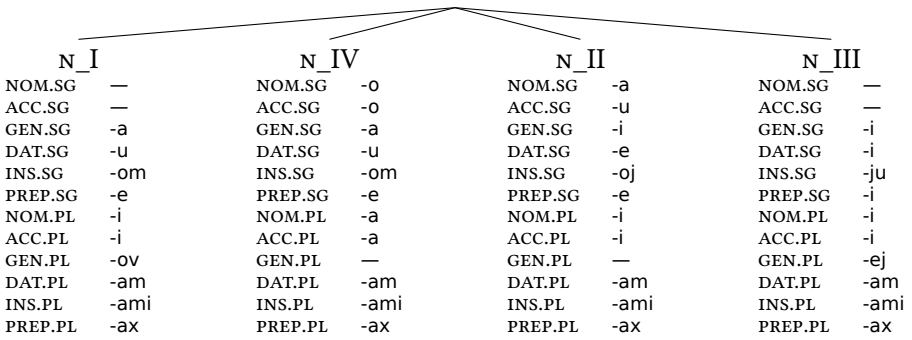
(critère 3) à l'intérieur d'une classe, des paradigmes concrets identiques ;

(critère 4) une structure implicative plate, où aucune case n'est plus prédictive qu'une autre.

Il suit de ces critères qu'il n'existe, dans un système canonique, aucun point commun entre deux microclasses. En effet, si deux microclasses A et B partagent un exposant ou un patron d'alternance, elles violent le critère 1. De plus, la case concernée par ce partage est alors moins prédictive que les autres (A et B ne peuvent être distingués à partir de celle-ci), violant le critère 4. Par ailleurs, en vertu du critère 2, un système de classes canonique ne peut présenter aucune entrée défective ou surabondante. Enfin, le critère 3 est vrai par définition de tout système de microclasses. Mais dans un système où chaque microclasse est entièrement distincte des autres, il n'existe pas de système de macroclasses non trivial et distinct des microclasses. En somme, un système flexionnel canonique prend la forme d'une partition de classes flexionnelles à un seul niveau, conformément à la description traditionnelle des classes flexionnelles. Décrire un système non canonique au moyen d'une partition en classes revient donc à masquer sa non canonicité dans la classification.

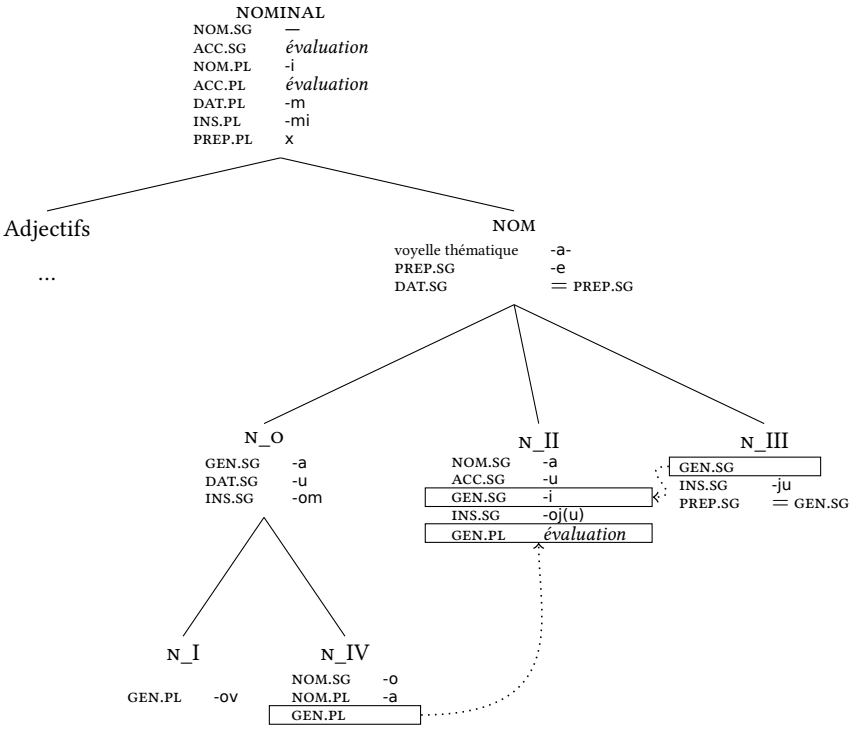
Reprenons l'exemple des classes flexionnelles du russe, décrites par Brown (1998) et Brown et Hippisley (2012), déjà discutées dans la section 4.2 du chapitre 4. L'exemple (58) présente ces classes, définies à partir d'affixes, sous la forme d'une partition.

(58)



Nous rappelons également, exemple (59), la structure proposée par Brown (1998). Nous avons vu que dans cette analyse, il existe deux façons différentes d'hériter d'un trait (par une règle de renvoi, ou par la hiérarchie des classes).

(59)

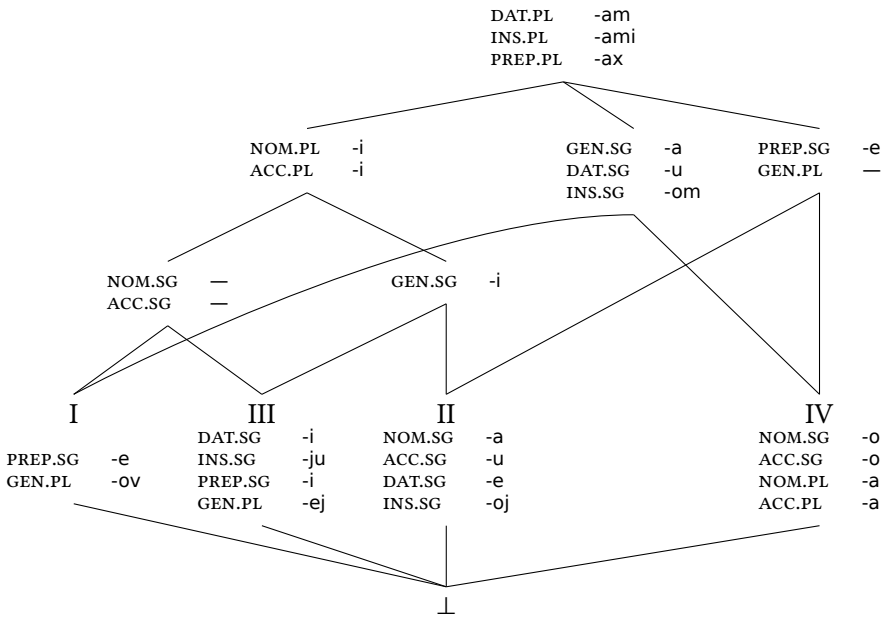


En (60), nous proposons une analyse qui rend compte de l'ensemble des points de similarité (entre les noms exclusivement) entre ces quatre classes. Cette analyse ne se permet pas d'autres moyens d'assigner une propriété à une classe que la hiérarchie elle-même : en conséquence,

elle ne comporte ni règles de renvoi entre cases, ni fonctions à évaluer [1]. L'information y est organisée entièrement et uniquement via la hiérarchisation des classes.

Ce type de hiérarchie monotone à héritage multiple forme un treillis. Il s'agit du même type de structure que la hiérarchie des types en HPSG (Flickinger 1987 ; Pollard et Sag 1994 ; Ginzburg et Sag 2000), ou que les hiérarchies de traits phonologiques (Chomsky et Halle 1968 ; Frisch 1997).

(60)



Dans la hiérarchie proposée en (60), chaque nœud intermédiaire existe en vertu d'un point de similarité entre classes, et toutes les similarités entre classes sont ainsi exprimées. On peut lire l'ensemble des informations pertinentes pour une classe en lisant l'ensemble de ses ancêtres. Les informations spécifiées sur les feuilles sont entièrement distinctives (elles sont propres à chaque classe seule). À l'opposé, les valeurs indiquées à la racine sont communes à toutes les classes. Dans la hiérarchie, les nœuds les plus hauts sont plus généraux que les nœuds les plus bas : ils sont moins spécifiés en termes d'information, mais concernent plus de classes. Les

1. Pour cet exemple, nous considérons les classes I à IV comme des microclasses, ce qui exclue de nombreux lexèmes dont la hiérarchie de Brown (1998) rend compte via les fonctions EVALUATION. Cela n'est absolument pas un problème de principe : le type de classification proposé vise à être étendu à toutes les microclasses d'un système. Rien n'empêcherait en principe non plus d'intégrer également les adjectifs à ce type d'analyse.

nœuds y sont donc ordonnés par généralité croissante.

Étant donnés deux nœuds de la hiérarchie qui définissent certaines propriétés, on peut savoir quelles classes ces deux ensembles de propriétés ont en commun en cherchant leur borne inférieure, c'est à dire leur plus grand enfant commun. Par exemple, le nœud {NOM.PL -i, ACC.PL -i} et le nœud {PREP.SG -e, GEN.PL -} ont pour borne inférieure la classe II. Mais la classe II et la classe IV ont pour borne inférieure le symbole $\bot$, qui symbolise l'absence de classes : ils n'ont aucune sous-classe en commun. Symétriquement, on peut se demander de n'importe quelle paire de nœuds, et donc des classes qu'ils recouvrent, quelles informations elles ont en commun, en cherchant leur borne supérieure, c'est à dire leur plus petit ancêtre commun.

La hiérarchie de l'exemple (60) exhibe précisément ce qui distingue ces classes de la situation canonique : tandis que la situation canonique ne présente aucun nœud intermédiaire, ces quelques classes du russe en présentent cinq. Tandis que dans la situation canonique, la hiérarchie n'a qu'une hauteur de un, cette hiérarchie a une hauteur maximale de trois (le chemin de la racine à la classe I emprunte trois arcs). Enfin, tandis que la situation canonique ne présente que de l'héritage simple, c'est à dire que mis à part la racine, on trouve un seul parent par nœud, dans cette hiérarchie, les nœuds inférieurs à la racine ont en moyenne 1.4 parents.

Les exemples (58) à (60) illustrent successivement trois modèles de classification d'expressivité croissante : une partition, un arbre, un treillis. Une organisation en partition fait la prédiction que les similarités entre microclasses sont inexistantes (c'est-à-dire qu'elles sont canoniques), ou tellement exceptionnelles qu'elles sont négligeables. Nous savons que cela est faux de tous les systèmes que nous avons observés.

Une organisation en arbre fait la prédiction qu'il existe du partage de propriétés entre microclasses, mais qu'il est rare ou exceptionnel qu'une classe partage des traits avec (au moins) deux classes qui n'ont pas ces traits en commun. Corbett (2009) définit comme suit la notion d'HÉTÉROCLISE :

> un petit nombre d'items [lexèmes] présentant des combinaisons de formes spécifiques à d'aures classes peuvent être traitées comme hétéroclites [2].

2. [En anglais dans le texte] « *a small number of items showing combinations of forms from other classes can be treated as heteroclites* ».

Dans nos termes, une microclasse hétéroclite serait une microclasse de petite taille héritant de plusieurs classes plus grandes qui ne sont elles-mêmes pas en relation d'ascendance hiérarchique. Il est difficile cependant d'estimer quantitativement quelles classes doivent être considérées comme petites ou grandes. Par ailleurs, comme on peut le voir dans l'exemple (60), l'héritage multiple peut apparaître entre n'importe quelles classes, quelles que soient leurs tailles relative.

Les phénomènes de partage de propriétés et d'héritages multiples ne sont pas exclusivement le propre des systèmes flexionnels dont l'organisation en classes est réputée complexe. Observons par exemple un petit sous ensemble des microclasses de l'anglais dites irrégulières, listées de (61) à (64). Le lexème en gras est celui que nous utilisons comme étiquette de microclasse.

(61) BEGET, **FORGET**

(62) BACKBITE, **BITE**

(63) BESTRIDE, DRIVE, OUTRIDE, OVERRIDE, **RIDE**, STRIDE, STRIVE, TEST-DRIVE

(64) FORGIVE, **GIVE**, MISGIVE

Les quatre microclasses envisagées sont petites, et la plus grande, (62) comporte plusieurs composés. Nous présentons dans la figure 6.1 la structure de similarité entre ces classes sous la forme de patrons d'alternance. Par souci de brièveté, nous ne présentons pas les contextes d'application. Puisque cet ensemble ne comporte pas le verbe TO BE, nous ne mentionnons dans cet exemple que les cases de paradigme PRS.OTHERS, PRS.3S, PST, PPRES et PPART.

Cette hiérarchie met en évidence le fait que, y compris au sein de petites classes, et dans un système flexionnel réputé peu complexe, il existe des partages qui donnent lieu à de l'héritage multiple. Nous appellerons ce type de partage PARTAGE HÉTÉROCLITE.

Nous nous proposons d'inférer automatiquement ces hiérarchies à héritage multiple à partir des patrons d'alternance, puis d'en étudier la canonicité.

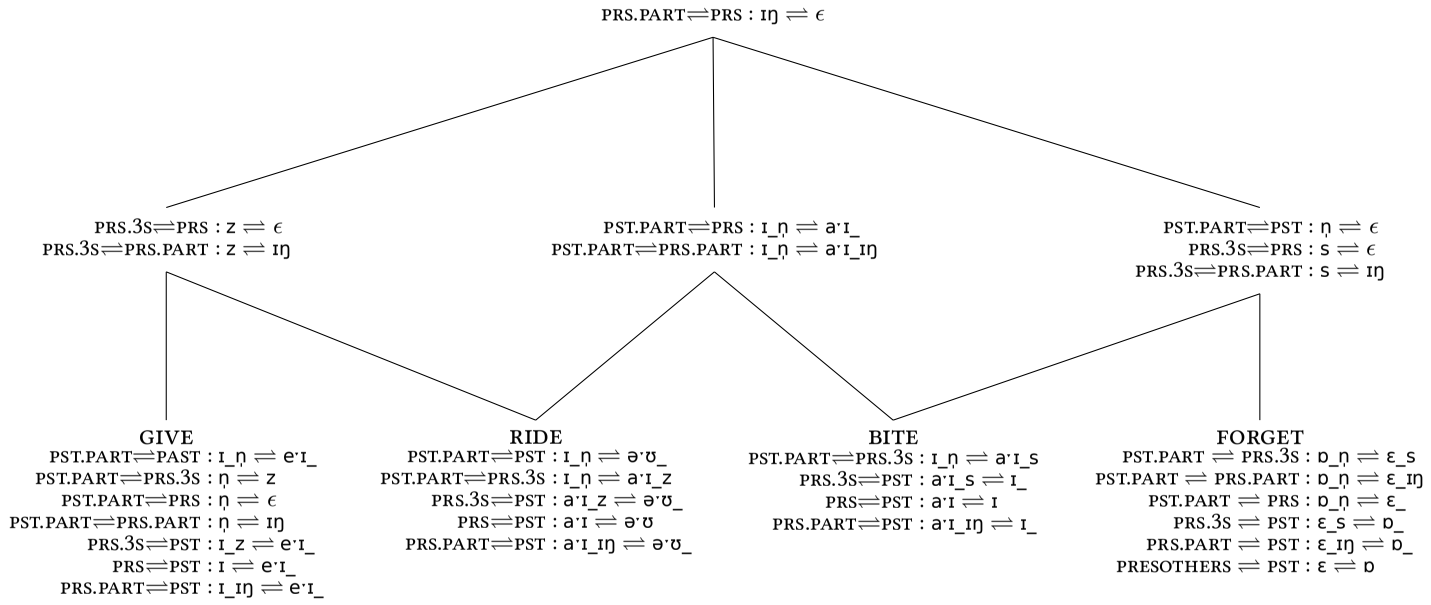


Figure 6.1 – Hiérarchie flexionnelle de quelques verbes hétéroclites de l'anglais.

6.2 Analyse formelle de concepts

Afin de produire la classification à partir des vecteurs de patrons assignés à chaque lexème, nous employons le formalisme mathématique nommé ANALYSE FORMELLE DE CONCEPTS (Ganter et Wille 1998). Ce formalisme permet d'étudier l'ensemble des relations pertinentes entre des objets (dans notre cas, des lexèmes ou des microclasses), et des propriétés qu'ils instancient (dans notre cas, des propriétés flexionnelles), sous la forme de classes nommées CONCEPTS et ordonnées entre elles en une hiérarchie conceptuelle.

Dans cette section, nous prenons pour exemple le même petit sous-ensemble des lexèmes de l'anglais, déjà discuté et présenté dans le tableau 6.1. Nous suivons principalement les définitions de Bělohlávek (2009), mais également celles de Ganter et Wille (1998).

lexème	PST	PAST.PART	PRS
DRIVE	/drə·ʊv/	/drɪvn̩/	/draˑɪv/
RIDE	/rəʊd/	/rɪdn̩/	/raˑɪd/
BITE	/bɪt/	/bɪtn̩/	/baˑɪt/
FORGET	/fəgɒt/	/fəgɒtn̩/	/fəgɛt/

TABLEAU 6.1 – Quelques sous-paradigmes des verbes de l'anglais.

Jusqu'ici, nous avons considéré que le comportement flexionnel d'un lexème peut être représenté par un vecteur de patrons d'alternances. La relation entre les lexèmes et les patrons peut également être représentée par une table d'incidence, comme dans le tableau 6.2, où les objets (lexèmes) sont indiqués en ligne, les propriétés (patrons) en colonne, et où une croix dans une case indique que l'objet de cette ligne instancie la propriété de cette colonne. Une telle table d'incidence représente un CONTEXTE FORMEL.

Un CONTEXTE FORMEL est un triplet $\langle X, Y, I \rangle$, où X et Y sont des ensembles non vides, et I est une relation binaire d'incidence entre X (les objets, en ligne) et Y (les propriétés, en colonne) : $I \subseteq X \times Y$. Pour tout $x \in X$ et $y \in Y$:

— $\langle x, y \rangle \in I$ indique que l'objet x a la propriété y,

	PST.PART⇌PRS		PST ⇌ PRS			PST.PART ⇌ PST	
	ɪ_ɳ ⇂↾ aɪ_	ɔ_ɳ ⇂↾ ɛ_	aɪ ⇂↾ eɪ	ɪ ⇂↾ aɪ	ɛ ⇂↾ ɔ	ɪ_ɳ ⇂↾ eɪ_	ɳ ⇂↾ ɛ
DRIVE	×		×			×	
RIDE	×		×			×	
BITE	×			×			×
FORGET		×			×		×

TABLEAU 6.2 – Contexte formel pour notre petit sous-paradigme de l'anglais, sous la forme d'une table d'incidence.

– $\langle x, y\rangle \notin I$ indique que x n'a pas y.

Dans la table d'incidence représentant un contexte $\langle X, Y, I\rangle$, on trouve une croix dans la case indiciée par i et j si et seulement si $\langle x_i, y_i\rangle \in I$. Ganter et Wille (1998) proposent de noter $\langle x, y\rangle \in I$ par xIy.

Pour tout sous-ensemble d'objets $A \subset X$, nous nous intéressons aux propriétés qu'ils ont en commun, et pour tout sous-ensemble de propriétés $B \subset Y$, nous nous intéressons aux objets qui les partagent tous. Nous définissons deux opérateurs « $\uparrow$ » et « $\downarrow$ », tels que [3] :

– $\uparrow: 2^X \to 2^Y$

– $A\uparrow = \{y \in Y |$ pour chaque $x \in A : xIy\}$

– $\downarrow: 2^Y \to 2^X$

– $B\downarrow = \{x \in X |$ pour chaque $y \in B : \langle xIy\}$

Si les objets A n'ont aucune propriété en commun, alors $A\uparrow = \emptyset$. De même, si aucun objet ne partage toutes les propriétés de B, alors $B\downarrow = \emptyset$. En conséquence, $\emptyset\uparrow = Y$ et $\emptyset\downarrow = X$.

Dans notre exemple, on a :

(65) {RIDE, DRIVE}↑ = {ɪ_ɳ ⇌ aˑɪ_, aˑɪ ⇌ əˑʊ, ɪ_ɳ ⇌ əˑʊ_}

3. Cette notation est celle de Bělohlávek (2009), tandis que Ganter et Wille (1998) notent les deux opérateurs $'$, et les ensembles $A\uparrow$ et $B\downarrow$ respectivement A' et B'.

(66) {ɪ_ɳ ⇌ aˑɪ_, aˑɪ ⇌ əˑʊ, ɪ_ɳ ⇌ əˑʊ_}↓ = {DRIVE, RIDE}

(67) {ɪ_ɳ ⇌ aˑɪ_}↓ = {DRIVE, RIDE}

(68) {aˑɪ ⇌ ɪ, ɛ ⇌ ɒ}↓ = ∅

Ces égalités peuvent se lire dans le tableau 6.2. Les lexèmes DRIVE et RIDE partagent toutes leur propriétés (65). Les trois patrons qu'ils partagent ne sont partagés que par eux (66). Le patron ɪ_ɳ ⇌ aˑɪ_ n'est lui aussi partagé que par ces deux lexèmes (67). Enfin, l'opérateur ↓ appliqué à des patrons concurrents pour la même alternance produit généralement un ensemble vide (68), sauf s'il existe des lexèmes surabondants présentant alternativement les deux patrons.

Ces opérateurs nous permettent de définir un CONCEPT FORMEL. Un concept formel dans $\langle X, Y, I\rangle$ est une paire $\langle A, B\rangle$ formée d'un ensemble d'objets $A \subseteq X$ appelé l'EXTENSION du concept, et un ensemble de propriétés $B \subseteq Y$ appelé l'INTENSION du concept, tels que $A \uparrow = B$ et $B \downarrow = A$. En somme, A est l'ensemble de tous les objets qui ont toutes les propriétés de B, et B est l'ensemble de toutes les propriétés communes à tous les objets de A.

Par exemple, ⟨{DRIVE,RIDE}, {ɪ_ɳ ⇌ aˑɪ_, aˑɪ ⇌ əˑʊ, ɪ_ɳ ⇌ əˑʊ_ }⟩ est un concept formel, en raison de (65) et (66), mais ⟨{DRIVE,RIDE}, {ɪ_ɳ ⇌ aˑɪ_ }⟩ n'en est pas un, car malgré (67), l'inverse n'est pas vrai (65).

Il est possible, à partir de la table d'incidence, de produire la liste de tous les concepts correspondants. Les exemples (69) à (75) fournissent la liste des concepts présents dans le tableau 6.2 :

(69) ⟨∅, {ɒ_ɳ ⇌ ɛ_ , ɪ_ɳ ⇌ əˑʊ_ , ɪ_ɳ ⇌ aˑɪ_ , aˑɪ ⇌ ɪ , ɳ ⇌ ϵ }⟩

(70) ⟨{BITE}, {ɪ_ɳ ⇌ aˑɪ_ , aˑɪ ⇌ ɪ , ɳ ⇌ ϵ }⟩

(71) ⟨{FORGET}, {ɒ_ɳ ⇌ ɛ_ , ɳ ⇌ ϵ, ɛ ⇌ ɒ }⟩

(72) ⟨{BITE, FORGET}, {ɳ ⇌ ϵ }⟩

(73) ⟨{DRIVE, RIDE}, {ɪ_ɳ ⇌ əˑʊ_ , ɪ_ɳ ⇌ aˑɪ_, aˑɪ ⇌ əˑʊ }⟩

(74) ⟨{BITE, DRIVE, RIDE}, {ɪ_ɳ ⇌ aˑɪ_ }⟩

(75) ⟨{BITE, DRIVE, FORGET, RIDE}, ∅⟩

Nous avions remarqué au sujet de la hiérarchie des classes nominales du russe (60) que les classes pouvaient être comparés en termes de spécificité. Les concepts peuvent de même être

ordonnés entre eux par leur spécificité. Étant donnés deux concepts $\langle A_1, B_1\rangle$ et $\langle A_2, B_2\rangle$ dans $\langle X, Y, I\rangle$, $\langle A_1, B_1\rangle$ est plus spécifique que $\langle A_2, B_2\rangle$ si et seulement si A_1 est un sous-ensemble de A_2, ce qui implique également que B_1 est un super-ensemble de B_2. On a donc : Nous avions remarqué au sujet de la hiérarchie des classes nominales du russe (60) que les classes pouvaient être comparés en termes de spécificité. Les concepts peuvent de même être ordonnés entre eux par leur spécificité. Étant donnés deux concepts $\langle A_1, B_1\rangle$ et $\langle A_2, B_2\rangle$ dans $\langle X, Y, I\rangle$, $\langle A_1, B_1\rangle$ est plus spécifique que $\langle A_2, B_2\rangle$ si et seulement si A_1 est un sous-ensemble de A_2, ce qui implique également que B_1 est un super-ensemble de B_2. On a donc :

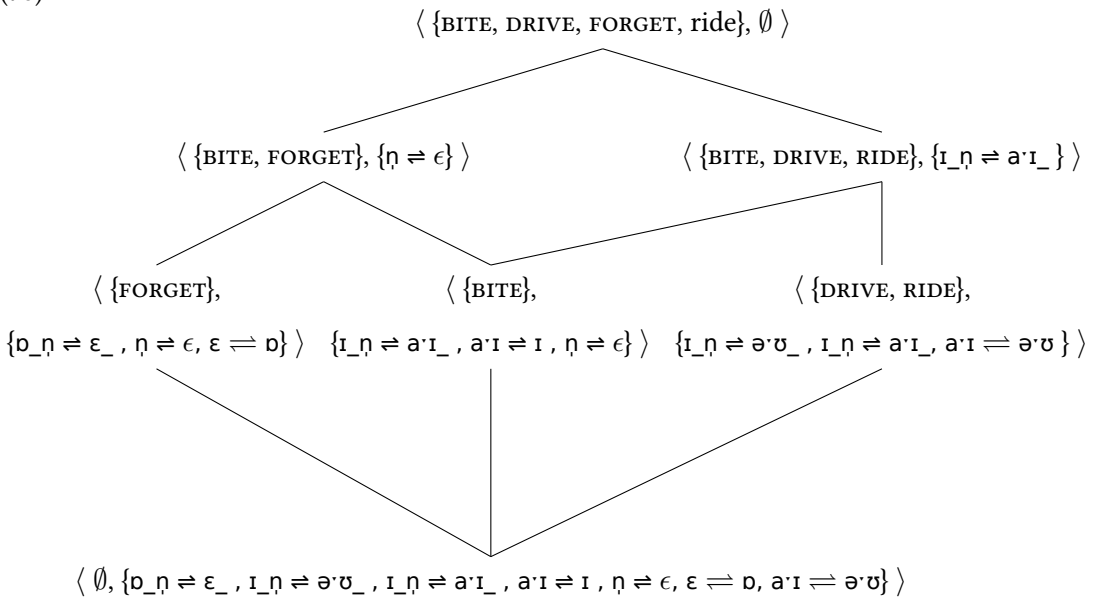
$$\begin{aligned}\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle &\iff A_1 \subseteq A_2\\ &\iff B_2 \subseteq B_1\end{aligned}$$

On peut dire que $\langle A_1, B_1\rangle$ est un sous-concept de $\langle A_1, B_1\rangle$. Par ailleurs, si $\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$ et qu'il n'existe aucun concept $\langle A_i, B_i\rangle$ dans $\langle X, Y, I\rangle$ tel que $\langle A_1, B_1\rangle \leq \langle A_i, B_i\rangle \leq \langle A_2, B_2\rangle$, alors $\langle A_1, B_1\rangle$ est un voisin inférieur immédiat de $\langle A_2, B_2\rangle$, et on note : $\langle A_1, B_1\rangle \prec \langle A_2, B_2\rangle$.

La collection des concepts formels d'un contexte $\langle X, Y, I\rangle$, associée à la relation d'ordre $\leq$, forme le TREILLIS DES CONCEPTS de $\langle X, Y, I\rangle$, noté $\mathcal{B}\langle X, Y, I\rangle$. Un ensemble ordonné par une relation d'ordre peut êre représenté par un diagramme de Hasse dans lequel chaque élément de l'ensemble forme un nœud d'une structure hiérarchique, si un élément est inférieur à un autre, il est représenté plus bas dans le diagramme. Les voisins immédiats sont reliés par un arc. De cette façon, pour toute paire de concepts c_1, c_2 dans $\langle X, Y, I\rangle$, on a $c_1 \leq c_2$ si on peut atteindre c_2 à partir de c_1 par un chemin ascendant.

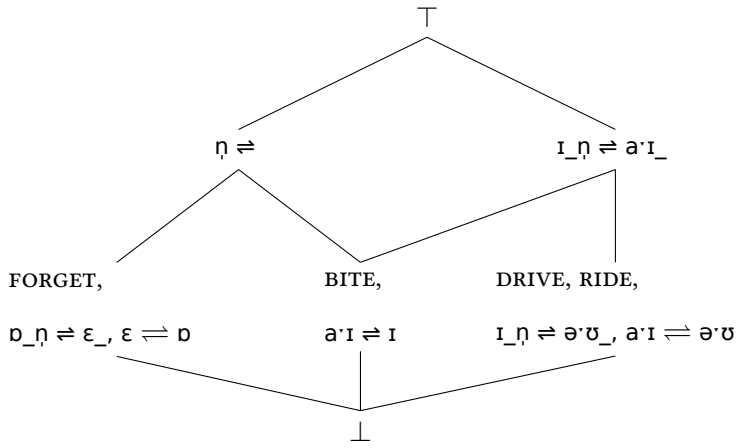
L'exemple (76) présente le diagramme de Hasse pour le treillis du contexte du tableau 6.2. Chaque nœud est annoté par son concept.

(76)



Cette notation est redondante : il n'est en fait pas nécessaire de répéter plus haut les objets qui ont déjà été rencontrés plus bas, car ils peuvent se déduire de la structure hiérarchique. Inversement, il n'est pas nécessaire de répéter plus bas les patrons d'alternance qui ont déjà été rencontrés plus haut. L'affichage dit réduit consiste à ne noter les objets et les propriétés sur la structure hiérarchique qu'au niveau des concepts qui les définissent. L'exemple (77) présente le même treillis que l'exemple (76), en affichage réduit.

(77)



Les treillis de concepts en affichage réduit peuvent se lire comme des hiérarchies monotones à héritage multiple. La hiérarchie en (6.1), produite sur un ensemble d'objets et de propriétés

légèrement plus grands, est également un treillis de concepts. Le concept inférieur, ou *infimum*, y a été omis, car il n'est pas intéressant pour nous : Il comporte toujours l'ensemble de toutes les propriétés pour intension, et l'ensemble vide pour extension.

6.3 Treillis de patrons d'alternance

Afin d'appliquer cette analyse aux systèmes flexionnels que nous étudions, nous prenons pour propriétés tous les couples d'une paire de cases et d'un patron d'alternance, et pour objets, toutes les microclasses. Un lexème défectif sera noté comme n'ayant aucun patron pour toutes les paires concernant une case où il est défectif. Un lexème surabondant sera noté comme ayant tous les patrons qu'il est susceptible d'instancier. Nous nous appuyons sur la librairie python *concepts* (Bank 2016) afin de générer les concepts et le treillis.

Les treillis obtenus sont de grande taille. Par exemple, la figure 6.2 dessine la structure générale des treillis de l'anglais et du français. Seuls les objets sont étiquetés, en affichage réduit, car les patrons d'alternance sont beaucoup trop nombreux. Ces exemples sont typiques de la situation pour l'ensemble des langues observées, c'est à dire qu'ils sont trop grands et trop fournis pour être explorés manuellement.

Ce fait à lui seul invalide l'hypothèse selon laquelle les systèmes réels se rapprocheraient du canon de Corbett : les systèmes flexionnels étudiés dans cette thèse sont très loin de s'organiser en partitions. Lorsque nous discutions de la structure des classes du russe, dans la première section du présent chapitre, nous avons proposé de mesurer la canonicité d'une hiérarchie de classe multiple au moyen de 3 mesures, que nous pouvons maintenant opérationnaliser.

— **Nombre de concepts** : dans la situation canonique, si un treillis a une base de b feuilles, il existe au total $b+1$ concepts (en ignorant l'infimum), c'est à dire que les feuilles ne sont reliées que par la racine. Plus le nombre de concepts est haut, plus le système enfreint le critère 1 (distinctivité).

— **Hauteur de la hiérarchie** : dans la situation canonique, le plus long chemin reliant la racine à une feuille n'emprunte qu'un seul arc. En fait, c'est le cas de tous les chemins entre la racine et une feuille. Mesurer la hauteur de la hiérarchie nous donne une infor-

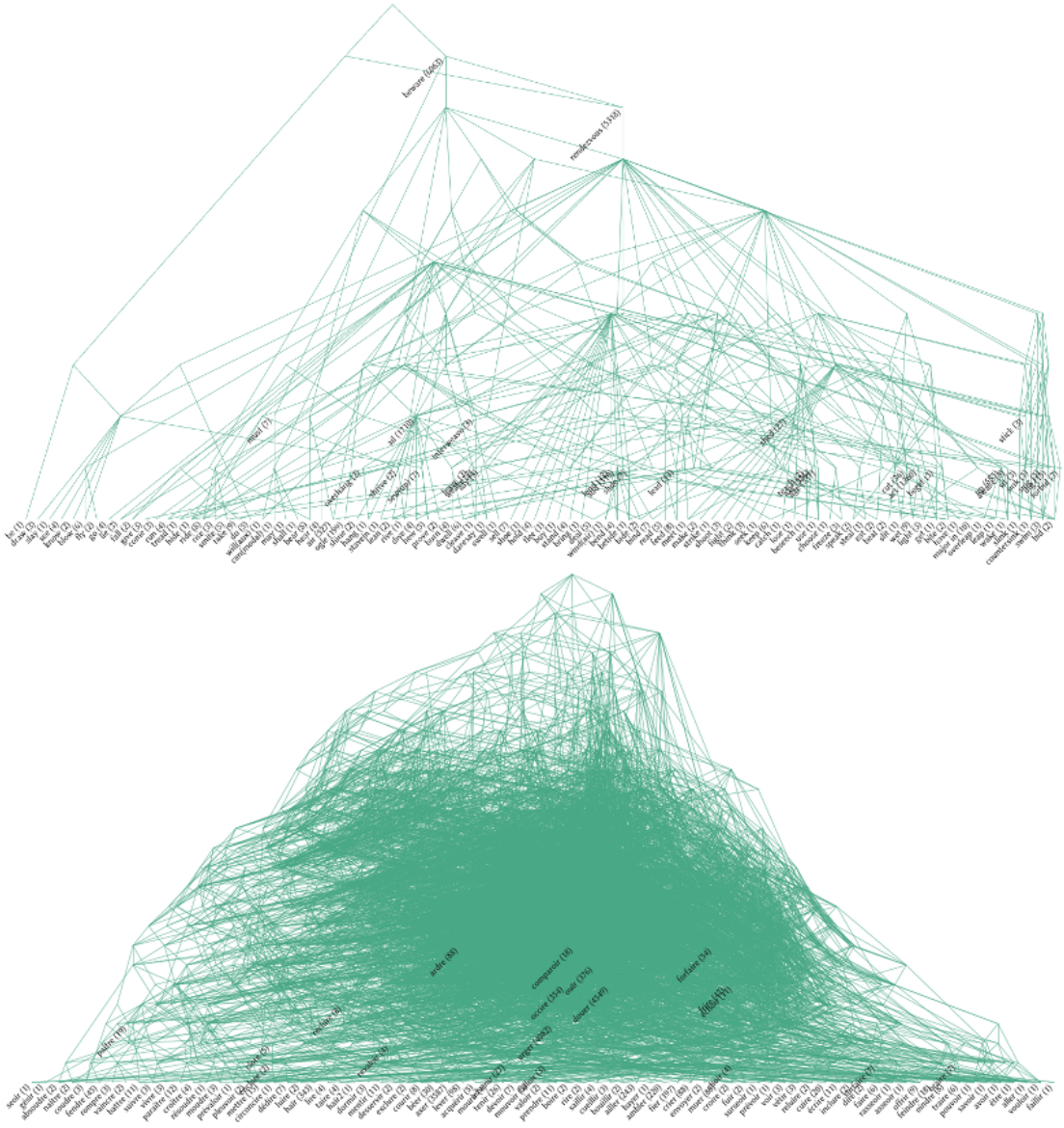


FIGURE 6.2 – Treillis des patrons pour l'anglais et le français, avec affichage réduit des objets uniquement.

mation sur le type de partage entre les microclasses. Une hiérarchie haute est organisée en classes et sous-classes successives, et présente de nombreuses implications entre les concepts (tout concept implique ses parents). Une hiérarchie peu haute présente moins de structure implicative entre les concepts, et pourrait présenter une série de classifications orthogonales, ou moins liées entre elles. En conséquence, une hiérarchie haute viole le critère 4 (structure implicative plate).

— **Degré moyen** : une structure hiérarchique arborescente, c'est à dire où chaque concept ne peut hériter que d'un parent, est plus canonique qu'une hiérarchie à héritage multiple. Nous mesurons le nombre moyen d'arcs entrant sur les concepts (en ignorant la racine), afin d'estimer à quel point ces structures s'éloignent d'un arbre. Un arbre présente un degré moyen de 1. Plus le degré est haut, et plus la structure viole le critère 1 (distinctivité) par des partages hétéroclite.

Le tableau 6.3 présente ces mesures pour chaque système, à l'exception de celui du navajo [4]. Pour chaque système, la taille de la base est assez proche du nombre de microclasses. Celle-ci ne peut en différer qu'en vertu de lexèmes surabondants ou défectifs. La prévalence de ces phénomènes est très dépendant de la constitution des données : par exemple les données CELEX de l'anglais présentent beaucoup de surabondance due à des variantes de l'anglais, tandis que les données du portugais présentent des paradigmes sans défectivité ni surabondance parce que la variation n'a pas été prise en compte par les auteurs de la ressource. Les variations de la taille de la base relativement au nombre de microclasses ne sont donc pas interprétables comme des distinctions typologiques.

Les degrés entrants dans chaque treillis sont toujours supérieurs à 2, sauf pour l'anglais où le degré est de 1.9, ce qui indique que l'héritage multiple y est la norme. Les systèmes de classes des langues étudiées sont donc tous très loin d'être des arbres.

Afin de pouvoir comparer les valeurs de hauteur et le nombre de concepts, nous calculons une hauteur relative et un nombre de concepts relatif (ou densité) dans les treillis.

4. Pour les systèmes bipartites, nous formons les contextes en accolant les contextes des patrons formés sur chaque sous-paradigme. En navajo, le nombre de propriétés est trop élevé pour pouvoir calculer le treillis en un temps raisonnable.

	Microclasses	Base	Hauteur	Degré	Concepts
Anglais	118	88	11	1.91	244
Chatino de Zenzontepec	99	98	8	2.65	524
Portugais	60	60	21	2.79	677
Arabe	367	302	33	3.65	10 125
Français	97	77	27	3.96	4 845
Chatino de Yaitepec	293	290	23	4.45	33 199
Russe	226	208	26	5.19	53 858

TABLEAU 6.3 – Mesures de canonicité sur les treillis de patrons d'alternance.

Étant donné un treillis comportant une base de b atomes et présentant une hauteur h, nous normalisons cette hauteur par la hauteur maximale possible sur b atomes, à savoir $b - 1$ (en ignorant les arcs en direction de l'infimum) :

$$\text{NH}(\mathcal{B}\langle X, Y, I\rangle) = \frac{h}{b-1} \tag{6.1}$$

Cette hauteur maximale correspond au cas le moins canonique, où le treillis constitue l'ensemble des parties de l'ensemble des b atomes. On compte alors au total $n = 2^b - 1$ concepts [5]. Nous normalisons donc le nombre de concepts d'un treillis par cette valeur maximale, et appelons cette mesure DENSITÉ. Soit n le nombre de concepts d'un treillis $\mathcal{B}\langle X, Y, I\rangle$ et b le nombre de ses atomes, la densité de ce treillis est :

$$\text{densité}(\mathcal{B}\langle X, Y, I\rangle) = \frac{n}{2^b - 1} \tag{6.2}$$

La figure 6.3 présente ces valeurs pour chaque système. La croissance de 2^b est telle que relativement au maximum concevable, nos treillis sont très peu fournis en nœuds, et ce même s'ils semblent très denses à l'œil nu (toutes les valeurs de densité sont inférieures à 10^{-10}). La hauteur relative, elle, varie d'un système à l'autre. D'une part les systèmes du chatino, de

5. Il faut à nouveau soustraire 1 pour ignorer l'infimum, c'est à dire l'ensemble vide pour l'extension. Notons que la taille de ce treillis maximale croît très rapidement en fonction de b.

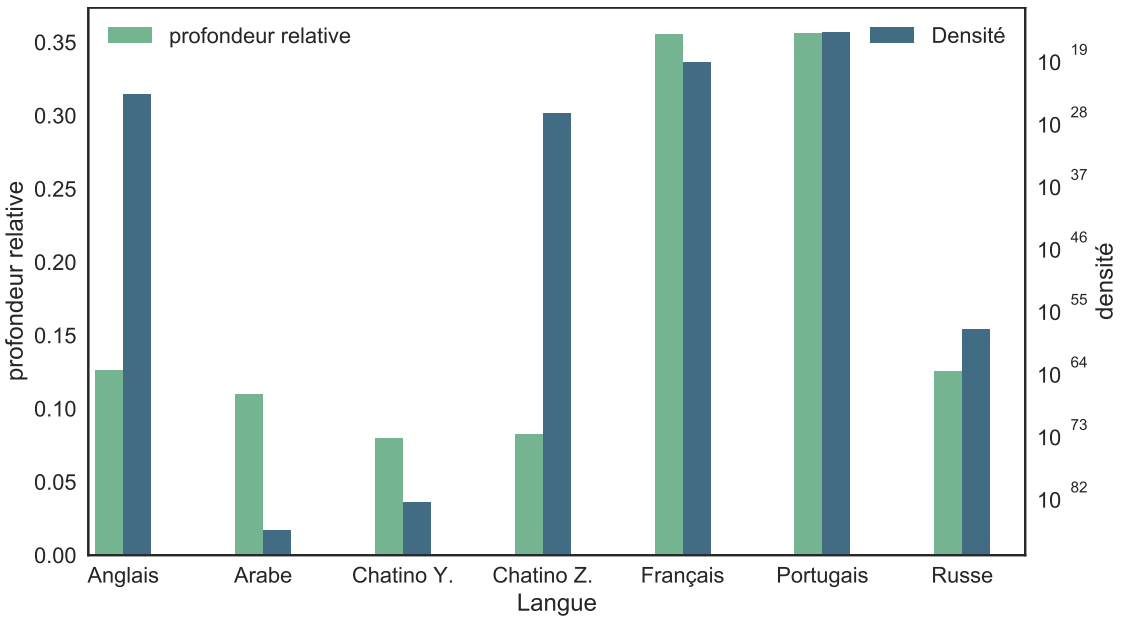


FIGURE 6.3 – Canonicité des treillis de patrons d'alternance.

l'arabe, du russe et de l'anglais présentent une hauteur inférieure à 0.15, tandis que les deux autres systèmes présentent une hauteur relative autour de 0.35, indicative d'un plus grand nombre de classes intermédiaires.

Dans l'ensemble, ces résultats montrent que les classifications obtenues sont très complexes à vue d'œil, et manifestement non canoniques. Elles permettent donc de rejeter sans hésitation l'hypothèse de la canonicité, selon laquelle des partitions seraient un modèle approprié des classes flexionnelles. Cependant, ces systèmes sont également beaucoup moins complexes que le maximum théoriquement possible.

6.4 Lisibilité des treillis de classes flexionnelles

Les classifications fondées sur des concepts reliant microclasses et patrons d'alternances contiennent trop de concepts pour être lisibles par un humain. Cela limite considérablement leur utilité pratique pour l'exploration qualitative des propriétés des systèmes flexionnels. Or elles sont redondantes de deux façons distinctes : d'une part, un grand nombre de concepts

	vole	oiseau	insecte	ailé	aquatique	rapace	parle
papillon	×		×	×			
perroquet	×	×		×			×
pingouin		×		×	×		
aigle	×	×		×		×	

Tableau 6.4 – Contexte de quelques animaux ailés.

n'introduit ni objets (microclasses), ni propriétés (patrons), et leur sémantique est celle d'une conjonction. D'autre part, nous avons dit que les patrons d'alternance calculés sur l'ensemble des paires de cases fournissent des caractérisations intentionnellement redondantes, permettant de capter l'ensemble des points de similarités possibles entre microclasses. Cette section propose une méthode permettant d'inférer des hiérarchies plus lisibles en opérant une simplification du système.

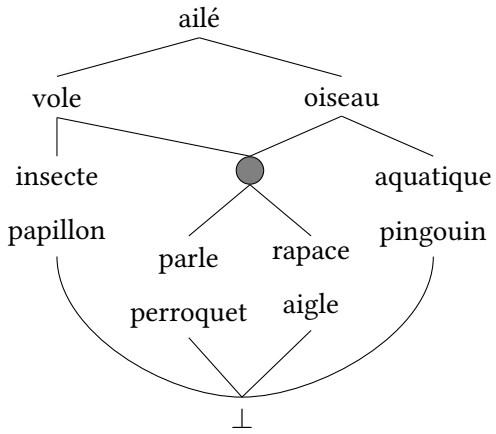
6.4.1 Hiérarchies des concepts objets et attributs

Dans un treillis de concepts, nous avons vu que l'affichage réduit permet de n'indiquer les objets et les propriétés que sur les concepts qui les définissent. Un concept introduisant un objet est un CONCEPT OBJET (*object concept* en anglais) et un concept introduisant une propriété est un CONCEPT PROPRIÉTÉ (*attribute concept* en anglais). Certains concepts n'introduisent ni objets, ni propriétés. Ainsi, dans un treillis de concepts représenté en affichage réduit, certains concepts n'ont aucune étiquette.

Le tableau 6.4 illustre la situation pertinente pour un contexte non-linguistique. Le treillis en (78) est le treillis des concepts de ce contexte. Le concept indiqué dans le treillis par un cercle gris, et dans le tableau par des cases grisées, est $\langle$ {perroquet, aigle}, {oiseau, vole} $\rangle$. Ce concept ne définit pas l'ensemble des animaux ailés qui volent, car le papillon en fait également partie. Il ne définit pas non plus l'ensemble des oiseaux, car le pingouin en est un. Ce n'est donc pas un concept propriété. De même, il ne définit pas l'ensemble des perroquets, car les perroquets ont également la propriété de parler. Il ne définit pas plus l'ensemble des aigles, car les aigles ont la

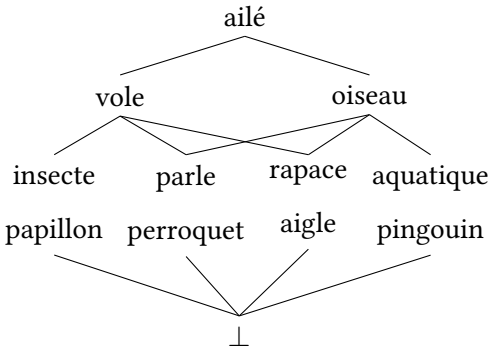
propriété d'être des rapaces. Ce concept a pour sémantique une conjonction : les perroquets et les aigles ont en commun d'appartenir à la fois au concept objet des animaux ailés qui volent et au concept objet des animaux ailés qui sont des oiseaux.

(78)



Petersen (2001) nomme AOC-POSET, ou *attribute object concept partial ordered set* l'ensemble des concepts objets et des concepts propriété d'un contexte, ordonnés par inclusion. Il s'agit d'un sous-ensemble des concepts du treillis. L'exemple (79) présente la hiérarchie résultant de cet ensemble. Elle se distingue de (78) par l'absence du concept ⟨ {perroquet, aigle}, {oiseau, vole} ⟩. À la place, des arcs supplémentaires ont été introduits. La hiérarchie obtenue peut se lire de la même façon que le treillis des concepts, mais il ne s'agit pas d'un treillis. En effet, en (78), toute paire de concepts avait une unique borne supérieure et une unique borne inférieure. En particulier, la borne haute pour ⟨ {perroquet}, {oiseau, vole, parle} ⟩ et ⟨ {aigle}, {oiseau, vole, rapace} ⟩ était le concept représenté en gris ⟨ {perroquet, aigle}, {oiseau, vole} ⟩. Dans la hiérarchie (79), il existe deux bornes hautes concurrentes pour ces concepts : ⟨ {perroquet, aigle, papillon}, {vole} ⟩ et ⟨ {perroquet, aigle, pingouin}, {oiseau} ⟩.

(79)



Selon Osswald et Petersen (2003), « Les *AOC-posets* fournissent une méthode très simple pour induire des hiérarchies d'héritages sans redondance à partir d'immenses bases de données [6] ». La taille de ces hiérarchies est bornée par la somme du nombre d'objets et du nombre d'attributs, ce qui les rend beaucoup plus compactes que les treillis de concepts. La perte des propriétés d'un treillis n'est pas un problème pour nos besoins, et nous pensons donc qu'il est utile de restreindre nos hiérarchies à ces sous-ensembles. Osswald et Petersen (2002) utilisent justement ces structures pour inférer des classifications linguistiques.

6.4.2 Choix d'un sous-ensemble de patrons

Un second type de redondance est à l'œuvre dans nos treillis de concepts fondés sur les patrons d'alternance : les patrons d'alternance permettent de prédire n'importe quelle forme à partir de n'importe quelle autre. Cette redondance est nécessaire afin de rendre compte de l'ensemble des points de similarités possibles entre les lexèmes. Face à de très grands paradigmes, il peut cependant être souhaitable de renoncer à certaines relations de similarité afin de pouvoir explorer manuellement les hiérarchies produites.

Albright (2002) propose une solution intuitivement satisfaisante à ce problème :

> Si nous disions simplement qu'une grammaire contient les règles qui relient les formes du paradigme entre elles, et n'ajoutions rien de plus (comme par exemple le font Bochner 1993, Barr 1994, et Neuvel et Singh 2001), nous nous retrouverions à devoir inclure dans la grammaire un nombre énorme de relations binaires.

6. [En anglais dans le texte] « *AOC-posets provide a very simple method to induce redundancy-free inheritance hierarchies from huge databases* ».

> Intuitivement, la réponse à ce problème est que les locuteurs n'apprennent probablement pas les règles pour toutes les relations binaires dans le paradigme, mais plutôt pour un sous ensemble des relations possibles [7].
> Albright (2002, p.8)

Il propose donc de restreindre les règles (nos patrons), de façon à ce qu'elles incluent toujours certaines cases de paradigmes, dites *bases*. La façon la plus radicale de réduire ainsi les alternances, qui est employée explicitement ou implicitement dans de nombreuses grammaires, est de s'appuyer sur une base unique. Albright (2002) nomme « HYPOTHÈSE DE LA BASE UNIQUE [8] » l'hypothèse selon laquelle en général, les formes d'un paradigmes sont toutes formées sur une unique forme. La figure 6.4 présente l'ensemble des relations qui seraient conservées pour les verbes du français si l'on considère l'infinitif comme case de base.

Cependant, ce choix réduit énormément le nombre d'alternances considérées, et perd tellement d'information que la structure hiérarchique résultante est très peu informative.

Une autre possibilité pour se concentrer sur certaines relations d'interprédictibilité entre cases consiste à suivre l'idée de Boyé (2000) et Bonami et Boyé (2003a), qui organisent les cases de paradigmes en arbres de dépendance implicative. Sagot (2013) propose une manière d'opérationnaliser cette idée. Il utilise l'entropie conditionnelle pour sélectionner un arbre couvrant d'entropie conditionnelle minimale dans le graphe de toutes les alternances possibles entre cases de paradigme.

Sagot (2013) représente les combinaisons entre patrons par un graphe orienté, dans lequel l'arc entre une case c_1 et une case c_2 porte pour poids l'entropie $H(c_1 \Rightarrow c_2)$. Afin de trouver l'arbre couvrant dans ce graphe, il emploie l'algorithme de Chu-Liu-Edmonds, qui requiert le choix d'une racine *a priori*. En français, il choisit l'infinitif. Il obtient donc un arbre enraciné dans lequel toutes les formes peuvent être déduites à partir de la racine, ou à partir d'un

7. [En anglais dans le texte] « *If we were to say simply that a grammar contains rules relating forms in the paradigm to one another and leave it at that (as, for example, Bochner 1993, Barr 1994, and Neuvel and Singh 2001 do), we would be left with an enormous number of pairwise relations to include in the grammar. Intuitively, the answer to this problem is that speakers probably do not learn rules for every pairwise relation in the paradigm, but rather for only a subset of the possible relations* ».

8. [En anglais dans le texte] « *single base hypothesis* ».

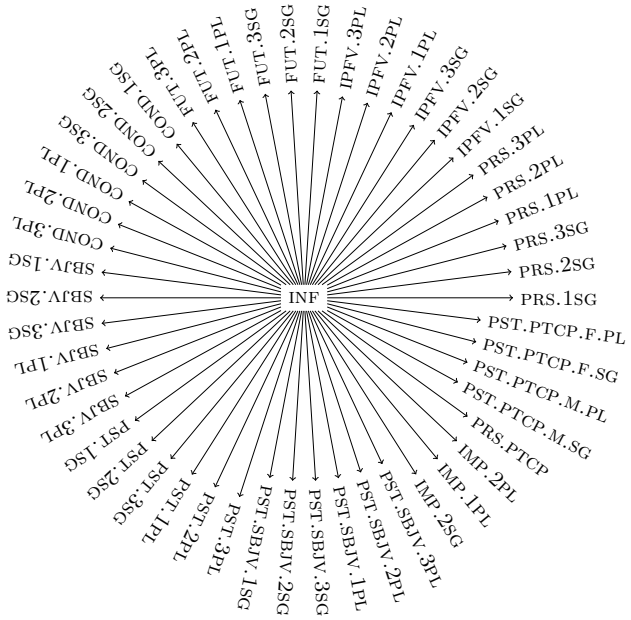


FIGURE 6.4 – Alternances impliquant l'infinitif au sein des paradigmes de verbes français.

de leur ancêtre. Nous présentons en figure 6.5 l'arbre qu'il obtient pour les verbes du français. Cet arbre représente une structure de paradigme optimale, c'est à dire qu'il présente pour chaque case ses meilleurs prédicteurs. Lorsque deux cases sont entièrement interprédictibles, elles sont indiquées sur un même nœud de l'arbre. À nouveau, cette solution produit une coupe trop importante dans le graphe pour fonder une classification hiérarchique des comportements flexionnels.

Nous proposons, plutôt que de regarder les alternances les plus prédictibles, de nous concentrer au contraire sur les alternances les moins prédictibles. Pour ce faire, nous réduisons notre attention aux alternances au sein des distillations de paradigmes produites au chapitre 3 sur la base de zones de prédictibilités quasi-catégoriques. Le procédé est simple : pour chaque zone d'interprédictibilité, on ne conserve qu'une case de paradigme. Cette stratégie a un avantage majeur : elle produit, comme nous l'avons vu au chapitre 3, une plus grande réduction du nombre de cases dans les grands paradigmes (de 51 à 9 en français) que dans les petits paradigmes (aucune réduction en chatino de Zenzontepec). La figure 6.6 indique les alternances considérées pour les verbes du français.

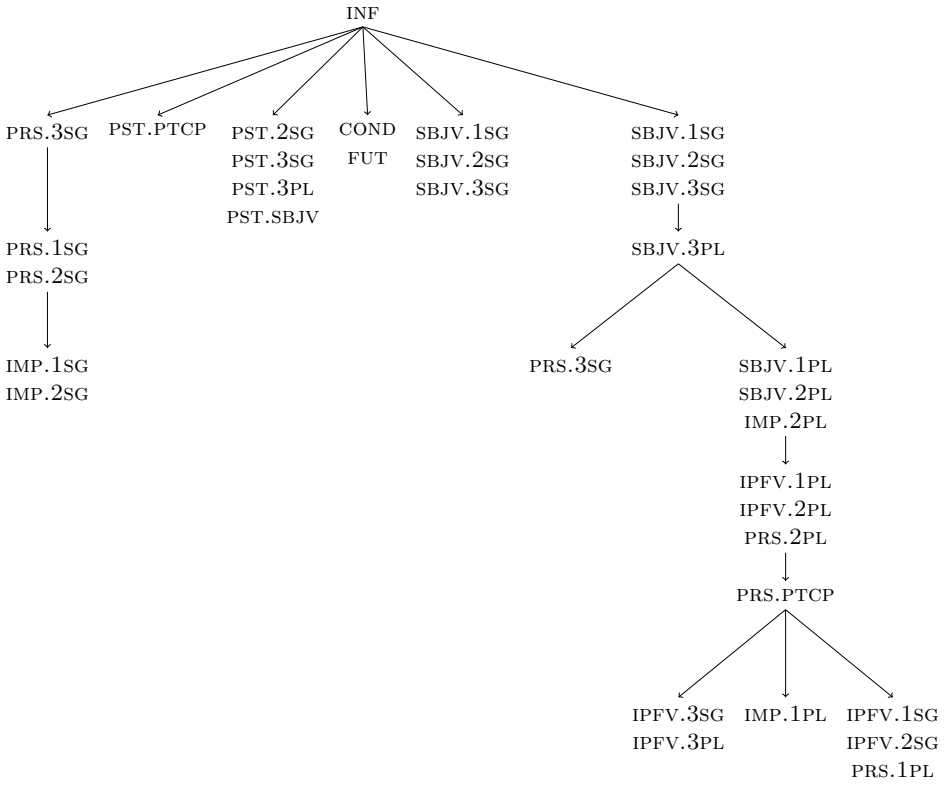


FIGURE 6.5 – Alternances dans l'arbre couvrant de prédictibilité maximale (français).

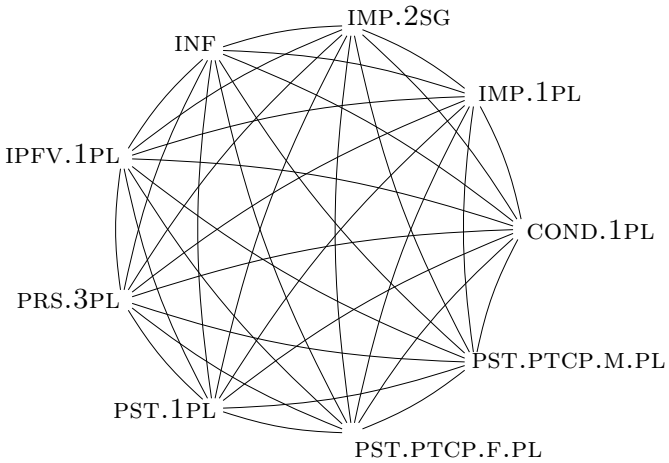


FIGURE 6.6 – Alternances à travers les zones d'interprédictibilité (français).

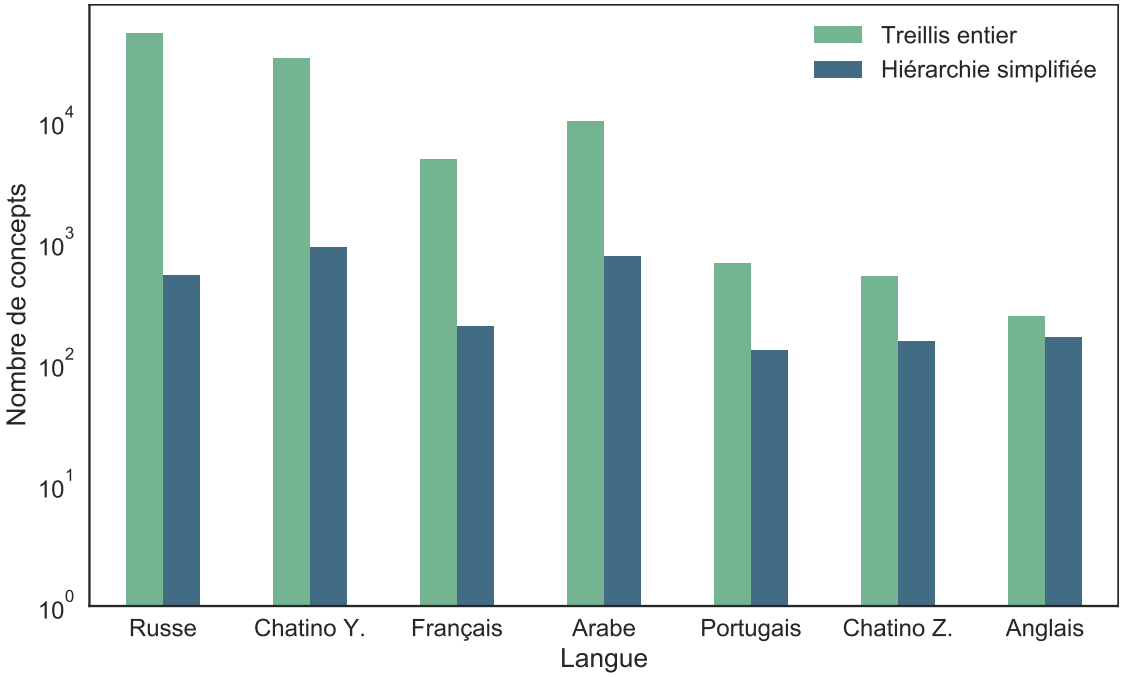


FIGURE 6.7 – Comparaison de la taille des classifications exhaustives et simplifiées (échelle logarithmique).

6.4.3 Bilan

Le résultat de ces deux simplicifications (hiérarchies AOC, sélection de cases en fonction des zones d'interprédictibilité) produit bien entendu une vue partielle du treillis. Cependant, la perte d'information est minimale, et les simplifications rendent possible une exploration manuelle des hiérarchies produites, ouvrant la possibilité d'en faire usage pour des analyses fines.

La méthode que nous avons présentée permet d'inférer des hiérarchies de taille beaucoup plus modeste que les treillis entiers calculés sur ces mêmes systèmes. La figure 6.7 et le tableau 6.5 comparent le nombre de concepts dans les treillis exhaustifs et les hiérarchies flexionnelles simplifiées. En russe, où la réduction est la plus dramatique, la hiérarchie obtenue comporte 0.9% des concepts de la hiérarchie complète. En anglais, à l'opposé, elle comporte 70% des concepts d'origine.

Langue	Treillis entier	Hiérarchie simplifiée	Proportion du treillis entier
Russe	53858	538	0.01
Chatino Y.	33199	922	0.03
Français	4845	201	0.04
Arabe	10125	775	0.08
Portugais	677	127	0.19
Chatino Z.	524	152	0.29
Anglais	244	164	0.67

TABLEAU 6.5 – Comparaison du nombre de concepts des classifications exhaustives et simplifiées.

6.4.4 Quelques exemples

Il n'est pas possible dans le cadre de cette thèse d'explorer systématiquement les hiérarchies obtenues pour les 8 systèmes étudiés. Dans cette section, nous nous contentons de commenter succinctement quelques-unes des hiérarchies simplifiées. Il n'est pas possible de discuter l'ensemble des concepts de ces hiérarchies. Afin de pouvoir explorer les hiérarchie, nous avons conçu une interface en HTML et javascript [9]. Celle-ci permet de se déplacer dans la hiérarchie, de zoomer sur des nœuds choisis par l'utilisateur. Le survol d'un nœud indique la taille de son extension, un exemple de lexème dans son extension, ainsi que l'ensemble des propriétés qu'il définit. Un clic sur un nœud active ou désactive sa visibilité, ainsi que celle de l'ensemble de ses enfants. Les captures d'écran de la figure 6.8 donnent un aperçu de ces fonctionnalités.

En ce qui concerne l'anglais, nous conservons les quatre cases de paradigme INF, PAST13, PPART, PRES3S et PRESPART, correspondant aux quatre zones d'interprédictibilité trouvées au chapitre 3. La figure 6.9 présente la hiérarchie résultante, qui comporte seulement 164 concepts, dont 117 coïncident avec une microclasse.

Le concept au sommet de la hiérarchie a l'ensemble vide pour intension, et l'ensemble des microclasses pour extension : il n'y a donc aucune propriété partagée par absolument tous les

9. Celle-ci s'appuie sur la librairie python mpld3 : http://mpld3.github.io/.

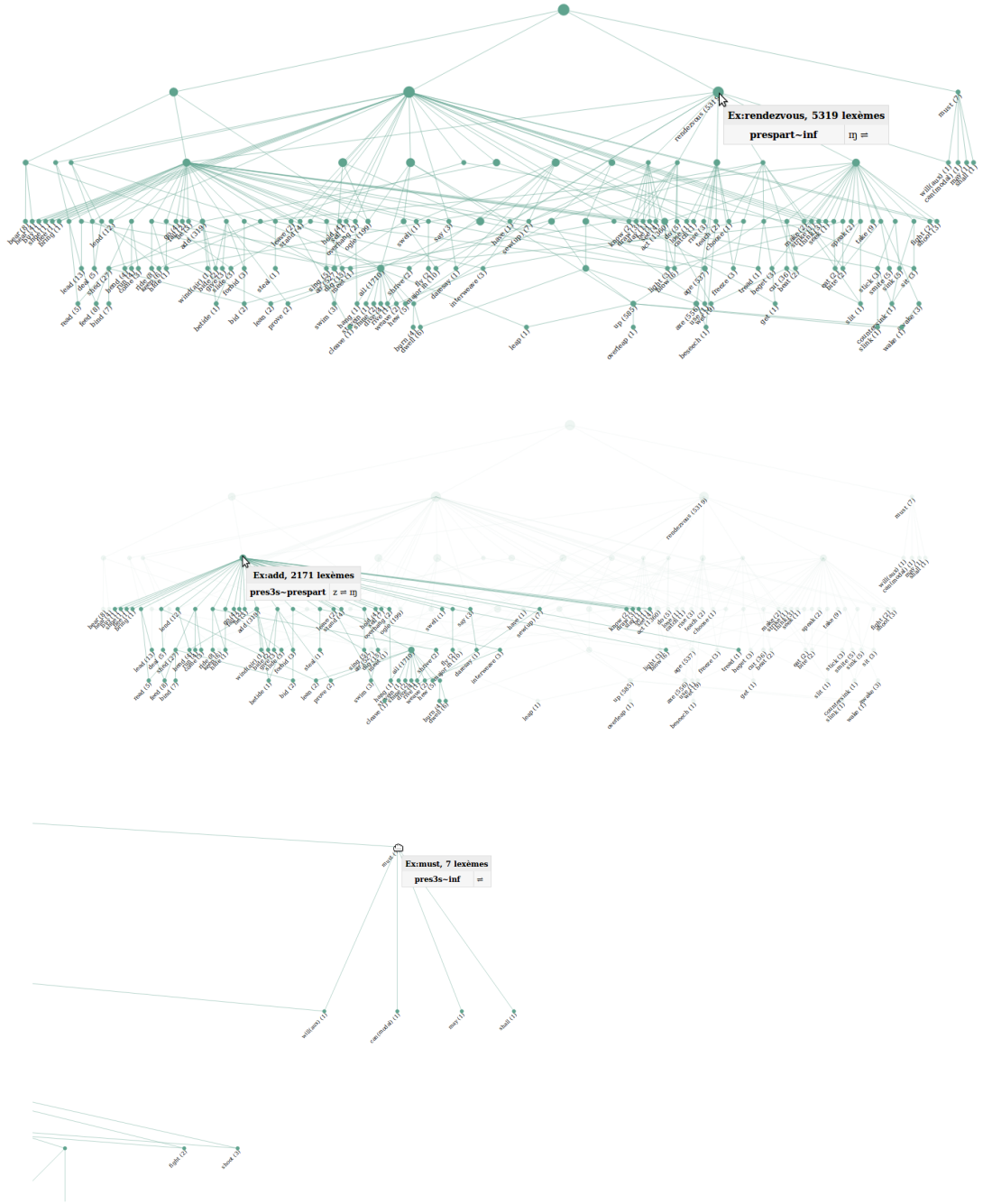


FIGURE 6.8 – Interface HTML d'exploration des hiérarchies.

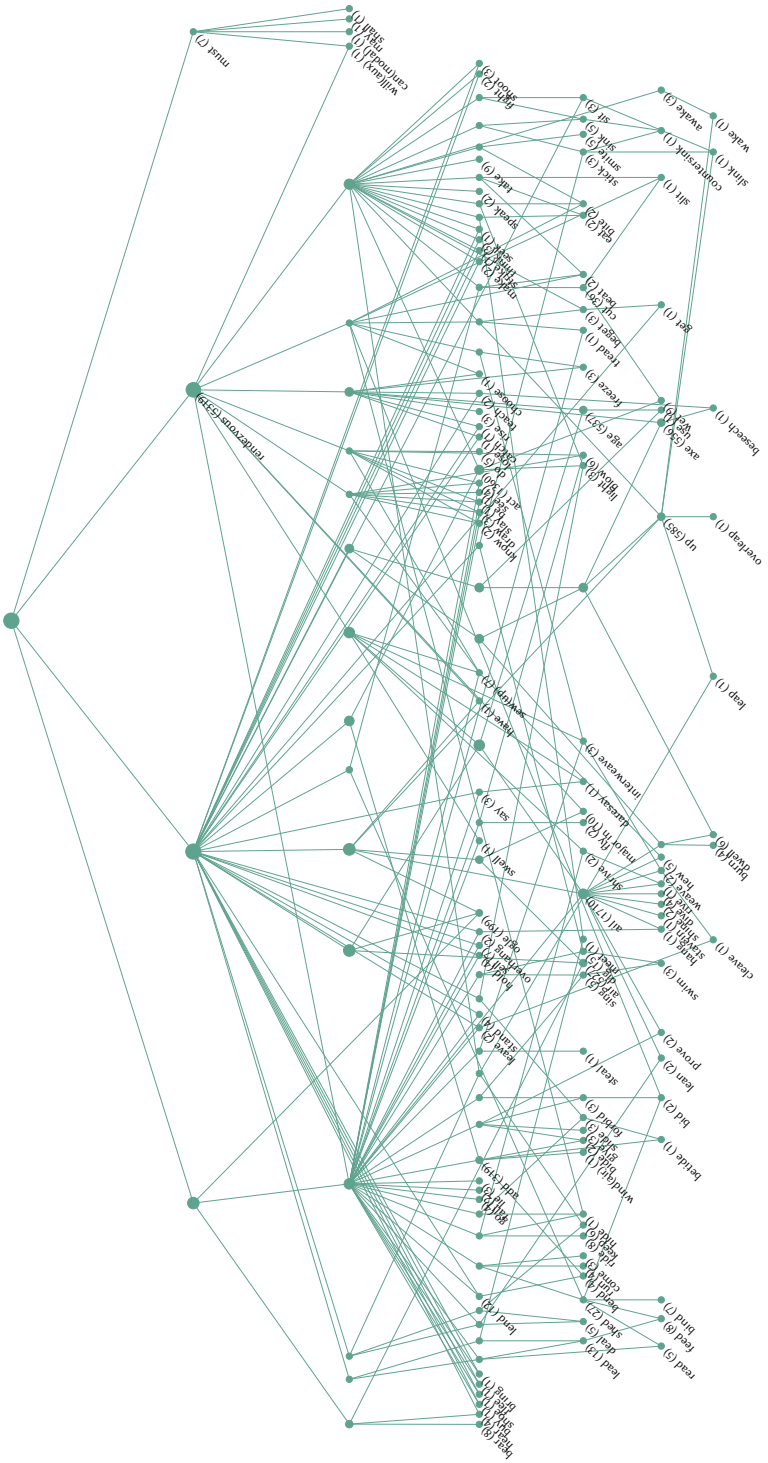


FIGURE 6.9 – Hiérarchie simplifiée des verbes de l'anglais.

verbes. Il a quatre enfants, que nous listons par taille décroissante de leur extension :

(i) 5950 lexèmes

— PPART⇌PAST13 :ϵ ⇌ ϵ/ X+

(ii) 5319 lexèmes (introduit RENDEZ-VOUS)

— PRESPART⇌INF :ɪŋ ⇌ ϵ/ X+_

(iii) 2909 lexèmes

— PRES3S⇌INF :z ⇌ ϵ/ X+_

(iv) 7 lexèmes (introduit MUST)

— PRES3S⇌INF :ϵ ⇌ ϵ/ X+[dlnrstzðθ]*

Le concept (iv) concerne les modaux, dont seul WILL partage quoique ce soit avec d'autres verbes, ceux du concept (ii). Le concept (i) concerne l'écrasante majorité des lexèmes, qui ont un participe passé identique à leur passé. Le concept (ii) a également une très grande extension. Il exclut 539 lexèmes du type BEAR et HEAR dans lesquels les /r/ disparaissent en finale à l'infinitif. Ces verbes suivent donc un patron PRESPART⇌INF :rɪŋ ⇌ ϵ. Il exclut également les modaux CAN, MAY, SHALL, MUST qui n'ont pas de participe présent, ainsi que 199 verbes comme OGLE qui instancient un patron [lmnŋ]ɪŋ ⇌ [m̩n̩l̩n̩], où les liquides finales deviennent syllabiques. Enfin le concept (iii) réunit les lexèmes qui forment leur troisième personne de présent en /-z/. Il s'oppose à des sous-concepts de (ii).

En ce qui concerne le français, nous obtenons la hiérarchie présentée dans la figure 6.10. Les alternances conservées concernent les cases de paradigme COND.1PL, IMP.1PL, IMP.2SG, INF, IPFV.1PL, PRS.3PL, PST.1PL, PST.PTCP.F.PL, et PST.PTCP.M.PL. Quoique cette hiérarchie semble toujours assez complexe, elle est beaucoup plus simple que la hiérarchie entière présentée en figure 6.2.

Une autre façon d'étudier les hiérarchies flexionnelles produites de cette façon consiste à les comparer à celles qui sont décrites par des linguistes. Pour le français, Kilani-Schoch et Dressler (2005) proposent une hiérarchie arborescente détaillée de l'organisation des verbes du français. Celle-ci s'organise en deux macroclasses, selon une division fondée sur la productivité des comportements flexionnels : la macroclasse I comporte les classes productives, et la

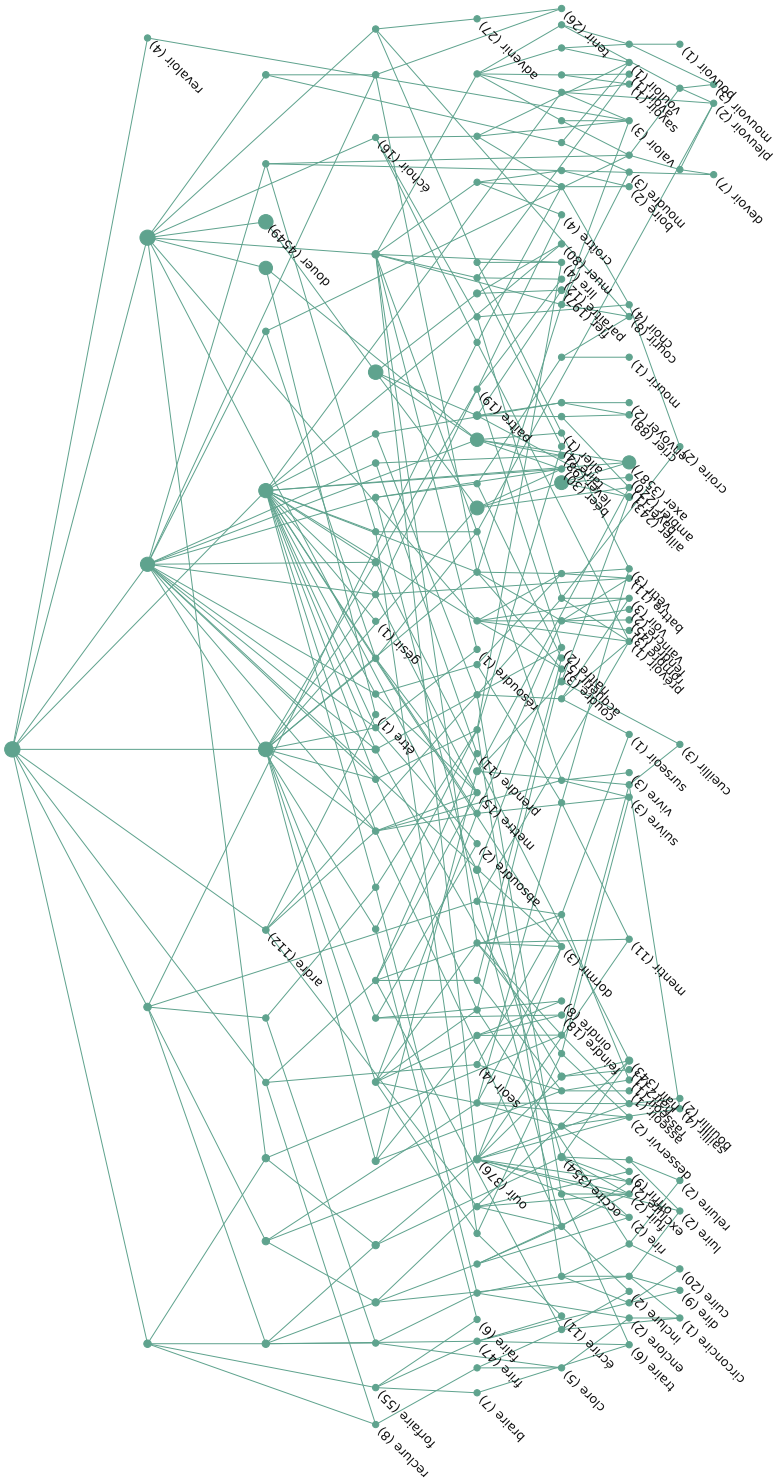


FIGURE 6.10 – Hiérarchie simplifiée des verbes du français.

macroclasse II comporte l'ensemble des classes non productives. Chacune est subdivisée à leur son tour en sous-classes, sous-sous classes, etc, jusqu'aux microclasses. Ces subdivisions sont cette fois fondées sur des propriétés d'exponence. Nous présentons succinctement la structure de la hiérarchie résultante dans la figure 6.11.

Dans leur description, Kilani-Schoch et Dressler (2005) proposent des implications par défaut, ou « conditions de structure paradigmatique [10] », pour chaque classe de la hiérarchie. Nous rappelons dans les exemples (80) à (82) quelques unes de ces implications présentées au chapitre 1.

(80) Macroclasse I : Infinitif /X+e/ ⇒
- Participe passé = /X+e/
- Passé simple première personne du singuliers = /X+e/
- Présent singulier = /X/
- Présent indicatif troisième personne du pluriel = /X/
- Subjonctif présent = /X/

(81) Classe I.1 : Imparfait parl+ɛ, futur parl+ər+e.

(82) Classe II.2 : Infinitif /Xwar/ ⇒
- Participe passé en /y/
- Passé simple en /y/
- par défaut, /wa/ fait partie de l'infinitif

Afin de comparer cette hiérarchie à celle que nous produisons automatiquement, nous augmentons notre hiérarchie avec de l'information provenant de celle de Kilani-Schoch et Dressler (2005). Plus précisément, nous créons une propriété pour chaque nœud (classe, sous-classe, etc) de leur hiérarchie, en excluant les microclasses pour simplifier la comparaison. Nous attribuons ces propriétés aux objets de notre classification (les 2549 verbes du français). Nous accolons ce contexte à celui des patrons (après sélection), et nous produisons une nouvelle hiérarchie. Dans celle-ci, que nous présentons en figure 6.12, les concepts peuvent introduire des patrons d'alternance (ils sont dessinés en vert clair dans la figure) ou des classes de Kilani-Schoch et Dressler (2005) (ils sont dessinés en violet foncé dans la figure). La hiérarchie comprend 208 concepts, dont 20 concepts appartenant à la classification de Kilani-Schoch et Dressler (2005). Quoique

10. [En anglais dans le texte] « *Paradigm Structure Conditions (PSCs)* ».

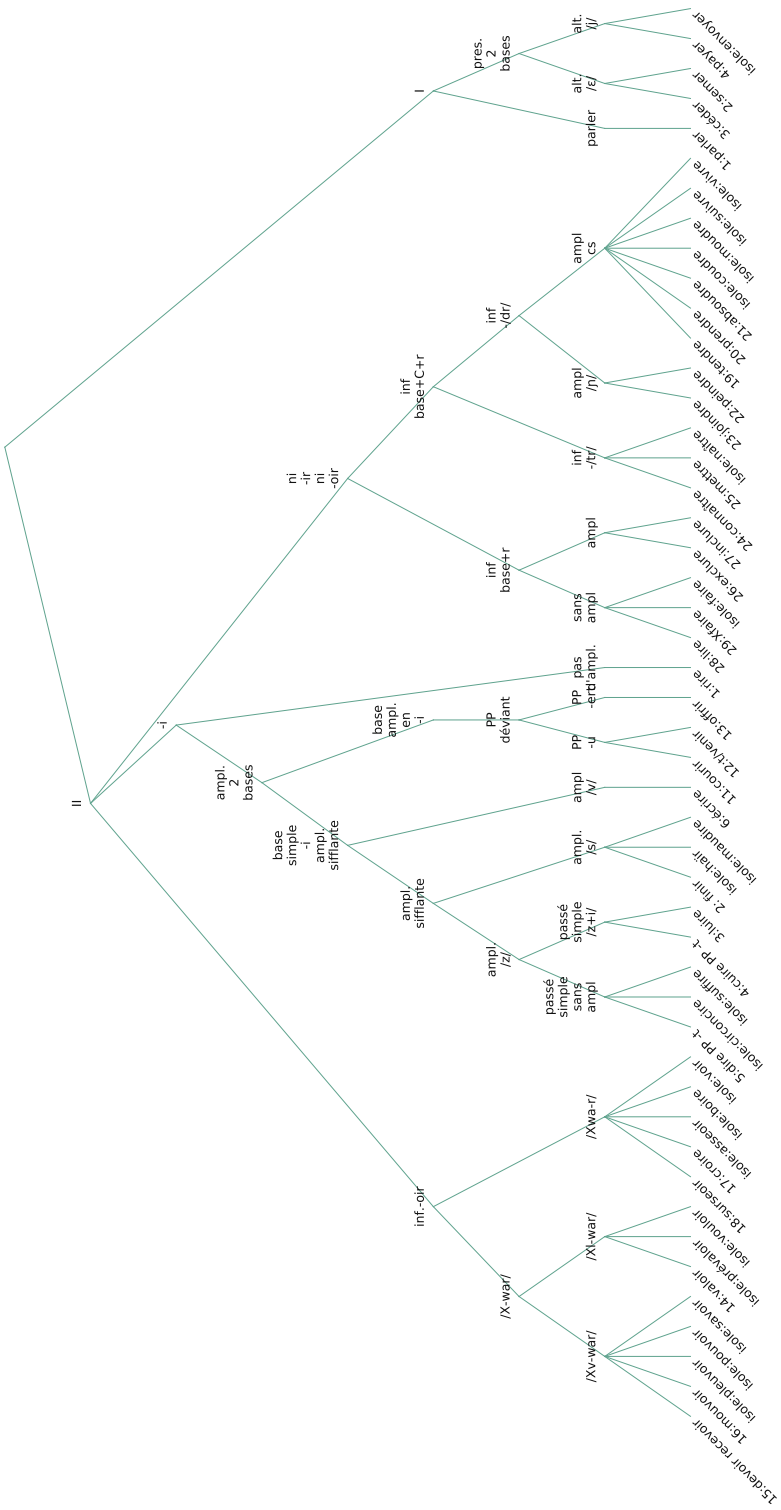


FIGURE 6.11 – Hiérarchie des verbes du français selon Kilani-Schoch et Dressler (2005).

ce soit théoriquement possible, nous ne trouvons aucun concept qui présente les deux types de propriétés.

Rappelons qu'il est possible de lire des implications dans les hiérarchies de concepts, car tout concept implique l'ensemble de ses ancêtres. Dans la hiérarchie comparée de la figure 6.12, l'intension de la racine est vide : il n'existe pas de propriétés communes à l'ensemble des lexèmes. Les deux macroclasses de Kilani-Schoch et Dressler (2005) sont les enfants de la racine. En conséquence, il n'y a pas d'implications catégoriques possibles à partir des macroclasses : appartenir à l'une ou l'autre macroclasse ne nous informe pas avec certitude sur les patrons instanciés. On peut en conclure que les macroclasses de Kilani et Dressler ne caractérisent pas des comportements morphophonologiques au sein des alternances retenues.

Cependant les implications qui constituent les conditions de structure paradigmatique sont des défauts, et non des règles catégoriques. Observons donc les principaux concepts qui impliquent d'appartenir à la macroclasse I. Nous listons les patrons définis par ces concepts en tableau 6.6. Pour chacun, nous indiquons la proportion des lexèmes de la macroclasse qui sont concernés par l'antécédent. Il existe d'autres implications valides, mais elles concernent au plus 12% des lexèmes de la macroclasse I. Pratiquement tous les lexèmes de la macroclasse I (99.93%) présentent un patron identité entre le participe passé et l'infinitif. De façon peu surprenante, puisque cette macroclasse est fondée sur un exposant affixal /-e/ à l'infinitif, les autres implications portent principalement sur des patrons qui concernent l'infinitif ou le participe passé en /-e/. Ces implications constituent de bons candidats à être des conditions de structure paradigmatique par défaut.

Nous présentons dans le tableau 6.7 tous les patrons des concepts qui impliquent la macroclasse II et concernent plus de 50% de ses lexèmes. L'hétérogénéité de la classe apparaît tout d'abord à travers l'absence d'implication quasi-catégoriques : le concept qui implique la macroclasse II avec le plus de fiabilité concerne seulement 70% de ses lexèmes. Par ailleurs, les patrons concernés sont de nature beaucoup plus variables que ceux qui impliquent la macroclasse 1.

L'observation des implications fiables dans la hiérarchie peut permettre de constituer des ensembles de *conditions de structure paradigmatique* obtenues automatiquement et dont la qualité est quantifiable (voir section 1.3.4). Nous laissons cette entreprise à de futurs travaux.

Figure 6.12 – Hiérarchie comparée des verbes du français.

Patrons	Proportion de la macroclasse I
INF⇌PST.PTCP.F.PL :ϵ ⇌ ϵ,	99.93 %
INF⇌PST.PTCP.M.PL :ϵ ⇌ ϵ	
IMP.1PL⇌INF :ɔ̃ ⇌ E ,	99.89 %
IMP.1PL⇌PST.PTCP.F.PL :ɔ̃ ⇌ E ,	
IMP.1PL⇌PST.PTCP.M.PL :ɔ̃ ⇌ E ,	
PST.1PL⇌IMP.1PL :am ⇌ ɔ̃ ,	
PST.1PL⇌INF :am ⇌ E ,	
PST.1PL⇌PST.PTCP.F.PL :am ⇌ E ,	
PST.1PL⇌PST.PTCP.M.PL :am ⇌ E	
IMP.2SG⇌INF :ϵ ⇌ E ,	89.65 %
IMP.2SG⇌PST.PTCP.F.PL :ϵ ⇌ E ,	
IMP.2SG⇌PST.PTCP.M.PL :ϵ ⇌ E ,	
PRS.3PL⇌INF :ϵ ⇌ E ,	
PRS.3PL⇌PST.1PL :ϵ ⇌ am ,	
PRS.3PL⇌PST.PTCP.F.PL :ϵ ⇌ E ,	
PRS.3PL⇌PST.PTCP.M.PL :ϵ ⇌ E ,	
PST.1PL⇌IMP.2SG :am ⇌ ϵ	
COND.1PL⇌INF :Øʁjɔ̃ ⇌ E ,	88.97 %
COND.1PL⇌PST.1PL :Øʁjɔ̃ ⇌ am ,	
COND.1PL⇌PST.PTCP.F.PL :Øʁjɔ̃ ⇌ E	
IPFV.1PL⇌INF :jɔ̃ ⇌ E ,	81.63 %
IPFV.1PL⇌PST.1PL :jɔ̃ ⇌ am	

TABLEAU 6.6 – Patrons définis par les principaux concepts qui impliquent l'appartenance à la macroclasse I.

Patrons	Proportion de la macroclasse I
PST.1PL⇌PST.PTCP.M.PL :m ⇌ ϵ	70.59 %
COND.1PL⇌INF :jɔ̃ ⇌ ϵ	66.28 %
PST.1PL⇌PST.PTCP.F.PL :m ⇌ ϵ	63.99 %
INF⇌PST.PTCP.M.PL :ʁ ⇌ ϵ	63.85 %
IMP.2SG⇌PST.PTCP.M.PL :ϵ ⇌ ϵ	63.13 %
COND.1PL⇌PST.PTCP.M.PL :ʁjɔ̃ ⇌ ϵ	62.12 %
IMP.2SG⇌INF :ϵ ⇌ ʁ	61.55 %
PST.1PL⇌INF :m ⇌ ʁ	57.82 %
COND.1PL⇌PST.1PL :ʁjɔ̃ ⇌ m	57.39 %
INF⇌PST.PTCP.F.PL :ʁ ⇌ ϵ	54.81 %
COND.1PL⇌PST.PTCP.F.PL :ʁjɔ̃ ⇌ ϵ	54.38 %
PST.1PL⇌IMP.2SG :m ⇌ ϵ	52.08 %
IMP.2SG⇌IMP.1PL :ϵ ⇌ sɔ̃ , IPFV.1PL⇌IMP.2SG :sjɔ̃ ⇌ ϵ, PRS.3PL⇌IMP.2SG :s ⇌ ϵ	51.94 %
IMP.2SG⇌PST.PTCP.F.PL :ϵ ⇌ ϵ	51.08 %

TABLEAU 6.7 – Patrons définis par les principaux concepts qui impliquent l'appartenance à la macroclasse I.

Passons maintenant au système du chatino de Zenzontepec. La figure 6.13 présente la hiérarchie simplifiées pour ce système. Puisqu'il n'y a pas de zones d'interprédictibilité, nous conservons les quatre cases de paradigme, et la simplification consiste seulement à ignorer les nœuds disjonctifs. La hiérarchie combine les patrons tonaux, dont les concepts sont indiqués en violet foncé, et les patrons segmentaux, dont les concepts sont indiqués en vert clair. Aucun concept ne présente à la fois des patrons tonaux et segmentaux. Cette observation témoigne de l'orthogonalité des deux systèmes. Outre quatre concepts introduisant des patrons identité tonaux qui dominent la hiérarchie, la hiérarchie est assez plate, indiquant qu'il existe peu de relations implicatives dans ce système, et en particulier, peu de relations implicatives entre le système tonal et le système segmental.

La figure 6.14 compare la hiérarchie des patrons segmentaux du chatino de Zenzontepec (nœuds vert clair) avec celle proposée par Campbell (2014, nœuds prune annotés). .

Nous listons ci-dessous les implications au sein de cette hiérarchie qui mettent en jeu une classe de Campbell (2014) et concernent au moins 20% des lexèmes du concept impliqué :

(i) Bc $\Longrightarrow$ HAB$\rightleftharpoons$PROG :i ⇌ e (53.38%)

(ii) Bc $\Longrightarrow$ CPL$\rightleftharpoons$PROG :ku ⇌ te (72.45%)

(iii) Bc $\Longrightarrow$ CPL$\rightleftharpoons$HAB :ku ⇌ ti (89.87%)

(iv) Bt $\Longrightarrow$ POT$\rightleftharpoons$HAB :ϵ ⇌ n (30.86%)

(v) At $\Longrightarrow$ CPL$\rightleftharpoons$HAB :kat ⇌ ty , CPL$\rightleftharpoons$POT :nkat ⇌ ty (50.0%)

(vi) CPL$\rightleftharpoons$PROG :ky ⇌ tey , POT$\rightleftharpoons$PROG :ch ⇌ ntey $\Longrightarrow$ By (57.14%)

(vii) HAB$\rightleftharpoons$PROG :ch ⇌ tey $\Longrightarrow$ By (55.1%)

(viii) HAB$\rightleftharpoons$PROG :u ⇌ e $\Longrightarrow$ A (53.89%)

(ix) CPL$\rightleftharpoons$HAB :ka ⇌ tu $\Longrightarrow$ A (52.33%)

FIGURE 6.13 – Hiérarchie simplifiée des verbes du zenzontepec.

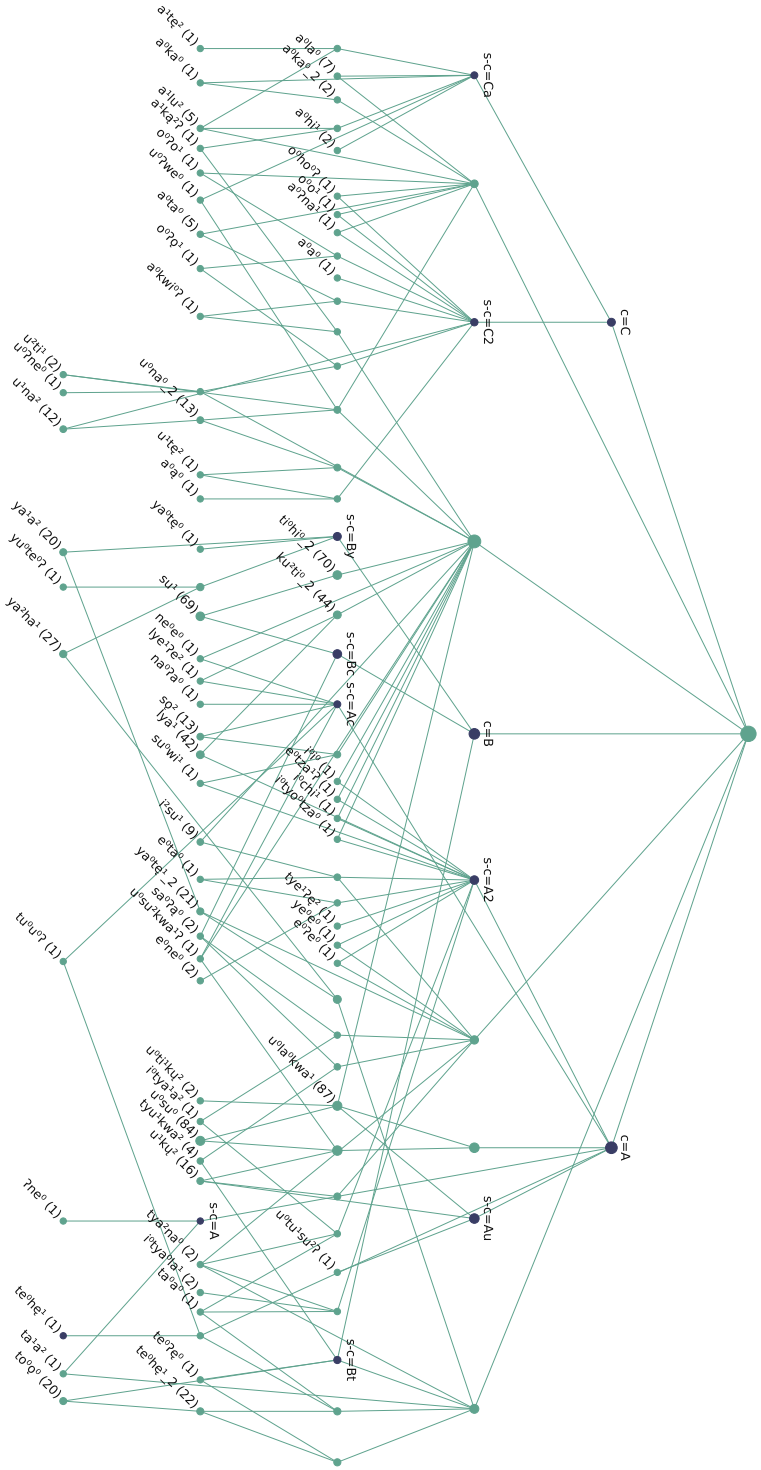


FIGURE 6.14 – Hiérarchie simplifiée des verbes du zenzontepec, comparée à celle de Campbell (2014).

(x) POT⇌PROG :ku ⇌ nte ⟹ A (46.11%)

(xi) CPL⇌POT :n_a ⇌ _u ⟹ A (44.56%)

(xii) CPL⇌POT :n_u ⇌ _a , CPL⇌PROG :ku ⇌ cha ⟹ Ca (38.89%)

(xiii) CPL⇌PROG :ϵ ⇌ ncha ⟹ Ca (33.33%)

(xiv) CPL⇌POT :ϵ ⇌ ka ⟹ Ca (33.33%)

(xv) CPL⇌HAB :ya ⇌ nti ⟹ C2 (20.69%)

Parmi les classes de premier niveau, seule la classe A apparaît dans ces implications. Toutes les sous-classes sauf Au et Ac y figurent. Nous pouvons donc en conclure qu'il n'existe pas de coïncidence exacte entre les concepts calculés sur la base des patrons et les classes de Campbell (2014), et les implications qui relient ces deux types de concepts ne sont jamais fondées sur une très grande proportion du concept cible, sauf pour Bc en (ii) et (iii). Ceci s'explique par le fait que Campbell (2011, 2014) fonde son système de classes sur celui de Kaufman (1989) pour le zapotec. Sa typologie est donc en grande partie historique. Nous ne trouvons pas de justification synchronique à cette classification dans la hiérarchie.

6.5 Conclusion

Dans cette section, nous avons montré qu'un modèle de classification flexionnelle fidèle aux données doit permettre l'évaluation de structures non-canoniques, ce qui n'est le cas ni des arbres ni des partitions de macroclasses. Nous avons proposé d'employer l'analyse formelle de concepts afin d'inférer une hiérarchie à héritage multiple dont chaque nœud constitue une classe flexionnelle. Les structures obtenues sont extrêmement fournies, et très éloignées du point canonique que constitue une partition de macroclasses. Cette propriété confirme

l'observation qu'une structure de macroclasses ne donne qu'une image extrêmement grossière d'un système flexionnel. Cependant les treillis obtenus sont également systématiquement beaucoup plus proche du canon que du maximum de non-canonicité imaginable. Cette observation, qui repose sur des mesures quantitatives, rejoint les observations de Carstairs (1987), Carstairs-McCarthy (1994) et Ackerman et Malouf (2015) : même dans les systèmes flexionnels qui semblent *a priori* très complexes, cette complexité est toujours très inférieure au maximum concevable. Enfin, nous avons décrit deux façons de simplifier les treillis de classes flexionnelles en hiérarchies de classes à héritage multiple, de façon à pouvoir les utiliser comme le point de départ d'analyses de morphologie fines. Ces analyses peuvent se fonder sur les implications fournies par la hiérarchie pour constituer de façon quantitative des hiérarchies comme celle que Kilani-Schoch et Dressler (2005) fournissent pour le français, ou comparer différentes propositions de classifications.

Conclusion

Tableau des résultats

Première partie : Alternances flexionnelles

La première partie de cette thèse a étudié les relations entre mots au sein des paradigmes de flexion. Nous avons représentés ces relations par des patrons d'alternances, qui rendent compte des contrastes formels entre les mots. Ce faisant, nous avons renoncé à une segmentation affixale, qui opère globalement sur l'ensemble du paradigme, en faveur de segmentations locales à chaque paire de cases. Ce changement d'échelle nous a permis de rendre compte de très nombreux points de similarité entre les comportements flexionnels des lexèmes, qui seraient ignorés par une analyse du type radical et exposant.

Le chapitre 2 présente un algorithme permettant d'inférer les patrons d'alternance efficacement sur de larges lexiques, sans préjuger du type d'exponence qu'ils présentent. Nous avons appliqué cet algorithme aux huit lexiques étudiés dans cette thèse. Suivant Albright et Hayes (1999, 2002), Albright et Hayes (2003) et Albright et Hayes (2006), les patrons se décomposent en deux séquences alternantes et un contexte d'application, et s'écrivent à la façon des règles phonologiques de Chomsky et Halle (1968). Le patron (83) présente une alternance fréquente entre le présent et le passé en anglais.

(83) $\epsilon \rightleftharpoons \text{ɪd} \;/\; \text{X+} \begin{bmatrix} \text{+dental} \\ \text{+anterior} \\ \text{-nasal} \\ \text{-continuant} \end{bmatrix} \_$

Le *Minimal Generalization Learner* de Albright et Hayes (1999) infère des patrons unidirectionnels. Il s'appuie sur une heuristique qui recherche un changement unique préférentiel-

lement suffixal, sinon préfixal, sinon interne. Il retient l'ensemble des généralisations intermédiaires. Ces propriétés posent problème pour nos besoins : en effet l'implémentation d'un biais d'alignement ne suffit pas à rendre compte de la plupart des alternances flexionnelles, qui peuvent présenter de l'alternance affixale et des alternances de base (russe, français, portugais), ou des alternances discontinues (arabe). L'usage d'une telle heuristique nuirait par ailleurs aux comparaisons typologiques. De plus, dans chaque lexique, nous nous sommes intéressés à l'ensemble des contrastes morphologiques présentés par chaque lexème. Nous avons calculé donc un ensemble de patrons d'alternances pour chaque paire de case au sein du paradigme, c'est à dire entre 6 (pour les 4 cases de paradigme du chatino de Zenzontepec) et 5886 (pour les 109 cases de paradigmes de l'arabe). Il est beaucoup trop coûteux en mémoire de conserver l'ensemble des généralisations intermédiaires dans ce contexte.

Nous avons proposé un algorithme qui n'est pas typologiquement biaisé en faveur des systèmes suffixaux, qui peut identifier des alternances de nature variées, et qui permet de calculer efficacement le très grand nombre de patrons d'alternance requis. Cet algorithme procède en trois étapes, répétées indépendamment pour chaque paire de case de paradigme :

(i) Aligner localement chaque paire de formes en minimisant une distance d'édition, et en déduire un patron d'alternance spécifique à cette paire de forme. Lorsqu'il existe plusieurs alignement optimaux, conserver l'ensemble de tous les patrons qui en découlent.

(ii) Fusionner tous les patrons qui partagent une même alternance structurelle, en exprimant leur contexte par une généralisation phonologique.

(iii) Scorer l'ensemble des patrons par le nombre de paires de formes qu'ils dérivent correctement. Pour chaque paire de forme, choisir parmi les patrons qui les dérivent correctement celui dont le score est le plus haut.

Cette procédure permet de maximiser les points de similarités entre lexèmes. L'étape de généralisation permet également d'identifier des alternances qui coïncident avec des opérations phonologiques régulières, comme une palatalisation ou un allongement. Nous avons présenté deux types de distances d'éditions susceptibles d'être employées en (i) : d'une part les distances de Levenshtein, d'autre par des distances pondérées par la similarité phonologique. Nous avons

évalué la qualité des généralisations trouvées au moyen d'une tâche de prédiction, en comparant l'alignement par distances d'édition aux alignement fixes à droite (changement préfixal) et à gauche (changement suffixal), ainsi qu'à la stratégie de Albright et Hayes (2002). L'évaluation montre que notre stratégie d'alignement produit les meilleurs résultats dans la plupart des langues, et des résultats très proche des meilleurs pour les autres langues. Nous n'avons pas trouvé différence de performance forte entre les deux stratégies d'alignement automatique. Nous avons conclu que notre algorithme produit en général de bonnes généralisation à travers des langues présentant des alternances variées.

L'évaluation produit des exactitudes inférieures à 80% pour le russe, le chatino et le navajo. Nous avons remarqué que ces trois systèmes flexionnels sont généralement décrits comme la conjonction de deux systèmes flexionnels : en russe, le système accentuel est distinct du système segmental, en chatino le système tonal se distingue du système segmental, et en navajo, le noyau verbal est formé de deux bases qui varient chacune de façon distincte. Nous avons proposé de rendre compte de chacun de ces lexiques flexionnels par deux représentation parallèles mais séparées, chacune représentative d'un des deux sous-systèmes. L'évaluation montre que cette manipulation améliore fortement l'évaluation sur chaque sous-système isolé, et que la prédiction jointe sur les paires de sous-systèmes améliore encore la prédiction.

Le chapitre 3 porte sur l'évaluation des relations implicatives au sein des paradigmes. Albright et Hayes (2002) emploient les patrons d'alternance pour simuler la prédiction d'un locuteur qui, connaissant une forme de paradigme, tente d'en prédire une autre. Cette tâche est cruciale pour les locuteurs, qui ne sont jamais exposés à l'ensemble des formes d'un lexème. (Ackerman, Blevins et Malouf 2009) nomment le problème résultant Problème de remplissage des cases de paradigme (*Paradigm Cell Filling Problem*).

Ackerman, Blevins et Malouf (2009) proposent de quantifier la difficulté du PCFP dans les paradigmes en se fondant sur une mesure de théorie de l'information. Ils emploient pour cela une entropie conditionnelle : connaissant la distribution d'affixes pour une case, quelle incertitude demeure pour deviner les affixes d'une autre case ? Ils nomment entropie paradigmatique la moyenne de ces entropies à travers l'ensemble du paradigme. La principale faiblesse de leur méthodologie est qu'elle se repose sur des segmentation affixales fournies par des grammaires

descriptives. Ces descriptions ne fournissent pas d'information de fréquence de type, sous-estiment à la fois la tâche des locuteurs, en masquant certaines opacités, et les informations disponibles, en ignorant toute une partie des mots (Bonami 2014 ; Bonami et Boyé 2014). Bonami et Boyé (2014) et Bonami et Luís (2014) proposent une mesure plus exacte du PCFP qui s'appuie sur les patrons d'alternance, et qu'ils nomment entropie implicative. Bonami et Beniamine (2016) étendent la formulation de l'entropie implicative aux prédictions depuis plusieurs formes, car il est rare que les locuteurs n'aient qu'une unique forme à leur disposition pour résoudre le PCFP. Dans le chapitre 3, nous avons également montré comment calculer l'entropie implicative des systèmes bipartites. En nous appuyant sur les résultats du chapitre 2, nous avons appliqué ces méthodologies à l'ensemble des huit lexiques étudiés dans cette thèse.

La principale conclusion de Ackerman, Blevins et Malouf (2009) est la conjecture d'entropie basse : à travers l'ensemble des systèmes dont ils évaluent l'entropie paradigmatique, l'entropie est uniformément basse. L'évaluation de l'entropie implicative sur les lexiques verbaux du français, du portugais, du chatino, de l'arabe, du navajo et le lexique nominal du russe confirment cette conjecture. Qui plus est, conformément aux résultats de Bonami et Beniamine (2016), l'entropie diminue de façon importante avec l'ajout de prédicteurs. La prédiction jointe est donc très utile dans le cadre du PCFP.

Nous avons observé que les entropies hautes s'organisent par case : dans tous les systèmes, nous avons trouvé des cases qui sont de mauvais prédicteurs. De façon intéressantes, ces cases sont en général facile à prédire. On peut penser qu'il s'agit des cases peu fréquentes, que les locuteurs sont parfois amenés à prédire, mais sur lesquelles ils ne fondent pas souvent leurs prédictions.

Nous avons remarqué que les entropies implicatives sont plus basses dans les systèmes présentant un plus grand nombre de cases de paradigme, et plus haute dans les petits systèmes. L'anglais, pour lequel nous avons surdifférencié les cases de paradigmes en raison du verbe TO BE, fait seul exception à cette tendance. Cette relation entre taille de paradigme et prédictibilité nous a mené à formuler une conjecture de stabilité de l'entropie implicative. Les cases de paradigmes sont un partitionnement construit des contextes syntaxiques en fonction des variations formelles. Un système présentant moins de cases ne présente pas moins de contextes syn-

taxiques, mais seulement moins de contrastes formels à travers ces contextes. En conséquence, en n'évaluant le PCFP qu'à travers les cases de paradigmes, nous ignorons toutes les prédictions syncrétiques, lorsqu'un locuteur connaissant la forme dans un contexte syntaxique emploie une forme identique dans un contexte syntaxique distinct. Si nous considérions non pas les cases de paradigmes mais les contextes d'emplois syntaxiques, l'entropie implicative moyenne pondérée à travers les langues devrait être relativement stable. Pour confirmer cette conjecture, il nous faudrait d'une part un échantillon de lexiques beaucoup plus grand,permettant de confirmer la relation, et d'autre part des données concernant les fréquences de cases de paradigmes dans chaque langue.

Deuxième partie : La structure de similarité des systèmes flexionnels

La deuxième partie de cette thèse a étudié la structure de similarité des systèmes flexionnels, et les modèles de classifications susceptibles d'en rendre compte. Nous avons caractérisé chaque lexème par l'ensemble des patrons d'alternances qu'il instancie. Chaque patron d'alternance représente donc potentiellement un point de similarité entre lexèmes.

Le chapitre 4 étudie les microclasses, c'est à dire les classes de lexèmes qui partagent exactement les mêmes patrons d'alternances. Les microclasses sont entièrement déterminées par l'inventaire de patrons. Empiriquement, nous avons montré qu'elles sont toujours nombreuses (de 52 en chatino du Zenzontepec à 367 en arabe). La distribution de leur fréquence de type (nombre de lexème par microclasse) suit une loi de puissance : on trouve quelques très grosses classes, et de nombreuses très petites classes. Nous avons vu qu'en arabe, en russe, en chatino de Yaitepec, pour les tons du chatino de Zenzontepec, pour les accents du russe, et en navajo, on trouve moins de cinq lexèmes par classes en moyenne, indiquant le faible pouvoir descriptif d'une classification en microclasses. En chatino, il se peut que cela soit lié à la petite taille des lexiques, qui comportent peut-être principalement des formes exemplaires. En général, à travers tous les lexiques étudiés, les microclasses sont trop nombreuses, et reflètent des variations trop petites, pour constituer une base intéressante pour un usage pédagogique, contrairement à la notion traditionnelle de classe flexionnelle. Nous avons proposé trois explorations de la structure de similarité des microclasses : le dessin de graphes de similarités saillantes, suivant

Sims et Parker (2016), l'inférence d'arbres d'héritage par défaut, et enfin le la classification ascendante hiérarchique fondée sur les distances entre classes.

Sims et Parker (2016) représentent les systèmes de classes flexionnelles par des graphes au sein desquelles chaque classe est un sommet, et les arcs représentent des partages de traits entre classes. Nous avons produit des graphes similaires fondés sur nos microclasses, et où la similarité représente les patrons communs entre microclasses. Nous avons trouvé trois types de réseaux : les verbes du zenzontepec forment un réseau très peu connecté, aussi bien du point de vue des alternances segmentales que tonales. Nous avons regroupé dans une seconde catégorie les systèmes dans lesquels la représentation en réseau laisse deviner des sous-réseaux bien connectés : il s'agit de l'arabe, du portugais, des segments du russe, et des tons du chatino de Yaitepec. Ces systèmes devraient constituer de bons candidats à l'analyse en macroclasses, puisqu'ils semblent bien présenter des groupements de microclasses distincts les uns des autres. Enfin, les systèmes de l'anglais, du français, et du navajo présentent des connections denses qui ne semblent pas se structurer en sous-classes.

Nous avons discuté de l'opportunité de former des arbres d'héritage par défaut pour représenter la parenté entre les classes. Les arbres d'héritage par défaut sont attirants principalement en vertu de deux propriétés : ils rendent compte intuitivement d'une notion de régularité (les classes les plus hautes dans la hiérarchie sont plus régulières), et ils fournissent une représentation très compacte d'un inventaire de microclasses. Nous avons proposé un algorithme qui permet de générer un arbre de microclasses par défaut, en privilégiant les microclasses les plus grandes, et en reliant les classes de façon à minimiser les distances. Nous avons défini la distance entre deux microclasses par le nombre de paires de paradigmes pour lesquelles ces classes instancient des patrons distincts. Nous avons illustré cet algorithme sur les noms du russe, et montré que la compacité de cette structure l'empêche de refléter les réseaux de similarité de façon intéressante dans une perspective abstractive.

Nous nous sommes alors tournés vers l'inférence d'un arbre binaire sous la forme d'un dendrogramme au moyen de la méthode UPGMA. L'algorithme ascendant réunit itérativement deux à deux les classes les plus similaires. Chaque fusion produit un niveau dans l'arborescence. L'arbre obtenu fournit un ordre sur les feuilles (les microclasses), ce qui facilite la visualisation

des cartes thermiques qui représentent les matrices de distances entre microclasses. Nous pouvons ainsi observer à l'œil nu l'organisation de similarité entre les microclasses, d'une façon plus subtile que l'image fournie par les graphes de microclasses. L'analyse des microclasses en arbre elle-même a cependant deux propriétés peu satisfaisantes : d'une part, elle opère des coupes drastiques dans le réseau de similarité (toutes les branches sont binaires, aucun héritage multiple n'est possible), et d'autre part, ces arbres ne peuvent être comparés aux analyses en macroclasses fournies par les linguistes ou la tradition grammaticale.

Le chapitre 5 décrit un algorithme permettant de choisir une partition de macroclasses en s'appuyant sur les vecteurs de patrons d'alternance. Nous nous sommes appuyé sur la longueur de description pour choisir une partition qui reflète au mieux les régularités des les données. Nous avons proposé une description probabiliste d'une partition de classes flexionnelles, qui rend compte de l'assignation des lexèmes à chaque classe, de la distribution des patrons dans chaque classe, et de la désambiguïsation nécessaire au sein des classes, lorsqu'il existe plusieurs patrons concurrents. Nous pouvons alors définir la longueur de description d'une partition de classes flexionnelles, et chercher une partition de longueur de description minimale. L'algorithme de recherche procède de façon ascendante. Il commence par un inventaire de microclasses, puis fusionne itérativement la paire de classes qui fournit la plus grande réduction en longueur de description, tant qu'il existe une telle paire. Cet algorithme n'est pas forcé de trouver une généralisation intermédiaires entre les microclasses et le système entier : il est tout à fait concevable qu'un système de microclasses, ou qu'un système formé d'une seule classe constituent des descriptions optimales d'un point de vue de la longueur de description. Nous avons appliqué cet algorithme aux systèmes pour lesquels nous disposions de descriptions en macroclasses provenant de la littérature linguistique ou des grammaires traditionnelles : le français, le portugais, les segments du russe et du chatino de Zenzontepec. Dans les quatre cas, nous avons bien trouvé des macroclasses distinctes des microclasses et de l'ensemble du système. En chatino de Zenzontepec, nous n'en avons trouvé que deux, en portugais, cinq, en français, six, en russe, neuf. Ces classes fournissent donc une généralisation importante sur les microclasses, beaucoup plus nombreuses. L'observation qualitative des classes obtenues révèle leur proximité avec les classes proposées dans la littérature. L'analyse automatisée fait ressor-

tir des régularités au sein des groupes moins fréquents et des similarités qui sont usuellement ignorées. L'évaluation quantitative confirme que les classifications obtenues sont très similaires aux classifications traditionnelles. Elles présentent toujours une longueur de description plus basse. À l'exception du chatino de Zenzontepec, elles sont meilleures en termes de scores de silhouette, et sont plus informatives sur les microclasses. Cependant, dans l'ensemble, les macroclasses ont toujours des scores de silhouette peu satisfaisants, et ne sont que très partiellement informatives. Nous en avons conclu que la stratégie proposée est efficace pour choisir une partition en macroclasse qui soit aussi bonne que possible, mais qu'en soi, une partition de macroclasse est généralement un modèle très imprécis de la structure des systèmes flexionnels, et ce y compris dans les systèmes comme le portugais qui se prêtent le mieux à ce type d'analyse.

Dans le chapitre 6, nous avons proposé de considérer comme classe flexionnelle tout groupe de lexème partageant maximalement un ensemble de patrons. Nous avons employé à cette fin l'analyse formelle de concepts. Celle-ci définit un concept formel comme une paire formée d'un ensemble de propriétés (nos patrons) et d'un ensemble d'objets (nos lexèmes), tels que tous les objets ont en commun toutes ces propriétés, mais aucune autre, et tels que ces propriétés sont partagées par exactement ces lexèmes, et aucun autre. Nous avons proposé que tout concept reliant des lexèmes et des patrons d'alternance constitue un point de similarité susceptible d'être remarqué ou utilisé par un locuteur. Chaque concept représente donc une classe flexionnelle, qui peut être d'une granularité variable. Les concepts peuvent être ordonnés entre eux par inclusion : un concept est supérieur à un autre si l'ensemble de ses objets contient l'ensemble des objets de l'autre. En raison de la définition des concepts, cela signifie également que l'ensemble des propriétés d'un concept supérieur à un autre est un sous-ensemble des propriétés de cet autre concept. En somme, cet ordre repose sur la généralité : les concepts les plus généraux (les plus grands) concernent beaucoup de lexèmes mais peu de propriétés, tandis que les concepts les plus spécifiques concernent plus de propriétés mais moins de lexèmes. L'ensemble des classes ordonnées par généralité forme un treillis au sens mathématique, qui peut se lire comme une hiérarchie à héritage multiple monotone. Cette représentation des système de classes nous a permis d'évaluer la distance entre les systèmes observés et les systèmes cano-

niques au sens de Corbett (2009). L'observation quantitative des treillis montre que l'héritage multiple, ignoré par une modélisation en macroclasses ou en arbres, constitue en fait le cas général. Nous avons montré que sur les systèmes étudiés, les treillis obtenus sont immenses, et très éloignés de l'idéal canonique d'une partition (Corbett 2009). Ils sont cependant encore beaucoup plus éloignés du maximum théorique de non-canonicité que du point de canonicité parfaite. Les treillis peuvent également être employés pour une exploration qualitative manuelle, à condition d'observer un sous-ensemble des concepts. Nous avons proposé deux façons complémentaires de filtrer les concepts, et avons présenté quelques exemples d'observations qualitatives possible sur ces structures qui illustrent leur utilisé pour la description des systèmes flexionnels.

Bilan général

Deux questions sous-tendent le travail mené dans cette thèse : d'une part, peut-on inférer des exposants automatiquement ? D'autre part, qu'est-ce qu'une classe flexionnelle ? Dans cette section, nous faisons le bilan de ce que nous avons appris quant à ces deux questions.

Les patrons d'alternance nous ont fourni un moyen opérationnel pour caractériser le comportement flexionnel formel des lexèmes sans avoir à nous prononcer sur les épineuses questions de segmentation en exposants et radicaux. Parce que le nombre de cases envisagées est très important, et parce qu'ils fournissent des informations très locales, les patrons d'alternance demeurent moins intuitifs qu'une segmentation affixale des données. Tandis qu'une segmentation en exposants et radicaux fournit une analyse plus courte que les données, la combinatoire du nombre de cases fait que les descriptions en patrons d'alternance sont plus grandes que les lexiques sur lesquels ils se fondent. Nous avons souvent été tenté de chercher des moyens de revenir à une segmentation globale du paradigme, fournissant une analyse compacte et intuitive des données. Nous avons envisagé dans ce but de nombreux algorithmes, fondés ou non sur les patrons, leurs contextes phonologiques, sur des alignements phonologiquement motivés entre les formes d'une case de paradigme, etc. Il existe de nombreuses façons d'obtenir une segmentation des formes qui respecte un certain principe ou une certaine mesure. Cependant, il existe

de nombreuses telles segmentations, et il reste difficile de les évaluer d'une façon satisfaisante. En général, même lorsqu'un algorithme fournit un résultat intuitif pour une langue donnée, ce n'est pas le cas pour les autres. Rien n'indique qu'il existe une *bonne* solution unique dans chaque lexique, ni qu'une bonne solution puisse être trouvée par une unique méthode dans n'importe quel système flexionnel. Par ailleurs, comme nous l'avons montré aux chapitres 1 et 3, les segmentations en radicaux et exposants fournissent des descriptions plus pauvres que les patrons d'alternances. Enfin il ne fait aucun doute que les locuteurs peuvent fonder leurs inférences sur les formes entières, y compris le matériel phonétique qui pourrait être attribué aux radicaux. Les impasses rencontrées à répétition dans la quête d'une segmentation idéale des données nous amène à conjecturer que celle-ci n'existe pas, et que les patrons constituent bien une meilleure représentation des contrastes flexionnels.

La définition des microclasses, fondées sur l'identité, est séduisante, car elle fournit une façon très simple de partitionner les lexèmes lorsque l'on connaît les patrons d'alternance. Cependant, en pratique, les microclasses sont beaucoup trop petites et nombreuses pour fournir le type d'intuitions qui ressortent généralement des descriptions en classes flexionnelles. Quoi qu'il soit possible d'inférer automatiquement des macroclasses, celles-ci ne fournissent qu'une vue extrêmement pauvre des relations de proximité entre les microclasses, et sont beaucoup moins informatives que les microclasses. Si l'on prend au sérieux l'entreprise de représenter l'ensemble des points de similarités intéressants au sein des paradigmes, on obtient les treillis de classes flexionnelles extrêmement fournis construits au chapitre 6. En somme, il est possible d'observer précisément la structuration des microclasses, mais les structures obtenues ne coïncident jamais avec des macroclasses, et elles sont plus complexe que l'inventaire des microclasses seul. En général, à travers les langues étudiées, nous ne trouvons pas de sous-classification spécifique qui mérite d'être mise en avant au détriment d'autres relations de similarité au sein des systèmes flexionnels. Nous en concluons que toute classification plus ambitieuse que le treillis complet des relations de similarité est une construction théorique, et non un objet pouvant être découvert. Il est possible de concevoir des classifications flexionnelles qui rendent compte de divers aspects des systèmes, mais il n'existe pas *les classes flexionnelles* d'un système spécifique. En particulier, nos résultats remettent en cause l'idée selon laquelle

l'observation des classes flexionnelles au sein d'un système flexionnel peut fournir une vue économique et intuitive de ce système.

Directions de recherche

À l'issue de cette thèse, nous discernons plusieurs directions de recherche : employer les outils développés dans le cadre de cette thèse pour une étude typologie fondée sur un éventail de langues représentatif ; explorer la notion d'exponence de façon non supervisée en se fondant sur des tâches réalistes pour les locuteurs ; envisager de partir de données qui n'ont pas été discrétisées.

L'étude menée dans cette thèse s'est fondée sur huit lexiques. L'échantillon est petit, et n'est pas représentatif d'une grande variété de familles de langues. Maintenant que les outils computationnels sont prêts, il serait intéressant de mener une étude similaire sur un échantillon de systèmes flexionnels appartenant à des langues choisies pour leur représentativité. Il existe deux obstacles à cette entreprise : d'une part, puisque nous avons besoin, en entrée, de grands lexiques informatisés, il existe un biais en faveur des langues écrites bien documentées. L'obtention de données utilisables par nos programmes pour les langues qui n'ont pas de tradition écrite ou qui sont peu documentées requerrait de constituer les lexiques nous-même à cet effet. D'autre part, nous avons constaté qu'il est impossible de se passer de l'évaluation qualitative des données par un expert de la langue. Celle-ci est d'autant plus nécessaire pour une étude approfondie de chaque système flexionnel. En conséquence, une telle étude typologique constitue un projet de plus grande envergure, qu'il n'est pas imaginable de mener dans le cadre d'une thèse.

Nous avons vu qu'il est vain de tenter de résoudre la question de l'exponence par une segmentation catégorique des formes en deux sous chaînes, le radical et l'exposant. La question demeure cependant de savoir sur quel matériel phonétique les locuteurs fondent leurs inférences flexionnelles, et quelles parties des mots formes sont plus ou moins informatives pour les locuteurs. Nous pensons que la reconnaissance de l'exponence est utile aux locuteurs pour résoudre deux problèmes : d'une part le PCFP, étudié dans cette thèse, et d'autre part

un problème symmétrique que nous nommerons PCRP, ou *Paradigm Cell Recognition Problem* (problème de reconnaissance des cases de paradigme) : Étant donné un mot-forme, qu'est-ce qui, dans sa forme, permet aux locuteurs d'identifier ses propriétés morpho-syntaxiques ? Bien évidemment, le contexte syntaxique joue toujours un rôle important dans la désambiguïsation des valeurs flexionnelles d'un mot forme. Cependant, les mots étant marqués d'un point de vue flexionnel, ils fournissent également des informations au locuteur. Dans un article récent, Malouf (2017) a entraîné un réseau de neurones récurrent à résoudre le PCFP. Il obtient des performances extrêmement bonnes (meilleures que celle d'un locuteur natif). L'avancement récent des méthodes permettant de comprendre l'organisation des réseaux de neurones pourrait permettre d'éclairer la façon dont un tel réseau de neurones *regarde* les mots-formes qu'il reçoit en entrée. Nous préparons un article qui propose de faire précisément cela pour le PCRP (Beniamine & Coavoux, en préparation). L'idée est d'entraîner un réseau de neurone récurrent à deviner la case de paradigme d'un mot forme, sans autre information que le lexème concerné et les phonèmes qui constituent le mot-forme. Nous utilisons un mécanisme d'*attention* (Bahdanau, Cho et Bengio 2014) pour observer les caractères (phonèmes) qui sont les plus utiles au réseau de neurones dans sa prédiction. Par ailleurs, signe que la question de l'exponence et celle des classes flexionnelles sont intimement liées, il est possible d'observer si ces systèmes organisent les lexèmes en classes flexionnelles. Qu'ils soient entraînés à simuler le PCFP ou le PCRP, ces réseaux de neurones prennent en entrée un identifiant de lexème. Le réseau de neurone est alors libre d'organiser des représentations vectorielles (*embeddings*) correspondant à ces lexèmes de façon à résoudre la tâche au mieux. L'étude de la structure de ces représentations peut fonder une étude non supervisée des classes flexionnelles. Il serait par exemple intéressant de savoir comment se distribuent les lexèmes dans l'espace vectoriel, et si la résolution du PCFP et du PCRP mènent à des organisations similaires.

Enfin, les méthodes employées dans cette thèse se fondent sur des données doublement discrétisées : les mots formes en notation phonétique constituent bien sur une simplification importante par rapport aux signaux auditifs que traitent véritablement les locuteurs. De plus, l'organisation de ces mots formes en paradigmes répond à des arbitrages donc nous avons discuté au chapitre 3. Il serait intéressant d'explorer soit des analyses fondées sur des don-

nées moins discrétisées, soit des discrétisations opérées automatiquement. Se débarasser de la phonémisation semble encore difficile au vu des capacités techniques actuelles, et requerrait une quantité d'enregistrements audio annotés manuellement qui n'existe probablement pour aucune langue. Il semble cependant déjà réaliste d'envisager de renoncer aux paradigmes préstructurés. Lee et Goldsmith (2013) ont ainsi tenté d'inférer des classes flexionnelles à partir d'ensembles de formes non structurés en paradigmes. Leur tentative échoue en partie en raison de problèmes de modélisation, et en partie parce qu'ils espéraient retrouver une structure de paradigme en se fondant exclusivement sur les propriétés formelles des mots fléchis. Nous pensons que l'inférence automatique de paradigme est possible, sous la forme d'une classification automatique (non supervisée) en corpus des contextes syntaxiques associés à des contrastes de formes.

Annexe

Annexe A

Les données

Cette thèse est consacrée à l'étude quantitative et automatisée de systèmes flexionnels. Pour chaque système étudié, nous prenons en entrée un lexique en notation phonémique accompagné d'une spécification des traits distinctifs caractérisant les phonèmes utilisés. Pour certains systèmes, nous utilisons les lexiques tels qu'ils sont distribués, avec très peu de post-traitements. Pour d'autres, un important travail de sélection, formatage, phonétisation, et organisation ont été nécessaires. Cette annexe documente la source de chaque lexique, ainsi que les choix opérés dans leur adaptation et leur phonétisation. Le tableau A.1 synthétise la taille de chaque paradigme, et le tableau A.2 présente la source du lexique, et la principale source pour la constitution de la spécification phonologique. Les traits employés s'inspirent toujours au moins en partie de la description distribuée par Hayes (2012).

Langue	Lexèmes	Cases de paradigme
Français	5249	51
Anglais	6064	8
Chatino de Zenzontepec	392	4
Chatino de Yaitepec	324	12
Portuguais européen	1996	69
Arabe	1018	109
Russe	1539	13
Navajo	2157	70

Tableau A.1 – Taille des lexiques flexionnels étudiés.

Langue	Lexique	Traits distinctifs
Français	Bonami, Caron et Plancq (2014)	Dell (1973)
Anglais	Baayen, Piepenbrock et Gulikers (1995)	Halle et Clements (1983)
Chatino de Zenzontepec	Feist et Palancar (2015)	Campbell (2014)
Chatino de Yaitepec	Feist et Palancar (2015)	Rasch (2002),Hayes (2012)
Portuguais Européen	Veiga, Candeias et Perdigão (2013)	Bonami et Luís (2014)
Arabe	Kirov et al. (2016)	Hayes (2012)
Russe	Brown (1998)	Hayes (2012)
Navajo	Young et Morgan (1987)	McDonough (2003), Hayes (2012)

Tableau A.2 – Sources pour les lexiques flexionnels étudiés.

A.1 Verbes du français

Notre lexique du français est constitué des verbes de Flexique (Bonami, Caron et Plancq 2014), qui est lui-même fondé sur Lexique (New, Pallier et al. 2001). Dans Flexique, les transcriptions phonétiques ont été corrigées manuellement et les paradigmes incomplets de Lexique ont été remplis semi-automatiquement. L'étude de la structure flexionnelles des paradigmes qui y figure nous a permis de contribuer au lexique Flexique en corrigeant des erreurs. Ces correctifs ont été intégrés à la ressource.

Le lexique est organisé en table de paradigmes, comme illustré sur un petit extrait dans le tableau A.3. Les lexèmes figurent en ligne et sont identifiés par une forme de citation. Les cases de paradigme sont indiquées en titres de colonnes. Les cases sont remplies avec les formes correspondant au lexème et à la case concernés. Certains lexèmes sont défectifs : la case porte alors la mention `#DEF#`.

Les cases de paradigmes sont au nombre de 51 et s'organisent en sept temps finis ayant chacun six personnes, ainsi que l'impératif qui présente seulement trois personnes et six cases non finies, comme indiqué dans le tableau A.4.

La définition des phonèmes se fonde sur Dell (1973). En raison de variations régionales, les données comportent une annotation inconsistante des voyelles moyennes. Nous neutralisons

	SBJV.3SG	PST.1SG	PRS.1SG	IPFV.1PL	...
lexème					...
dorloter	dɔʁlɔt	dɔʁlOtE	dɔʁlɔt	dɔʁlOtjɔ̃	...
déférer	dEfɛʁ	dEfEʁE	dEfɛʁ	dEfEʁjɔ̃	...
retraverser	ʁətʁavɛʁs	ʁətʁavɛʁsE	ʁətʁavɛʁs	ʁətʁavɛʁsjɔ̃	...
rabrouer	ʁabʁu	ʁabʁuE	ʁabʁu	ʁabʁujɔ̃	...
retrouver	ʁətʁuv	ʁətʁuvE	ʁətʁuv	ʁətʁuvjɔ̃	...
clore	kloz	#DEF#	klo	#DEF#	...
choir	#DEF#	ʃy	ʃwa	#DEF#	...
...	...	...	...	...	...

TABLEAU A.3 – Extrait de sept lexèmes et quatre cases de paradigmes issus du lexique du français.

	présent	passé simple	imparfait	futur	conditionnel	subjonctif présent	subjonctif passé	impératif
1SG	PRS.1SG	PST.1SG	IPFV.1SG	FUT.1SG	COND.1SG	SBJV.1SG	PST.SBJV.1SG	—
2SG	PRS.2SG	PST.2SG	IPFV.2SG	FUT.2SG	COND.2SG	SBJV.2SG	PST.SBJV.2SG	IMP.2SG
3SG	PRS.3SG	PST.3SG	IPFV.3SG	FUT.3SG	COND.3SG	SBJV.3SG	PST.SBJV.3SG	—
1PL	PRS.1PL	PST.1PL	IPFV.1PL	FUT.1PL	COND.1PL	SBJV.1PL	PST.SBJV.1PL	IMP.1PL
2PL	PRS.2PL	PST.2PL	IPFV.2PL	FUT.2PL	COND.2PL	SBJV.2PL	PST.SBJV.2PL	IMP.2PL
3PL	PRS.3PL	PST.3PL	IPFV.3PL	FUT.3PL	COND.3PL	SBJV.3PL	PST.SBJV.3PL	—

infinitif	participe présent	participe passé
INF	PRS.PTCP	PST.PTCP.M.SG, PST.PTCP.M.PL, PST.PTCP.F.SG, PST.PTCP.F.PL

TABLEAU A.4 – Organisation des cases de paradigmes le lexique du français.

donc les oppositions entre /e/ et /ɛ/, /ø/ et /œ/, /o/ et /ɔ/, que nous notons respectivement par les symboles /E/, /Ø/ et /O/. Nous ajoutons à l'inventaire de Dell (1973) la voyelle moyenne inférieure antérieure arrondie nasalisée /œ̃/, le schwa /ə/ ainsi qu'un trait [±antérieur] qui concerne les voyelles. Nous représentons le schwa comme une voyelle non arrière, non antérieure et non arrondie. Les tableaux A.5 et A.5 présentent la décomposition en traits obtenue. Les cases vides indiquent qu'un trait n'est pas pertinent pour un phonème donné.

	sonant	syllabique	consonantique	continu	nasal	haut	bas	arrière	arrondi	antérieur	coronal	voisé	rel.ret.
p	-	-	+	-	-	-		-		+	-	-	-
b	-	-	+	-	-	-		-		+	-	+	-
t	-	-	+	-	-	-		-		+	+	-	-
d	-	-	+	-	-	-		-		+	+	+	-
k	-	-	+	-	-	+		+		-	-	-	-
g	-	-	+	-	-	+		+		-	-	+	-
f	-	-	+	+	-	-		-		+	-	-	+
v	-	-	+	+	-	-		-		+	-	+	+
s	-	-	+	+	-	-		-		+	+	-	+
z	-	-	+	+	-	-		-		+	+	+	+
ʃ	-	-	+	+	-	+		-		-	+	-	+
ʒ	-	-	+	+	-	+		-		-	+	+	+
m	+	-	+	+	+	-		-		+	-	+	-
n	+	-	+	+	+	-		-		+	+	+	-
ɲ	+	-	+	+	+	+		-		-	-	+	-
ŋ	+	-	+	+	+	+		+		-	-	+	-
ʁ	+	-	+	+	-	-		+		-	-	+	+
l	+	-	+	+	-	-		-		+	+	+	+

Tableau A.5 – Traits distinctifs employés pour le français (consonnes).

	sonant	syllabique	consonantique	continu	nasal	haut	bas	arrière	arrondi	antérieur	coronal	voisé	rel.ret.
j	+	-	-	+	-	+	-	-	-	-	-	+	+
w	+	-	-	+	-	+	-	+	+	-	-	+	+
ɥ	+	-	-	+	-	+	-	-	+	-	-	+	+
i	+	+	-	+	-	+	-	-	-			+	+
y	+	+	-	+	-	+	-	-	+			+	+
u	+	+	-	+	-	+	-	+	+			+	+
E	+	+	-	+	-	-		-	-			+	+
e	+	+	-	+	-	-		-	-			+	+
ɛ	+	+	-	+	-	-		-	-			+	+
Ø	+	+	-	+	-	-		-	+			+	+
ə	+	+	-	+	-	-		-	+			+	+
ø	+	+	-	+	-	-		-	+			+	+
œ	+	+	-	+	-	-		-	+			+	+
O	+	+	-	+	-	-		+	+			+	+
o	+	+	-	+	-	-		+	+			+	+
ɔ	+	+	-	+	-	-		+	+			+	+
a	+	+	-	+	-	-	+	+	-			+	+
ɛ̃	+	+	-	+	+	-	+	-	-			+	+
œ̃	+	+	-	+	+	-	+	-	+			+	+
ɑ̃	+	+	-	+	+	-	+	+	-			+	+
ɔ̃	+	+	-	+	+	-	+	+	+			+	+

Tableau A.6 – Traits distinctifs employés pour le français (voyelles et glides).

A.2 Verbes du Portuguais (Portugal)

Le lexique verbal du portugais européen que nous utilisons se fonde sur le dictionnaire de prononciation de Veiga, Candeias et Perdigão (2013) [1]. La définition des traits distinctifs, proviennent de Bonami et Luís (2014). Le lexique de Veiga, Candeias et Perdigão (2013) présente exactement une forme par case de paradigme : aucun phénomène de surabondance (plus d'une forme par case) ni de défectivité (absence de forme dans une case) n'est pris en compte. Le tableau A.7 présente un petit extrait de quatre verbes et quatres cases de paradigmes sélectionnés aléatoirement. Le lexique que nous utilisons est le même que celui que nous avons employé dans dans Bonami et Beniamine (2016). Il comporte les 2000 verbes les plus fréquents du corpus journalistique CETEMPúblico (Santos et Rocha 2001). Après filtrage de verbes dupliqués, le lexique comporte 1996 entrées verbales.

	FutImpIndic3	PretMqpfIndic3	PresIndic4	PretPerfIndic4	...
RETRATAR	rətrɐtɛɾɐ	rətrɐtaɾɐ	rətrɐtemuʃ	rətrɐtamuʃ	...
REQUALIFICAR	rəkwɛlifikɛɾɐ	rəkwɛlifikaɾɐ	rəkwɛlifikemuʃ	rəkwɛlifikamuʃ	...
PENDER	pẽdəɾɐ	pẽdɛɾɐ	pẽdemuʃ	pẽdemuʃ	...
CORPORIZAR	kuɾpuɾizɛɾɐ	kuɾpuɾizaɾɐ	kuɾpuɾizemuʃ	kuɾpuɾizamuʃ	...
...	...	...	...	...	...

Tableau A.7 – Extrait de quatre lexèmes et quatre cases de paradigmes issus du lexique du portugais.

Les 69 cases de paradigme s'organisent en 14 combinaisons de modes et temps, dont 11 ont chacun six personnes. Le tableau A.8 synthétise cette organisation, ainsi que le nom de chaque case dans nos données.

La transcription des phonèmes est standardisée, et « correspond à une réalisation possible en contexte formel avec un débit relativement lent, ce qui maximise la quantité de voyelles réalisées » (Bonami et Luís 2014, p. 10). La décomposition des phonèmes en traits distinctifs

1. Des ressources plus récentes sont accessibles à l'addresse http://lsi.co.it.pt/spl/resources.html

	Indicatif				
	Futur	Présent	Imparfait	Plus que parfait	Parfait
1	FutImpIndic1	PresIndic1	PretImpIndic1	PretMqpfIndic1	PretPerfIndic1
2	FutImpIndic2	PresIndic2	PretImpIndic2	PretMqpfIndic2	PretPerfIndic2
3	FutImpIndic3	PresIndic3	PretImpIndic3	PretMqpfIndic3	PretPerfIndic3
4	FutImpIndic4	PresIndic4	PretImpIndic4	PretMqpfIndic4	PretPerfIndic4
5	FutImpIndic5	PresIndic5	PretImpIndic5	PretMqpfIndic5	PretPerfIndic5
6	FutImpIndic6	PresIndic6	PretImpIndic6	PretMqpfIndic6	PretPerfIndic6

	subjunctif		
	Futur	Présent	Imparfait
1	FutConj1	PresConj1	PretImpConj1
2	FutConj2	PresConj2	PretImpConj2
3	FutConj3	PresConj3	PretImpConj3
4	FutConj4	PresConj4	PretImpConj4
5	FutConj5	PresConj5	PretImpConj5
6	FutConj6	PresConj6	PretImpConj6

	Conditionnel	Impératif	Infinitif personnel
1	Condicional1	Imperativo1	InfinitPessoal1
2	Condicional2	Imperativo2	InfinitPessoal2
3	Condicional3	Imperativo3	InfinitPessoal3
4	Condicional4	Imperativo4	InfinitPessoal4
5	Condicional5	Imperativo5	InfinitPessoal5
6	Condicional6	Imperativo6	InfinitPessoal6

Participe passé	Gérondif	Infinitif
PartPasssm	Gerúndio	Infinitivo

Tableau A.8 – Organisation des cases de paradigmes le lexique du portugais.

provient également de Bonami et Luís (2014), qui se sont fondés sur l'analyse de Mateus et d'Andrade (2000). La transcription utilisée diffère de la description de Mateus et d'Andrade (2000) en trois points : les semi-voyelles sont notées par les voyelles hautes correspondantes, dont elles ne sont donc pas distinguées ; la voyelle centrale non basse est transcrite /ə/ plutôt que /ɨ/ ; enfin nos données ignorent l'accent. Les traits employés sont présentés dans les tableaux A.10 et A.9. Notons que dans ce fichier, tous les traits distinctifs ne sont pas binaires : les valeurs de certains traits s'échelonnes sur trois niveaux. Nous notons ici "-", "+" et "++".

	sonority	syllabic	consonantic	anterior	coronal	back	high	low	round	atr	nasal	lateral	continuant	voiced	strident
i	+	+	-	-	-	-	+	-	-	+	-	-	+	+	-
ɨ	+	+	-	-	-	+	+	-	-	+	-	-	+	+	-
e	+	+	-	-	-	-	-	-	-	+	-	-	+	+	-
ɛ	+	+	-	-	-	-	-	-	-	-	-	-	+	+	-
a	+	+	-	-	-	-	-	+	-	-	-	-	+	+	-
ə	+	+	-	-	-	+	-	-	-	-	-	-	+	+	-
ɐ	+	+	-	-	-	+	-	+	-	-	-	-	+	+	-
o	+	+	-	-	-	++	-	-	+	+	-	-	+	+	-
ɔ	+	+	-	-	-	++	-	-	+	-	-	-	+	+	-
u	+	+	-	-	-	++	+	-	+	+	-	-	+	+	-
ĩ	+	+	-	-	-	-	+	-	-	+	+	-	+	+	-
ɐ̃	+	+	-	-	-	+	-	+	-	-	+	-	+	+	-
ẽ	+	+	-	-	-	-	-	-	-	+	+	-	+	+	-
õ	+	+	-	-	-	++	-	-	+	+	+	-	+	+	-
ũ	+	+	-	-	-	++	+	-	+	+	+	-	+	+	-

Tableau A.9 – Traits distinctifs employés pour le portugais (voyelles et glides).

	sonority	syllabic	consonantic	anterior	coronal	back	high	low	round	ATR	nasal	lateral	continuant	voiced	strident
p	-	-	+	+	-	+	-	-	-	-	-	-	-	-	-
t	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-
k	-	-	+	-	-	++	+	-	-	-	-	-	-	-	-
b	-	-	+	+	-	+	-	-	-	-	-	-	-	+	-
d	-	-	+	+	+	-	-	-	-	-	-	-	-	+	-
g	-	-	+	-	-	++	+	-	-	-	-	-	-	+	-
f	-	-	+	+	-	+	-	-	-	-	-	-	+	-	+
s	-	-	+	+	+	-	-	-	-	-	-	-	+	-	+
ʃ	-	-	+	-	+	+	+	-	-	-	-	-	+	-	+
v	-	-	+	+	-	+	-	-	-	-	-	-	+	+	+
z	-	-	+	+	+	-	-	-	-	-	-	-	+	+	+
ʒ	-	-	+	-	+	+	+	-	-	-	-	-	+	+	+
m	+	-	+	+	-	+	-	-	-	-	+	-	-	+	-
n	+	-	+	+	+	-	-	-	-	-	+	-	-	+	-
ɲ	+	-	+	-	+	-	-	-	-	-	+	-	-	+	-
l	+	-	+	+	+	-	-	-	-	-	-	+	-	+	+
ʎ	+	-	+	-	+	-	-	-	-	-	-	+	-	+	+
ɫ	+	-	+	+	+	+	-	-	-	-	-	+	-	+	+
r	+	-	+	-	-	++	-	-	-	-	-	-	+	+	+
ɾ	+	-	+	-	-	++	-	-	-	-	-	-	-	+	+

Tableau A.10 – Traits distinctifs employés pour le portugais (consonnes).

A.3 Verbes du Chatino

Les deux lexiques du chatino proviennent tous deux de la base de donnée flexionnelle otomangue (*Oto-Manguean Inflectional Database*) de Feist et Palancar (2015). Les fichiers utilisés nous ont été fournis par Enrique L. Palancar.

A.3.1 Verbes du chatino de Yaitepec

Les données du chatino de Yaitepec de la base de donnée flexionnelle de Feist et Palancar (2015) proviennent de Rasch (2002). La thèse de Rasch (2002) constitue la principale documentation existante sur le chatino de Yaitepec. Les verbes du chatino de Yaitepec se fléchissent pour quatre valeurs d'aspect et de mode : le complétif, le potentiel, l'habituel et le progressif, chacun comportant trois personnes. Le tableau A.11 présente les première personnes de quatre verbes sélectionnés aléatoirement.

	1POT	1HAB	1PROG Opt	1CPL	...
SLYAʔ	ʃʌ̃ɔ̃ʔⁿ˩	ɳʃʌ̃ɔ̃ʔⁿ˩	ɳsʌ̃ɔ̃ʔⁿ˩	m̥sʌ̃ɔ̃ʔⁿ˩	...
XKA	ʃkɔ̃nˀ	ɳʃkɔ̃nˀ	ɳʃkɔ̃nˀ	m̥ʃkɔ̃nˀ	...
CHUʔ	kʲʧ̈ũʔⁿ˩	ɳʧ̈ũʔⁿ˩	ɳʧ̈ũʔⁿ˩	m̥ʧ̈ũʔⁿ˥	...
KWIʔ	ckʷĩʔⁿ˥	ɳɟkʷĩʔⁿ˥	ɳɟkʷĩʔⁿ˥	jkʷĩʔⁿ˥	...
...	...	...	...	...	...

TABLEAU A.11 – Extrait de quatre lexèmes issus du lexique du chatino de Yaitepec.

Les données de Rasch (2002) sont orthographiques. Nous avons élaboré des règles de phonémisation en collaboration avec Enrique Palancar, en nous fondant en grande partie sur les suggestions proposées par Rasch (2002). Nous résumons les règles employées pour les consommes dans le tableau A.12. Nous indiquons pour chaque phonème la notation de Rasch (2002), la transcription suggérée dans sa description, la transcription que nous employons, et les propriétés du phonème.

Les mots du chatino de Yaitepec sont tous monosyllabiques, et peuvent comporter une

Rasch (2002) (ortho.)	Rasch (2002) (API)	notre transcription	commentaire
x	ʃ	ʃ	fricative postalvéolaire non voisée
ch	ʧ	ʧ	affriquée postalvéolaire non voisée
j	h	h	fricative glottale non voisée
tz	ʦ	ʦ	affriquée alvéolaire non voisée
g	g	g	occlusive vélaire voisée
d	ð	ð	fricative dental voisée
nw	—	m	nasale bilabiale
y	—	j	semi voyelle
jy	—	hʲ	fricative glottale non voisée, palatalisée
7y	—	ʔʲ	occlusive glottale, palatalisée
ny	ñ,nʸ	ɲ	nasale palatale voisée
ly	lʸ	ʎ	palatale laterale voisée
7w	—	ʔʷ	occlusive glottale, labialisée
7n	—	ʔⁿ	glottale occlusive, vélarisée
jw	—	hʷ	fricative glottale non voisée, labialisée
gw	gw	gʷ	occlusive vélaire voisée, labialisée
t, (m/n)tt	t	t	occlusive alvéolaire non voisée
(m/n)t	?	d	occlusive alvéolaire voisée
k, (m/n)kk	k	k	occlusive vélaire non voisée
(m/n)k, g	g	g	occlusive vélaire voisée
ty, (m/n)tty	c	c	occlusive palatale non voisée
(m/n)ty	ɟ	ɟ	occlusive palatale voisée
kw, (m/n)kkw	kʷ	kʷ	occlusive vélaire non voisée, labialisée
(m/n)kw, gw	gw	gʷ	occlusive vélaire voisée, labialisée
(C)m	m̩	m̩	nasale bilabiale, syllabique
(C)n	n̩	n̩	nasale alvéolaire, syllabique
kuw	—	kw	cluster consonantique
tij	—	tj	cluster consonantique

Tableau A.12 – Résumé des règles de phonémisation employées pour les consonnes

Non nasalisé	Nasalisé
i	ĩ
e	ɛ̃
u	ũ
o	ũ
a	ɔ̃

Tableau A.13 – Nasalisation des voyelles en chatino de Yaitepec

voyelle initiale réduite. Rasch (2002, p. 24-25) décrit ainsi la structure syllabique d'un mot :

(84) (n/m) (C ($V_{[reduced]}$)) C V (ʔ)(n)

Les deux voyelles réduites sont /i/ et une forme de schwa. Nos données ne marquent que la première, que nous transcrivons /ⁱ/ comme le fait Rasch (2002). Par ailleurs, en position non réduite, les voyelles sont nasalisées après /n/, /ɲ/, et /ʔⁿ/ et avant /n/ ou /ʔⁿ/. La nasalisation peut changer la qualité de la voyelle, comme indiqué dans le tableau A.13.

Enfin, les tons, qui peuvent apparaître en combinaison, sont notés 1, 2, 3, et 4 dans nos données. Nous les notons respectivement ˥, ˧, ˦ et ˩. Afin de décomposer les tons en traits distinctifs, nous leur assignons une valeur négative pour tous les traits segmentaux, et nous créons trois traits TopScale, MidScale et BotScale. Les tons 1 et 2 sont [+TopScale], les 2 et 3 sont [+MidScale] et les tons 3 et 4 sont [+BotScale]. Nous indiquons dans les tableaux A.14 et A.15 les décompositions en traits des voyelles et des consonnes, qui sont principalement fondées sur Hayes (2012).

A.3.2 Verbes du chatino de Zenzontepec

Les données du chatino de Zenzontepec de la base de donnée flexionnelle de Feist et Palancar (2015) proviennent des données d'Eric Campbell. Le système flexionnel du chatino de Zenzontepec est documenté principalement dans Campbell (2011, 2014, 2016). Les verbes du chatino de Zenzontepec se fléchissent pour quatre valeurs d'aspect et de mode : le complétif, le potentiel, l'habituel et le progressif. Ils ne sont pas marqués pour la personne, il n'y a donc que

	segmental	syllabic	stress	long	consonantal	sonorant	continuant	delayed release	approximant	nasal	voice	spread gl	constr gl	labial	round	coronal	anterior	distributed	strident	lateral	dorsal	high	low	front	back	tense
n	+	-	-	-	+	+	-		-	+	+	-	-	-	-	+	+	-	-	-	-	-	-	-	-	
ʔ	+	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-				-	-	-	-	-	-	
ʔʷ	+	-	-	-	+	-	-	-	-	-	-	-	+	-	+	-				-	-	-	-	-	-	
ʔː	+	-	-	+	+	-	-	-	-	-	-	-	+	-	-	-				-	-	-	-	-	-	
ʔʲ	+	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-				-	-	+	-	+	-	
ʔⁿ	+	-	-	-	+	-	-	-	-	+	+	-	+	-	-	-				-	-	-	-	-	-	
d	+	-	-	-	+	-	-	-	-	-	+	-	-	-	-	+	+	-	-	-	-	-	-	-	-	
ɲ	+	-	-	-	+	+	-		-	+	+	-	-	-	-	+	-	+	-	-	+	+	-	+	-	
ɟ	+	-	-	-	+	-	-	-	-	-	+	-	-	-	-	+	-	+	-	-	+	+	-	+	-	
ɡ	+	-	-	-	+	-	-	-	-	-	+	-	-	-	-	-				-	+	+	-	-	-	
ɡʷ	+	-	-	-	+	-	-	-	-	-	+	-	-	-	+	-				-	+	+	-	-	-	
ʦ	+	-	-	-	+	-	-	+	-	-	-	-	-	-	-	+	+	-	+	-	-	-	-	-	-	
t	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	
c	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	+	+	-	+	-	
ʧ	+	-	-	-	+	-	-	+	-	-	-	-	-	-	-	+	-	+	+	-	-	-	-	-	-	
k	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-				-	+	+	-	-	-	
kʷ	+	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-				-	+	+	-	-	-	
l	+	-	-	-	+	+	+		+	-	+	-	-	-	-	+	+	-	-	+	-	-	-	-	-	
z	+	-	-	-	+	-	+	+	-	-	+	-	-	-	-	+	+	-	+	-	-	-	-	-	-	
ð	+	-	-	-	+	-	+	+	-	-	+	-	-	-	-	+	+	+	-	-	-	-	-	-	-	
ʎ	+	-	-	-	+	+	+		+	-	+	-	-	-	-	+	-	+	-	+	+	+	-	+	-	
s	+	-	-	-	+	-	+	+	-	-	-	-	-	-	-	+	+	-	+	-	-	-	-	-	-	
ʃ	+	-	-	-	+	-	+	+	-	-	-	-	-	-	-	+	-	+	+	-	-	-	-	-	-	
j	+	-	-	-	-	+	+		+	-	+	-	-	-	-	-				-	+	+	-	+	-	+
h	+	-	-	-	-	-	+	+	-	-	-	+	-	-	-	-				-	-	-	-	-	-	
hʷ	+	-	-	-	-	-	+	+	-	-	-	+	-	-	+	-				-	-	-	-	-	-	
hʲ	+	-	-	-	-	-	+	+	-	-	-	+	-	-	-	-				-	-	+	-	+	-	
w	+	-	-	-	-	+	+		+	-	+	-	-	+	+	-				-	+	+	-	-	+	+
n̥ː	+	+	-	+	+	+	-		-	+	+	-	-	-	-	+	+	-	-	-	-	-	-	-	-	
m̥	+	+	-	-	+	+	-		-	+	+	-	-	+	-	-				-	-	-	-	-	-	
n̥	+	+	-	-	+	+	-		-	+	+	-	-	-	-	+	+	-	-	-	-	-	-	-	-	

Tableau A.14 – Traits distinctifs employés pour le chatino de Yaitepec (consonnes).

	segmental	syllabic	stress	long	consonantal	sonorant	continuant	delayed release	approximant	nasal	voice	spread gl	constr gl	LABIAL	round	CORONAL	anterior	distributed	strident	lateral	DORSAL	high	low	front	back	tense
u	+	+	+	-	-	+	+		+	-	+	-	-	+	+	-				-	+	+	-	-	+	+
uː	+	+	+	+	-	+	+		+	-	+	-	-	+	+	-				-	+	+	-	-	+	+
ũ	+	+	+	-	-	+	+		+	+	+	-	-	+	+	-				-	+	+	-	-	+	+
ũː	+	+	+	+	-	+	+		+	+	+	-	-	+	+	-				-	+	+	-	-	+	+
o	+	+	+	-	-	+	+		+	-	+	-	-	+	+	-				-	+	-	-	-	+	+
oː	+	+	+	+	-	+	+		+	-	+	-	-	+	+	-				-	+	-	-	-	+	+
e	+	+	+	-	-	+	+		+	-	+	-	-	-	-	-				-	+	-	-	+	-	+
eː	+	+	+	+	-	+	+		+	-	+	-	-	-	-	-				-	+	-	-	+	-	+
ĩ	+	+	+	-	-	+	+		+	+	+	-	-	-	-	-				-	+	+	-	+	-	+
a	+	+	+	-	-	+	+		+	-	+	-	-	-	-	-				-	+	-	+	-	-	
aː	+	+	+	+	-	+	+		+	-	+	-	-	-	-	-				-	+	-	+	-	-	
ɔ̃	+	+	+	-	-	+	+		+	+	+	-	-	+	+	-				-	+	-	-	-	+	-
ɔ̃ː	+	+	+	+	-	+	+		+	+	+	-	-	+	+	-				-	+	-	-	-	+	-
ɛ̃	+	+	+	-	-	+	+		+	+	+	-	-	-	-	-				-	+	-	-	+	-	-
ɛ̃ː	+	+	+	+	-	+	+		+	+	+	-	-	-	-	-				-	+	-	-	+	-	-
i	+	+	+	-	-	+	+		+	-	+	-	-	-	-	-				-	+	+	-	+	-	+
ⁱ	+	+	-	-	-	+	+		+	-		-	-	-	-	-				-	+	+	-	+	-	+

Tableau A.15 – Traits distinctifs employés pour le chatino de Yaitepec (voyelles).

quatre cases de paradigme. Le tableau A.16 présente les formes de quatre verbes sélectionnés aléatoirement.

	cpl	pot	hab	prog
TU¹KWA²	nku⁰tu¹kwa²	tyu¹kwa²	ntyu¹kwa²	nte⁰tu¹kwa²
U⁰TYE⁰LE⁰	nka⁰tye²le¹	ku⁰tye⁰le⁰	ntu⁰tye⁰le⁰	nte⁰tye²le¹
U⁰TA¹ʔĄ²	nka⁰ta¹ʔą²	ku⁰ta¹ʔą²	ntu⁰ta¹ʔą²	nte⁰ta¹ʔą²
U⁰SA¹NA²	nka⁰sa¹na²	ku⁰sa⁰na¹	ntu⁰sa⁰na¹	nte⁰sa¹na²
...	...	...	...	...

TABLEAU A.16 – Extrait de quatre lexèmes issus du lexique du chatino de Zenzontepec.

La transcription employée pour les segments est celle d'origine. Nous marquons par /⁰/ toutes les syllabes qui portent un ton bas, non explicitement marqué dans le lexique d'origine, afin de pouvoir observer les alternances tonales. Afin de décomposer les tons en traits distinctifs, nous leur assignons une valeur négative pour tous les traits segmentaux, nous les laissons sous-spécifiés pour tous les trait segmentaux, les notons [-SEGMENT] et ajoutons un trait TONE qui prend respectivement les valeurs 0, 1 et 2 pour les tons /⁰/, /¹/ et /²/. Nous présentons dans le tableau A.17 la décomposition en traits des consonnes et des voyelles.

	segment	tone	syllabic	stress	long	consonantal	sonorant	continuant	delayed release	approximant	tap	trill	nasal	voice	spread gl	constr gl	labial	round	labiodental	coronal	anterior	distributed	strident	lateral	dorsal	high	low	front	back	tense
p	+		-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-				-	-					
b	+		-	-	-	+	-	+	+	-	-	-	-	+	-	-	+	-	-	-				-	-					
m	+		-	-	-	+	+	-		-	-	-	+	+	-	-	+	-	-	-				-	-					
t	+		-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-					
tz	+		-	-	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	+	+	-	+	-	-					
s	+		-	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-	+	+	-	+	-	-					
l	+		-	-	-	+	+	+		+	-	-	-	+	-	-	-	-	-	+	+	-	-	+	-					
n	+		-	-	-	+	+	-		-	-	-	+	+	-	-	-	-	-	+	+	-	-	-	-					
r	+		-	-	-	+	+	+		+	+	-	-	+	-	-	-	-	-	+	+	-	-	-	-					
ty	+		-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	+	+	-	+	-	
ch	+		-	-	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	+	+	-	-					
x	+		-	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-	+	-	+	+	-	-					
ly	+		-	-	-	+	+	+		+	-	-	-	+	-	-	-	-	-	+	+	-	-	+	+	+	-	+	-	
ny	+		-	-	-	+	+	-		-	-	-	+	+	-	-	-	-	-	+	+	-	-	-	+	+	-	+	-	
y	+		-	-	-	-	+	+		+	-	-	-	+	-	-	-	-	-	-				-	+	+	-	+	-	+
ky	+		-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-				-	+	+	-	+	-	
k	+		-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-				-	+	+	-	-	-	
kw	+		-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	+	-	-				-	+	+	-	-	-	
w	+		-	-	-	-	+	+		+	-	-	-	+	-	-	+	+	-	-				-	+	+	-	-	+	+
ʔ	+		-	-	-	+	-	-	-	-	-	-	-	-	-	+	-	-	-	-				-	-					
h	+		-	-	-	-	-	+	+	-	-	-	-	-	+	-	-	-	-	-				-	-					
a	+		+	-	-	-	+	+		+	-	-	-	+	-	-	-	-	-	-				-	+	-	+	-	-	
o	+		+	-	-	-	+	+		+	-	-	-	+	-	-	+	+	-	-				-	+	-	-	-	+	+
e	+		+	-	-	-	+	+		+	-	-	-	+	-	-	-	-	-	-				-	+	-	-	+	-	+
u	+		+	-	-	-	+	+		+	-	-	-	+	-	-	+	+	-	-				-	+	+	-	-	+	+
i	+		+	-	-	-	+	+		+	-	-	-	+	-	-	-	-	-	-				-	+	+	-	+	-	+
ą	+		+	-	-	-	+	+		+	-	-	+	+	-	-	-	-	-	-				-	+	-	+	-	-	
ǫ	+		+	-	-	-	+	+		+	-	-	+	+	-	-	+	+	-	-				-	+	-	-	-	+	+
ę	+		+	-	-	-	+	+		+	-	-	+	+	-	-	-	-	-	-				-	+	-	-	+	-	+
ų	+		+	-	-	-	+	+		+	-	-	+	+	-	-	+	+	-	-				-	+	+	-	-	+	+
į	+		+	-	-	-	+	+		+	-	-	+	+	-	-	-	-	-	-				-	+	+	-	+	-	+

Tableau A.17 – Traits distinctifs employés pour le chatino de Zenzontepec.

A.4 Verbes de l'arabe

Le lexique verbal de l'arabe a été extrait et normalisé à partir du wiktionnaire par Kirov et al. (2016) dans le cadre du projet unimorph [2]. Le tableau A.18 présente les formes de quatre lexèmes et quatre cases de paradigmes sélectionnées aléatoirement. Nos données comportent 1018 lexèmes.

	SBJV..PASS.M/F.2.D	IND.PST.ACT.M/F.1.P	IND.PRS.ACT.F.3.D	IMP.ACT.M.2.P	...
RAʾā	turaja:	raʔajna:	taraja:ni	raw	...
MAṯṯALA	tumaθθala:	maθθalna:	tumaθθila:ni	maθθilu:	...
INQAṭAʿA	#DEF#	inqatˁaʕna:	tanqatˁiʕa:ni	inqatˁiʕu:	...
RĀDDA ; RĀDADA	tura:dada: ; tura:dda:	ra:dadna:	#DEF#	ra:ddu: ; ra:didu:	...
...	...	...	...	...	...

TABLEAU A.18 – Extrait de quatre lexèmes issus du lexique de l'arabe.

Les verbes de l'arabe présentent 109 combinaisons possibles de mode, temps, voix, genre, personne et nombre. Nous synthétisons cette organisation dans le tableau A.19.

Dans les données de Kirov et al. (2016), les formes verbales de l'arabe sont transcrites au moyen d'une romanisation phonologique propre au Wiktionnaire. Nous avons traduit celle-ci automatiquement en API en nous fondant principalement sur la description de la romanisation du wiktionnaire [3]. Le tableau A.20 présente la correspondance entre caractères romanisés, orthographiques, et la transcription que nous avons adoptée. La transcription en API a été validée manuellement par une locutrice native sur 300 items. Les traits distinctifs utilisés sont fondés sur Hayes (2012), et sont synthétisés dans le tableau A.21.

2. Les données sont disponibles en ligne à l'adresse https://unimorph.github.io/.

3. Cette description peut être consultée à l'adresse https://en.wiktionary.org/wiki/Wiktionary:About_Arabic, section « *Comparative table of romanizations preferred and dispreferred by the English Wiktionary* ».

	indicatif			
	actif		passif	
	présent (imperfectif)	passé (perfectif)	présent (imperfectif)	passé (perfectif)
M/F.1.S	IND.PRS.ACT.M/F.1.S	IND.PST.ACT.M/F.1.S	IND.PRS.PASS.M/F.1.S	IND.PST.PASS.M/F.1.S
F.2.S	IND.PRS.ACT.F.2.S	IND.PST.ACT.F.2.S	IND.PRS.PASS.F.2.S	IND.PST.PASS.F.2.S
M.2.S	IND.PRS.ACT.M.2.S	IND.PST.ACT.M.2.S	IND.PRS.PASS.M.2.S	IND.PST.PASS.M.2.S
F.3.S	IND.PRS.ACT.F.3.S	IND.PST.ACT.F.3.S	IND.PRS.PASS.F.3.S	IND.PST.PASS.F.3.S
M.3.S	IND.PRS.ACT.M.3.S	IND.PST.ACT.M.3.S	IND.PRS.PASS.M.3.S	IND.PST.PASS.M.3.S
M/F.2.D	IND.PRS.ACT.M/F.2.D	IND.PST.ACT.M/F.2.D	IND.PRS.PASS.M/F.2.D	IND.PST.PASS.M/F.2.D
M.3.D	IND.PRS.ACT.M.3.D	IND.PST.ACT.M.3.D	IND.PRS.PASS.M.3.D	IND.PST.PASS.M.3.D
F.3.D	IND.PRS.ACT.F.3.D	IND.PST.ACT.F.3.D	IND.PRS.PASS.F.3.D	IND.PST.PASS.F.3.D
M/F.1.P	IND.PRS.ACT.M/F.1.P	IND.PST.ACT.M/F.1.P	IND.PRS.PASS.M/F.1.P	IND.PST.PASS.M/F.1.P
F.2.P	IND.PRS.ACT.F.2.P	IND.PST.ACT.F.2.P	IND.PRS.PASS.F.2.P	IND.PST.PASS.F.2.P
M.2.P	IND.PRS.ACT.M.2.P	IND.PST.ACT.M.2.P	IND.PRS.PASS.M.2.P	IND.PST.PASS.M.2.P
F.3.P	IND.PRS.ACT.F.3.P	IND.PST.ACT.F.3.P	IND.PRS.PASS.F.3.P	IND.PST.PASS.F.3.P
M.3.P	IND.PRS.ACT.M.3.P	IND.PST.ACT.M.3.P	IND.PRS.PASS.M.3.P	IND.PST.PASS.M.3.P

	jussif		subjonctif		impératif
	actif	passif	actif	passif	actif
M/F.1.S	JUSS.ACT.M/F.1.S	JUSS.PASS.M/F.1.S	SBJV..ACT.M/F.1.S	SBJV..PASS.M/F.1.S	
F.2.S	JUSS.ACT.F.2.S	JUSS.PASS.F.2.S	SBJV..ACT.F.2.S	SBJV..PASS.F.2.S	IMP.ACT.F.2.S
M.2.S	JUSS.ACT.M.2.S	JUSS.PASS.M.2.S	SBJV..ACT.M.2.S	SBJV..PASS.M.2.S	IMP.ACT.M.2.S
F.3.S	JUSS.ACT.F.3.S	JUSS.PASS.F.3.S	SBJV..ACT.F.3.S	SBJV..PASS.F.3.S	
M.3.S	JUSS.ACT.M.3.S	JUSS.PASS.M.3.S	SBJV..ACT.M.3.S	SBJV..PASS.M.3.S	
M/F.2.D	JUSS.ACT.M/F.2.D	JUSS.PASS.M/F.2.D	SBJV..ACT.M/F.2.D	SBJV..PASS.M/F.2.D	IMP.ACT.M/F.2.D
M.3.D	JUSS.ACT.M.3.D	JUSS.PASS.M.3.D	SBJV..ACT.M.3.D	SBJV..PASS.M.3.D	
F.3.D	JUSS.ACT.F.3.D	JUSS.PASS.F.3.D	SBJV..ACT.F.3.D	SBJV..PASS.F.3.D	
M/F.1.P	JUSS.ACT.M/F.1.P	JUSS.PASS.M/F.1.P	SBJV..ACT.M/F.1.P	SBJV..PASS.M/F.1.P	
F.2.P	JUSS.ACT.F.2.P	JUSS.PASS.F.2.P	SBJV..ACT.F.2.P	SBJV..PASS.F.2.P	IMP.ACT.F.2.P
M.2.P	JUSS.ACT.M.2.P	JUSS.PASS.M.2.P	SBJV..ACT.M.2.P	SBJV..PASS.M.2.P	IMP.ACT.M.2.P
F.3.P	JUSS.ACT.F.3.P	JUSS.PASS.F.3.P	SBJV..ACT.F.3.P	SBJV..PASS.F.3.P	
M.3.P	JUSS.ACT.M.3.P	JUSS.PASS.M.3.P	SBJV..ACT.M.3.P	SBJV..PASS.M.3.P	

Tableau A.19 – Organisation des cases de paradigmes le lexique de l'arabe.

Rom.	Orth.	API
ā	ا	aː
ʾa ʾu ʾ	أ	ʔa ʔu ʔ
ʾi	إ	ʔi
ʾā	آ	ʔaː
b	ب	b
t	ت	t
ṯ	ث	θ
j	ج	ʤ
ḥ	ح	ħ
ḵ	خ	x
d	د	d
ḏ	ذ	ð
r	ر	r
z	ز	z
s	س	s
š	ش	ʃ
ṣ	ص	sˤ
ḍ	ض	dˤ
ṭ	ط	tˤ
ẓ	ظ	ðˤ
ʿ	ع	ʕ
ḡ	غ	ɣ
f	ف	f
q	ق	q

Rom.	Orth.	API
k	ك	k
l	ل	l
m	م	m
n	ن	n
h	ه	h
w ū o ō	و	w uː o oː
ʾ	ؤ	ʔ
y ī e ē	ي	j iː e eː
ā	ى	aː
ʾ	ئ	ʔ
ʾ	ء	ʔ
a at	ة	a at
āh āt	ةا	aːh aːt
a	ـَ	a
u	ـُ	u
i	ـِ	i
an	ـًا; ـًى	an
un	ـٌ	un
in	ـٍ	in
aw	ـَو	aw
ū	ـُو	uː
ay	ـَي	aj
ī	ـِي	iː
ā	ـٰ	aː

Tableau A.20 – Résumé des règles de phonémisation employées pour l'arabe

	syllabic	long	consonantal	sonorant	continuant	delayed release	approximant	trill	nasal	voice	spread gl	constr gl	labial	round	labiodental	coronal	anterior	distributed	strident	lateral	dorsal	high	low	front	back	tense
h	-	-	-	-	+	+	-	-	-	-	+	-	-	-	-	-				-	-	-	-	-		
j	-	-	-	+	+		+	-	-	+	-	-	-	-	-	-				-	+	+	-	+	-	+
w	-	-	-	+	+		+	-	-	+	-	-	+	+	-	-				-	+	+	-	-	+	+
b	-	-	+	-	-	-	-	-	-	+	-	-	+	-	-	-				-	-	-	-	-		
d	-	-	+	-	-	-	-	-	-	+	-	-	-	-	-	+	+	-	-	-	-	-	-	-		
dˤ	-	-	+	-	-	-	-	-	-	+	-	-	-	-	-	+	+	-	-	-	-	-	+	-	+	
f	-	-	+	-	+	+	-	-	-	-	-	-	+	-	+	-				-	-	-	-	-		
k	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-				-	+	+	-	-		
l	-	-	+	+	+		+	-	-	+	-	-	-	-	-	+	+	-	-	+	-	-	-	-		
m	-	-	+	+	-		-	-	+	+	-	-	+	-	-	-				-	-	-	-	-		
n	-	-	+	+	-		-	-	+	+	-	-	-	-	-	+	+	-	-	-	-	-	-	-		
q	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-				-	+	-	-	-	+	
r	-	-	+	+	+		+	+	-	+	-	-	-	-	-	+	+	-	-	-	-	-	-	-		
s	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	+	+	-	+	-	-	-	-	-		
sˤ	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	+	+	-	+	-	-	-	+	-	+	
t	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-		
tˤ	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	+	-	+	
x	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-				-	+	+	-	-		
z	-	-	+	-	+	+	-	-	-	+	-	-	-	-	-	+	+	-	+	-	-	-	-	-		
ð	-	-	+	-	+	+	-	-	-	+	-	-	-	-	-	+	+	+	-	-	-	-	-	-		
ðˤ	-	-	+	-	+	+	-	-	-	+	-	-	-	-	-	+	+	+	-	-	-	-	+	-	+	
ħ	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-				-	+	-	+	-	+	
ɣ	-	-	+	-	+	+	-	-	-	+	-	-	-	-	-	-				-	+	+	-	-		
ʃ	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	+	-	+	+	-	-	-	-	-		
ʔ	-	-	+	-	-	-	-	-	-	-	-	+	-	-	-	-				-	-	-	-	-		
ʕ	-	-	+	-	+	-	-	-	-	+	-	-	-	-	-	-				-	+	-	+	-	+	
ʤ	-	-	+	-	-	+	-	-	-	+	-	-	-	-	-	+	-	+	+	-	-	-	-	-		
θ	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-		
a	+	-	-	+	+		+	-	-	+	-	-	-	-	-	-				-	+	-	+	-	-	
aː	+	+	-	+	+		+	-	-	+	-	-	-	-	-	-				-	+	-	+	-	-	
i	+	-	-	+	+		+	-	-	+	-	-	-	-	-	-				-	+	+	-	+	-	+
iː	+	+	-	+	+		+	-	-	+	-	-	-	-	-	-				-	+	+	-	+	-	+
u	+	-	-	+	+		+	-	-	+	-	-	+	+	-	-				-	+	+	-	-	+	+
uː	+	+	-	+	+		+	-	-	+	-	-	+	+	-	-				-	+	+	-	-	+	+

Tableau A.21 – Traits distinctifs employés pour l'arabe.

A.5 Verbes du navajo

Notre lexique du Navajo s'appuie sur des tables verbales extraites de Young et Morgan (1987), en collaboration avec Joyce McDonough. Le fichier des bases$_2$ a été fourni par Joyce McDonough, et les bases$_1$ ont été saisies manuellement par Benjamin Goehring. Un important travail de filtrage et de normalisation a été effectué sur ces fichiers, qui comportaient des erreurs de frappe (certaines provenant du dictionnaire), de reconnaissance optique de caractères et d'encodage. À plusieurs étapes de ce travail, nous avons validé quelques centaines de formes manuellement, conjointement avec Joyce McDonough afin d'ajuster les manipulations automatisées du lexique. 231 formes erronées dans le fichier des bases$_1$ ont été corrigées manuellement à la suite de ces validations. Nous avons également mené une évaluation semi-automatique en cherchant les n-grammes de caractères rares dans les formes, tant sur la notation orthographique qu'après phonémisation. Les verbes du navajo se conjuguent en mode (futur, imperfectif, optatif, perfectif et répétitif) et en personne, comme indiqué dans le tableau A.22.

Le dictionnaire spécifie, pour chaque entrée verbale, cinq parties principales, qui correspondent à la première personne pour chacun des modes (base$_1$ et base$_2$ combinées). La base$_2$ ne varie pas en personne, contrairement à la base$_1$. Des tableaux synthétisent les variations de la base$_1$ à travers tous les modes et les personnes. La génération des formes entières pour tous les modes et personnes a donc nécessité de segmenter les parties principales, et de réassembler les bases correspondantes pour chaque case de paradigme. Il existe des phénomènes de phonologie réguliers à l'interface entre ces bases, en particulier l'harmonie consonantique et vocalique (McDonough 2003), que nous avons implémentés sous forme de règles. Le résultat de ces règles a été vérifié manuellement. Nous excluons 904 verbes pour lesquels la conjugaison de la base$_2$ est irrégulière et ne figure pas dans les tables. L'inclusion de ces verbes nécessiterait un travail de saisie beaucoup plus important. Nous présentons dans le tableau A.23 quatre verbes et quatre cases de paradigmes, sélectionnés aléatoirement dans les données. Le lexique comporte des formes défectives ainsi que des formes surabondantes.

Nous avons implémenté une transcription automatique des formes orthographiques en notation phonémique, en suivant un guide de transcription établi par Joyce McDonough (et révisé

	optatif	itératif	futur	parfait	imparfait
1	OPT.1	ITER.1	FUT.1	PFV.1	IPFV.1
2	OPT.2	ITER.2	FUT.2	PFV.2	IPFV.2
3	OPT.3	ITER.3	FUT.3	PFV.3	IPFV.3
3A	OPT.3A	ITER.3A	FUT.3A	PFV.3A	IPFV.3A
3O	OPT.3O	ITER.3O	FUT.3O	PFV.3O	IPFV.3O
3I	OPT.3I	ITER.3I	FUT.3I	PFV.3I	IPFV.3I
3S	OPT.3S	ITER.3S	FUT.3S	PFV.3S	IPFV.3S
1DL	OPT.1DL	ITER.1DL	FUT.1DL	PFV.1DL	IPFV.1DL
2DL	OPT.2DL	ITER.2DL	FUT.2DL	PFV.2DL	IPFV.2DL
1PL	OPT.1PL	ITER.1PL	FUT.1PL	PFV.1PL	IPFV.1PL
2PL	OPT.2PL	ITER.2PL	FUT.2PL	PFV.2PL	IPFV.2PL
3PL	OPT.3PL	ITER.3PL	FUT.3PL	PFV.3PL	IPFV.3PL
3APL	OPT.3APL	ITER.3APL	FUT.3APL	PFV.3APL	IPFV.3APL
3OPL	OPT.3OPL	ITER.3OPL	FUT.3OPL	PFV.3OPL	IPFV.3OPL

Tableau A.22 – Organisation des cases de paradigmes le lexique du navajo.

	FUT.3o	PFV.1dl	FUT.1	FUT.3	...
ʾADINIIKÁÁH	#DEF#	ʔatini:k͡xai	#DEF#	#DEF#	...
BIʾNIISHʾÁÁH	jitíʔnóoʔah	piʔni:ʔah	pitíʔnéeʃʔah	pitíʔnóoʔah	...
ʾIISHʾAAH	ʔi:to:ʔá:ɬ	ʔi:ʔą́	ʔate:ʃʔá:ɬ	ʔato:ʔá:ɬ	...
YÁÁBÍDZIISTSʾIN	já:jízto:tsʼí:ɬ	já:pítsi:ɬtsʼin	já:píizte:stsʼí:ɬ	#DEF#	...
...	...	...	...	...	...

Tableau A.23 – Extrait de quatre lexèmes issus du lexique du navajo.

Ortho	IPA
shx	ʃh
sh	ʃ
zh	ʒ
t	t͡x
k	k͡x
dz	ʦ
dl	t͡ɬ
j	ʧ
ch’	ʧ’
ch	ʧʰ
tɬ	t͡ɬ
tɬ	t͡ɬʰ
ts’	ʦ’

Ortho	IPA
ts	ʦʰ
gh	ɣ
aa	aː
ą	ą
ąą	ąː
ąą́	ąː
áą́	áː
ee	eː
ęę	ęː
ęę́	ęː
éé	éː
oo	oː
ǫǫ	ǫː

Ortho	IPA
ǫ́ǫ́	ǫ́ː
óó	óː
ii	iː
įį	įː
íį́	íː
íí	íː
d	t
t’	ť
g	k
b	p
ɬ	ɬ
y	j
’	ʔ

Tableau A.24 – Résumé des règles de phonémisation employées pour le navajo

selon les validations manuelles). Nous synthétisons la correspondance entre caractères orthographiques et phonologiques dans le tableau A.24. Les traits distinctifs utilisés sont fondés sur Hayes (2012), et sont synthétisés dans les tableaux A.25 et A.26.

	syllabic	htone	long	consonantal	sonorant	continuant	delayed release	approximant	nasal	voice	spread gl	constr gl	labial	coronal	anterior	strident	lateral	dorsal	high	low	front	back
ɣ	-		-	+	-	+	+	-	-	+	-	-	-	-			-	+	+			
k	-		-	+	-	-	-	-	-	-	-	-	-	-			-	+	+			
k’	-		-	+	-	-	-	-	-	-	-	+	-	-			-	+	+			
k͡x	-		-	+	-	-	+	-	-	-	-	-	-	-			-	+	+			
l	-		-	+	+	+		+	-	+	-	-	-	+	+	-	+	-				
ɬ	-		-	+	-	+	+	-	-	-	-	-	-	+	+	-	+	-				
m	-		-	+	+	-		-	+	+	-	-	+	-			-	-				
n	-	-	-	+	+	-		-	+	+	-	-	-	+	+	-	-	-				
ń	-	+	-	+	+	-		-	+	+	-	-	-	+	+	-	-	-				
p	-		-	+	-	-	-	-	-	-	-	-	+	-			-	-				
s	-		-	+	-	+	+	-	-	-	-	-	-	+	+	+	-	-				
ʃ	-		-	+	-	+	+	-	-	-	-	-	-	+	-	+	-	-				
t	-		-	+	-	-	-	-	-	-	-	-	-	+	+	-	-	-				
t’	-		-	+	-	-	-	-	-	-	-	+	-	+	+	-	-	-				
t͡ɬ	-		-	+	-	-	+	-	-	-	-	-	-	+	+	-	+	-				
t͡ɬ’	-		-	+	-	-	+	-	-	-	-	+	-	+	+	-	+	-				
t͡ɬʰ	-		-	+	-	-	+	-	-	-	+	-	-	+	+	-	+	-				
ts	-		-	+	-	-	+	-	-	-	-	-	-	+	+	+	-	-				
ts’	-		-	+	-	-	+	-	-	-	-	+	-	+	+	+	-	-				
tsʰ	-		-	+	-	-	+	-	-	-	+	-	-	+	+	+	-	-				
tʃ	-		-	+	-	-	+	-	-	-	-	-	-	+	-	+	-	-				
tʃ’	-		-	+	-	-	+	-	-	-	-	+	-	+	-	+	-	-				
tʃʰ	-		-	+	-	-	+	-	-	-	+	-	-	+	-	+	-	-				
t͡x	-		-	+	-	-	+	-	-	-	-	-	-	-			-	+	+			
z	-		-	+	-	+	+	-	-	+	-	-	-	+	+	+	-	-				
ʒ	-		-	+	-	+	+	-	-	+	-	-	-	+	-	+	-	-				
ʔ	-		-	+	-	-	-	-	-	-	-	+	-	-			-	-				
x	-		-	+	-	+	+	-	-	-	-	-	-	-			-	+	+			
h	-		-	-	-	+	+	-	-	-	+		-	-			-	-				
j	-		-	-	+	+		+	-	+	-		-	-			-	+	+	-	+	-
w	-		-	-	+	+		+	-	+	-		+	-			-	+	+	-	-	+

Tableau A.25 – Traits distinctifs employés pour le navajo (consonnes).

	syllabic	htone	long	consonantal	sonorant	continuant	delayed release	approximant	nasal	voice	spread gl	constr gl	LABIAL	CORONAL	anterior	strident	lateral	DORSAL	high	low	front	back
a	+	-	-	-	+	+		+	-	+			-	-			-	+	-	+	-	-
á	+	+	-	-	+	+		+	-	+			-	-			-	+	-	+	-	-
ą	+	-	-	-	+	+		+	+	+			-	-			-	+	-	+	-	-
ą́	+	+	-	-	+	+		+	+	+			-	-			-	+	-	+	-	-
aː	+	-	+	-	+	+		+	-	+			-	-			-	+	-	+	-	-
áː	+	+	+	-	+	+		+	-	+			-	-			-	+	-	+	-	-
ąː	+	-	+	-	+	+		+	+	+			-	-			-	+	-	+	-	-
ą́ː	+	+	+	-	+	+		+	+	+			-	-			-	+	-	+	-	-
e	+	-	-	-	+	+		+	-	+			-	-			-	+	-	-	+	-
é	+	+	-	-	+	+		+	-	+			-	-			-	+	-	-	+	-
ę	+	-	-	-	+	+		+	+	+			-	-			-	+	-	-	+	-
ę́	+	+	-	-	+	+		+	+	+			-	-			-	+	-	-	+	-
eː	+	-	+	-	+	+		+	-	+			-	-			-	+	-	-	+	-
éː	+	+	+	-	+	+		+	-	+			-	-			-	+	-	-	+	-
ęː	+	-	+	-	+	+		+	+	+			-	-			-	+	-	-	+	-
ę́ː	+	+	+	-	+	+		+	+	+			-	-			-	+	-	-	+	-
i	+	-	-	-	+	+		+	-	+			-	-			-	+	+	-	+	-
í	+	+	-	-	+	+		+	-	+			-	-			-	+	+	-	+	-
į	+	-	-	-	+	+		+	+	+			-	-			-	+	+	-	+	-
į́	+	+	-	-	+	+		+	+	+			-	-			-	+	+	-	+	-
iː	+	-	+	-	+	+		+	-	+			-	-			-	+	+	-	+	-
íː	+	+	+	-	+	+		+	-	+			-	-			-	+	+	-	+	-
įː	+	-	+	-	+	+		+	+	+			-	-			-	+	+	-	+	-
į́ː	+	+	+	-	+	+		+	+	+			-	-			-	+	+	-	+	-
o	+	-	-	-	+	+		+	-	+			+	-			-	+	-	-	-	+
ó	+	+	-	-	+	+		+	-	+			+	-			-	+	-	-	-	+
ǫ	+	-	-	-	+	+		+	+	+			+	-			-	+	-	-	-	+
ǫ́	+	+	-	-	+	+		+	+	+			+	-			-	+	-	-	-	+
oː	+	-	+	-	+	+		+	-	+			+	-			-	+	-	-	-	+
óː	+	+	+	-	+	+		+	-	+			+	-			-	+	-	-	-	+
ǫː	+	-	+	-	+	+		+	+	+			+	-			-	+	-	-	-	+
ǫ́ː	+	+	+	-	+	+		+	+	+			+	-			-	+	-	-	-	+

Tableau A.26 – Traits distinctifs employés pour le navajo.

A.6 Noms du russe

Notre lexique du russe provient de données générées et fournies par Dunstan Brown, dans une romanisation orthographique. La transcription des formes en notation phonémique a été faite selon les indications de Dunstan Brown, comme synthétisé dans le tableau A.27. Des vérifications manuelles ont permis de corriger des erreurs dans les données d'entrée.

Le paradigme nominal du russe comporte six cas et deux nombres. Nos données distinguent également un locatif singulier alternatif. Le tableau A.28 synthétise les 13 cases de paradigme obtenues. Nous présentons dans le tableau A.29 les formes qui apparaissent dans le lexique pour quatre lexèmes et quatre cases de paradigmes sélectionnés aléatoirement. Le lexique comporte au total 1539 lexèmes. La décomposition en traits est fondée sur Hayes (2012) et est synthétisée dans les tableaux A.30 et A.31.

Transcription	IPA
(u)	(effacement, instrumental archaïque)
^	(effacement, symbole de concaténation)
_	(effacement)
@”	ˈ
-&	ˈ
ch’	t͡ɕ
ch	t͡ɕ
shch’ ou shch	ɕː
e	e, (i)
i	i
o	o, e
’ ou @’	ʲ
c	t͡s
sh	ʃ
zh	ʒ
k	k

Transcription	IPA
s	s
x	x
g	g
f	f
j	j
l	l
m	m
d	d
n	n
p	p
r	r
t	t
u	u
v	v
z	z
a	a
b	b

Tableau A.27 – Résumé des règles de phonémisation employées pour le russe

	singulier	puriel
nominatif	SG.NOM	PL.NOM
accusatif	SG.ACC	PL.ACC
génitif	SG.GEN	PL.GEN
datif	SG.DAT	PL.DAT
instrumental	SG.INS	PL.INS
locatif	SG.LOC	PL.LOC
locatif 2	SG.LOC2	

Tableau A.28 – Organisation des cases de paradigmes le lexique du russe.

	SG.INS	SG.GEN	SG.NOM	PL.LOC	...
Xoz'ajka	xozʲájkoj	xozʲájkʲi	xozʲájka	xozʲájkax	...
Prep'atstv'ijo	prʲepʲátstvʲijem	prʲepʲátstvʲija	prʲepʲátstvʲije	prʲepʲátstvʲijax	...
Strana	stranój	straní	straná	stránax	...
Operac'ija	opʲerátsijej	opʲerátsiji	opʲerátsija	opʲerátsijax	...
...	...	...	...	...	...

Tableau A.29 – Extrait de quatre lexèmes issus du lexique du russe.

	syllabic	stress	long	consonantal	sonorant	continuant	delayed release	approximant	trill	nasal	voice	labial	round	labiodental	coronal	anterior	distributed	strident	lateral	dorsal	high	low	front	back	tense	palatal
j	-	-	-	-	+	+	-	+	-	-	+	-	-	-	-				-	+					+	+
k	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-				-	+						-
kʲ	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-				-	+						+
t	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-						-
tʲ	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-						+
p	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-				-	-						-
pʲ	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-				-	-						+
g	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-				-	+						-
gʲ	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-				-	+						+
d	-	-	-	+	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-						-
dʲ	-	-	-	+	-	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-						+
b	-	-	-	+	-	-	-	-	-	-	+	+	-	-	-				-	-						-
bʲ	-	-	-	+	-	-	-	-	-	-	+	+	-	-	-				-	-						+
ʦ	-	-	-	+	-	-	+	-	-	-	-	-	-	-	+	+	-	+	-	-						-
ʨ	-	-	-	+	-	-	+	-	-	-	-	-	-	-	+	+	+	+	-	+						+
x	-	-	-	+	-	+	+	-	-	-	-	-	-	-	-				-	+						-
xʲ	-	-	-	+	-	+	+	-	-	-	-	-	-	-	-				-	+						+
ʃ	-	-	-	+	-	+	+	-	-	-	-	-	-	-	+	-	+	+	-	-						-
s	-	-	-	+	-	+	+	-	-	-	-	-	-	-	+	+	-	+	-	-						-
sʲ	-	-	-	+	-	+	+	-	-	-	-	-	-	-	+	+	-	+	-	-						+
f	-	-	-	+	-	+	+	-	-	-	-	+	-	+	-				-	-						-
fʲ	-	-	-	+	-	+	+	-	-	-	-	+	-	+	-				-	-						+
ʒ	-	-	-	+	-	+	+	-	-	-	+	-	-	-	+	-	+	+	-	-						-
z	-	-	-	+	-	+	+	-	-	-	+	-	-	-	+	+	-	+	-	-						-
zʲ	-	-	-	+	-	+	+	-	-	-	+	-	-	-	+	+	-	+	-	-						+
v	-	-	-	+	-	+	+	-	-	-	+	+	-	+	-				-	-						-
vʲ	-	-	-	+	-	+	+	-	-	-	+	+	-	+	-				-	-						+
n	-	-	-	+	+	-		-	-	+	+	-	-	-	+	+	-	-	-	-						-
nʲ	-	-	-	+	+	-		-	-	+	+	-	-	-	+	+	-	-	-	-						+
m	-	-	-	+	+	-		-	-	+	+	+	-	-	-				-	-						-
mʲ	-	-	-	+	+	-		-	-	+	+	+	-	-	-				-	-						+
l	-	-	-	+	+	+		+	-	-	+	-	-	-	+	+	-	-	+	-						-
lʲ	-	-	-	+	+	+		+	-	-	+	-	-	-	+	+	-	-	+	-						+
r	-	-	-	+	+	+		+	+	-	+	-	-	-	+	+	-	-	-	-						-
rʲ	-	-	-	+	+	+		+	+	-	+	-	-	-	+	+	-	-	-	-						+
ɕː	-	-	+	+	-	+	+	-	-	-	-	-	-	-	+	+	+	+	-	+						+

Tableau A.30 – Traits distinctifs employés pour le russe (consonnes).

	syllabic	stress	long	consonantal	sonorant	continuant	delayed release	approximant	trill	nasal	voice	labial	round	labiodental	coronal	anterior	distributed	strident	lateral	dorsal	high	low	front	back	tense	palatal
e	+	-	-	-	+	+		+	-	-	+	-	-	-	-				-	+	-	-	+	-	+	
a	+	-	-	-	+	+		+	-	-	+	-	-	-	-				-	+	-	+	-	-		
i	+	-	-	-	+	+		+	-	-	+	-	-	-	-				-	+	+	-	+	-	+	
o	+	-	-	-	+	+		+	-	-	+	+	+	-	-				-	+	-	-	-	+	+	
u	+	-	-	-	+	+		+	-	-	+	+	+	-	-				-	+	+	-	-	+	+	
é	+	+	-	-	+	+		+	-	-	+	-	-	-	-				-	+	-	-	+	-	+	
á	+	+	-	-	+	+		+	-	-	+	-	-	-	-				-	+	-	+	-	-		
í	+	+	-	-	+	+		+	-	-	+	-	-	-	-				-	+	+	-	+	-	+	
ó	+	+	-	-	+	+		+	-	-	+	+	+	-	-				-	+	-	-	-	+	+	
ú	+	+	-	-	+	+		+	-	-	+	+	+	-	-				-	+	+	-	-	+	+	

Tableau A.31 – Traits distinctifs employés pour le russe (voyelles).

A.7 Verbes de l'anglais

Le lexique de l'anglais est constitué des formes verbales de CELEX2 (Baayen, Piepenbrock et Gulikers 1995). Les notations SAMPA ont été transcrites automatiquement en IPA. Quelques formes manquantes ont été ajoutées manuellement. Ces données présentent beaucoup de surabondance, souvent due à des variantes dialectales. Lorsque la surabondance est sytématique à travers le paradigme d'un lexème en raison de deux bases concurrentes, nous dédoublons l'entrée, par exemple pour SECOND sur la base /sɪkɒnd/ ou sur la base /sɛkənd/. Les verbes concernés sont BOW, CONJURE, PROCESS, RECOUNT, REJOIN, ROW, et SECOND. Nous supprimons les /ɹ/ finaux qui sont sujets à variation dialectales et sont notés /(ɹ)/ dans les données CELEX2. Nous fusionnons les verbes dont l'ensemble des formes sont homonymes, comme par exemple RAIN, REIGN et REIN (111 verbes sont concernés, et ramenés à 54 après fusion). Les entrées verbales de CELEX2 comprennent un grand nombre de verbes prépositionnels. Nous ignorons 2144 entrées qui sont des variantes prépositionnelles de 883 bases déjà présentes dans le lexique. Pour 279 autres verbes prépositionnels, nous retirons les prépositions de la transcription phonologique et fusionnons si nécessaire les lignes identiques. Nous obtenons 6064 lexèmes distincts au to-

tal. Le tableau A.32 présente quatre verbes sélectionnés aléatoirement pour quatre cases de paradigme.

	INF	PRESOTHERS	PRES3S	PAST13	...
CHIP	tʃɪp	tʃɪp	tʃɪps	tʃɪpt	...
DISSEMBLE	dɪsɛmbl̩	dɪsɛmbl̩	dɪsɛmbl̩z	dɪsɛmbl̩d	...
CAGE	keːɪdʒ	keːɪdʒ	keːɪdʒɪz	keːɪdʒd	...
CLAW	klɔː	klɔː	klɔːz	klɔːd	...
...	...	...	...	...	...

TABLEAU A.32 – Extrait de quatre lexèmes issus du lexique de l'anglais.

La plupart des verbes de l'anglais ne présentent que cinq cases de paradigme (le présent troisième personne, l'infinitif/autres formes du présent, le participe passé, le participe présent, le passé). Cependant le verbe TO BE distingue le présent première personne de l'infinitif et des autres formes du présent, ainsi que le passé première et troisième personne des autres formes du passé. Afin d'obtenir une grille de paradigme homogène, nous comptons donc huit cases de paradigme : participe passé, infinitif, présent première personne, présent troisième personne, autres cases du présent, participe présent, passé première et troisième personne, autres cases du passé.

La décomposition des phonèmes en traits distinctifs se fonde principalement sur la description de Halle et Clements (1983), avec quelques ajustements. Pour noter le contraste entre /s/ et /θ/, /z/ et /ð/, nous adoptons le trait [±strident] proposé par Chomsky et Halle (1968). Halle et Clements (1983) proposent une représentation non segmentale pour noter les affriquées /ʧ/ et /ʤ/. Comme nous avons besoin de faire cette distinction en termes de traits, nous optons pour la solution de Chomsky et Halle (1968) qui les notent comme les occlusives /t/ et /d/, mais avec le trait [+strident]. Nos données comportent des voyelles longues, que nous notons [+tense]. Elles contiennent également les diphtongues /æ/, /aɪ/, /aʊ/, /eɪ/, /əʊ/, /ɛə/, /ɪə/, /ɔɪ/, /ʊə/. Nous ajoutons des traits [±diphtong j], [±diphtong ə] et [±diphtong w]. Pour chaque diphtongue, nous l'encodons comme sa voyelle initiale, avec le trait [+tense] ainsi que le trait de diphtongaison

correspondant à la seconde voyelle. Par exemple, /aɪ/ est encodé comme un /a/ avec [+tense] et [+diphtong j]. La syllabicité des consonnes /m/, /n/, /ŋ/ et /l/ est également notée dans les données, nous ajoutons donc un trait [± syllabic]. Enfin, nous notons le schwa /ə/ comme une voyelle non haute, non b asse, non arrière, non labiale, non tendue et non nasale. Le système résultant est synthétisé dans les tableaux A.33 et A.34.

	strident	syllabic	sonorant	consonantic	high	low	back	labial	tense	nasal	continuant	coronal	lateral	anterior	spread	voice	diphtong j	diphtong ə	diphtong w
p	-	-	-	+				+			-	-		+		-			
b	-	-	-	+				+			-	-		+		+			
f	+	-	-	+				+			+	-		+		-			
v	+	-	-	+				+			+	-		+		+			
t	-	-	-	+				-			-	+		+		-			
d	-	-	-	+				-			-	+		+		+			
θ	-	-	-	+				-			+	+		+		-			
ð	-	-	-	+				-			+	+		+		+			
s	+	-	-	+				-			+	+		+		-			
z	+	-	-	+				-			+	+		+		+			
ʃ	+	-	-	+				-			+	+		-		-			
ʧ	+	-	-	+				-			-	+		-		-			
ʒ	+	-	-	+				-			+	+		-		+			
ʤ	+	-	-	+				-			-	+		-		+			
k	-	-	-	+	+	-	+	-			-	-		-		-			
g	-	-	-	+	+	-	+	-			-	-		-		+			
x	-	-	-	+	+	-	+	-			+	-		-		-			
m	-	-	+	+				+		+	-	-	-	+	-				
m̩	-	+	+	+				+		+	-	-	-	+	-				
n	-	-	+	+				-		+	-	+	-	+	-				
n̩	-	+	+	+				-		+	-	+	-	+	-				
ŋ	-	-	+	+	+		+	-		+	-	-	-	-	-				
ŋ̩	-	+	+	+	+		+	-		+	-	-	-	-	-				
l	-	-	+	+				-		-	-	+	+	+	-				
l̩	-	+	+	+				-		-	-	+	+	+	-				
r	-	-	+	+				-		-	+	+	-	+	-				
j		-	+	-	+		-	-		-	+	+	-	-	-				
w		-	+	-	+		+	+		-	+	-	-	-	-				
h	-	-	+	-						-	+		-		+				

Tableau A.33 – Traits distinctifs employés pour l'anglais (consonnes).

	strident	syllabic	sonorant	consonantic	high	low	back	labial	tense	nasal	continuant	coronal	lateral	anterior	spread	voice	diphtong j	diphtong ə	diphtong w
ɛ		+	+	-	-	-	-	-	-	-							-	-	-
ə		+	+	-	-	-	-	-	-	-							-	-	-
əˑʊ		+	+	-	-	-	-	-	-	-							-	-	+
ɛˑə		+	+	-	-	-	-	-	+	-							-	+	-
eˑɪ		+	+	-	-	-	-	-	+	-							+	-	-
æ		+	+	-	-	+	-	-	-	-							-	-	-
a		+	+	-	-	+	-	-	+	-							-	-	-
ʌ		+	+	-	-	-	+	-	-	-							-	-	-
ɜː		+	+	-	-	-	+	-	+	-							-	-	-
ɑː		+	+	-	-	+	+	-	+	-							-	-	-
aˑʊ		+	+	-	-	+	+	-	+	-							-	-	+
aˑɪ		+	+	-	-	+	+	-	+	-							+	-	-
ɪ		+	+	-	+	-	-	-	-	-							-	-	-
ɪˑə		+	+	-	+	-	-	-	-	-							-	+	-
i̇ː		+	+	-	+	-	-	-	+	-							-	-	-
ɔː		+	+	-	-	-	+	+	-	-							-	-	-
ɔˑɪ		+	+	-	-	-	+	+	+	-							+	-	-
ɒ		+	+	-	-	+	+	+	-	-							-	-	-
ʊ		+	+	-	+	-	+	+	-	-							-	-	-
ʊˑə		+	+	-	+	-	+	+	-	-							-	+	-
uː		+	+	-	+	-	+	+	+	-							-	-	-
æ̃		+	+	-	-	+	-	-	-	+							-	-	-
æ̃ː		+	+	-	-	+	-	-	+	+							-	-	-
ɑ̃ː		+	+	-	-	+	+	-	+	+							-	-	-
ɔ̃ː		+	+	-	-	-	+	+	+	+							-	-	-

Tableau A.34 – Traits distinctifs employés pour l'anglais (voyelles).

Annexe B

Classifications hiérarchiques des microclasses (UPGMA)

Dans la section 4.3.2 du chapitre 4, nous avons montré comment il est possible de mesurer des distances entre microclasses en se fondant sur les patrons d'alternance, et comment l'algorithme UPGMA peut être employé pour produire des classifications arborescentes fondées sur ces distances. Nous avons observé les matrices de distances et les arbres résultants pour le français, le portugais européen, le russe et le chatino de Zenzontepec. Cette annexe présente les figures correspondantes pour les verbes de l'arabe (figure B.1), du chatino de Yaitepec (figures B.2 et B.3), de l'anglais (figure B.4) et du navajo (figure B.5 et B.6).

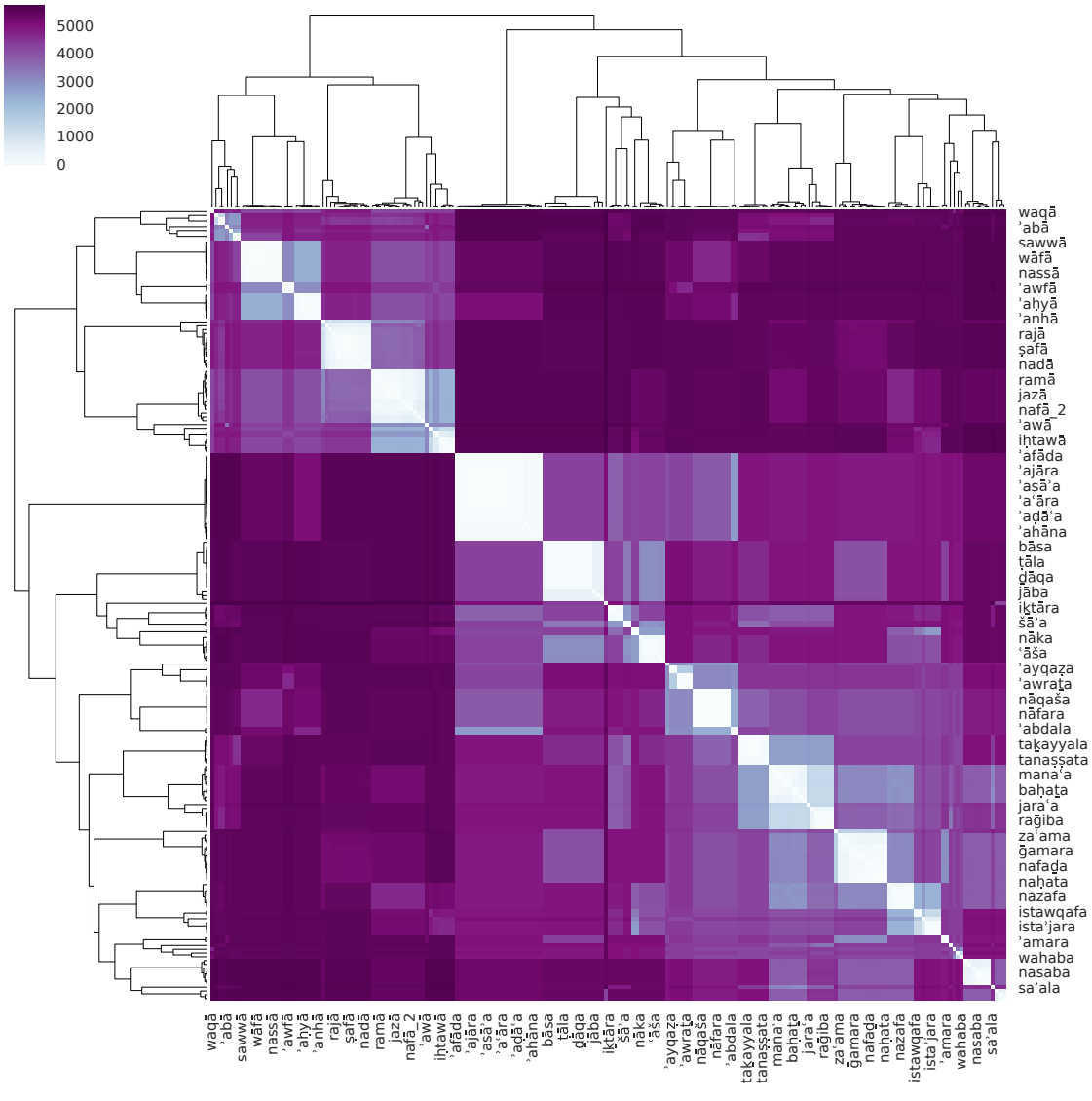


FIGURE B.1 – Classification hiérarchique des verbes de l'arabe par l'algorithme UPGMA.

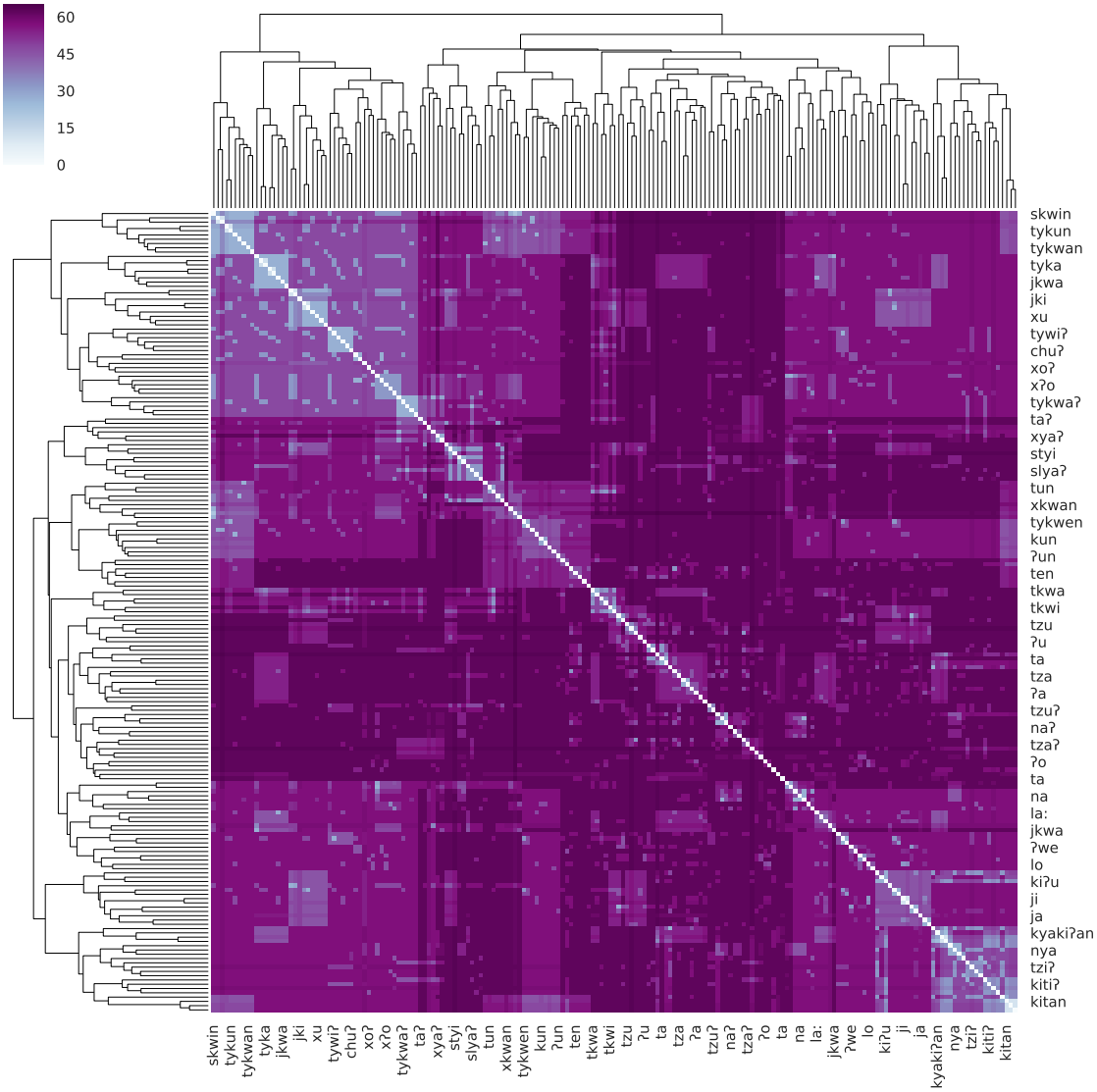


FIGURE B.2 – Classification hiérarchique des verbes du chatino de Yaitepec par l'algorithme UPGMA (tons).

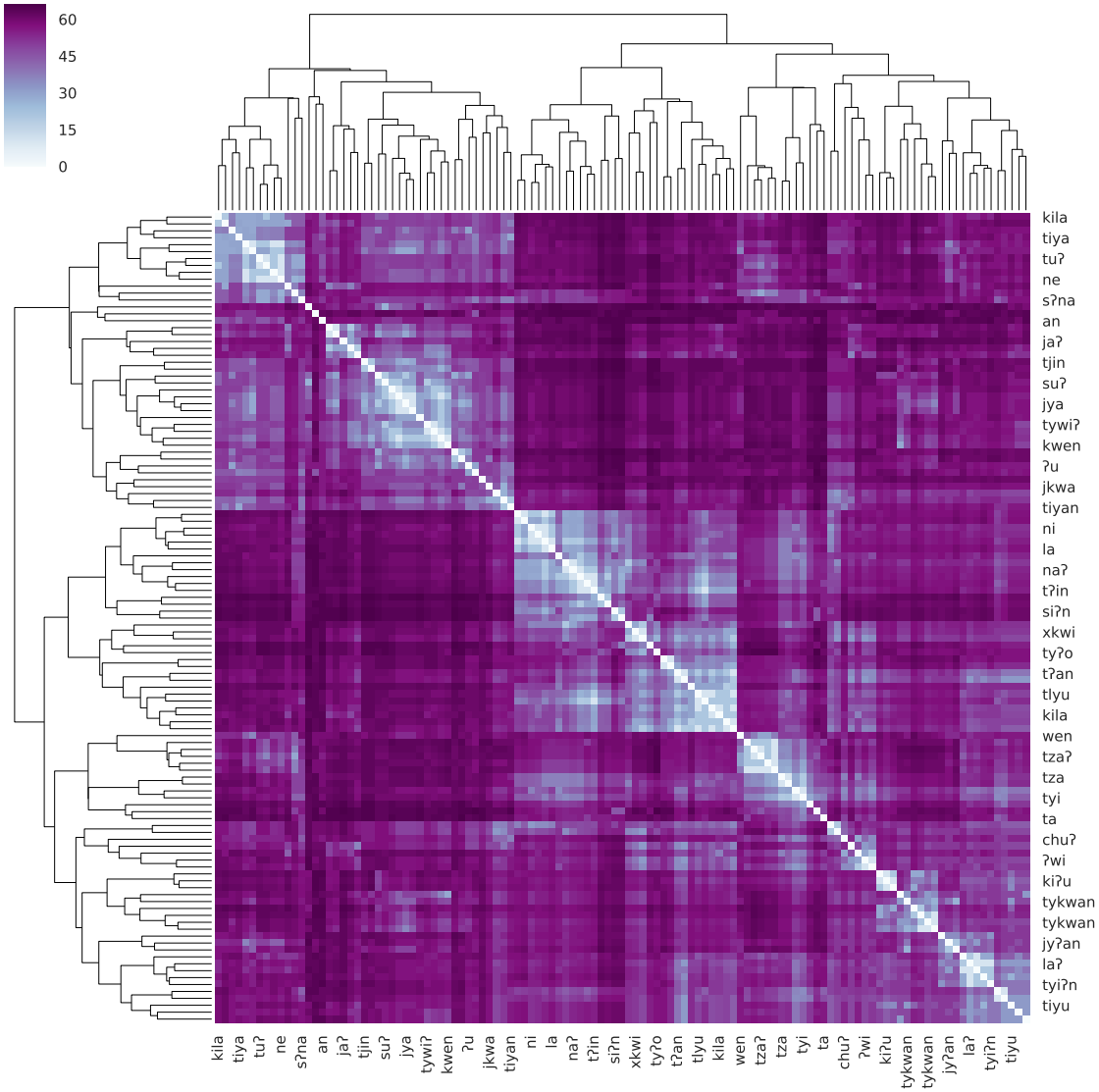


FIGURE B.3 – Classification hiérarchique des verbes du chatino de Yaitepec par l'algorithme UPGMA (tons).

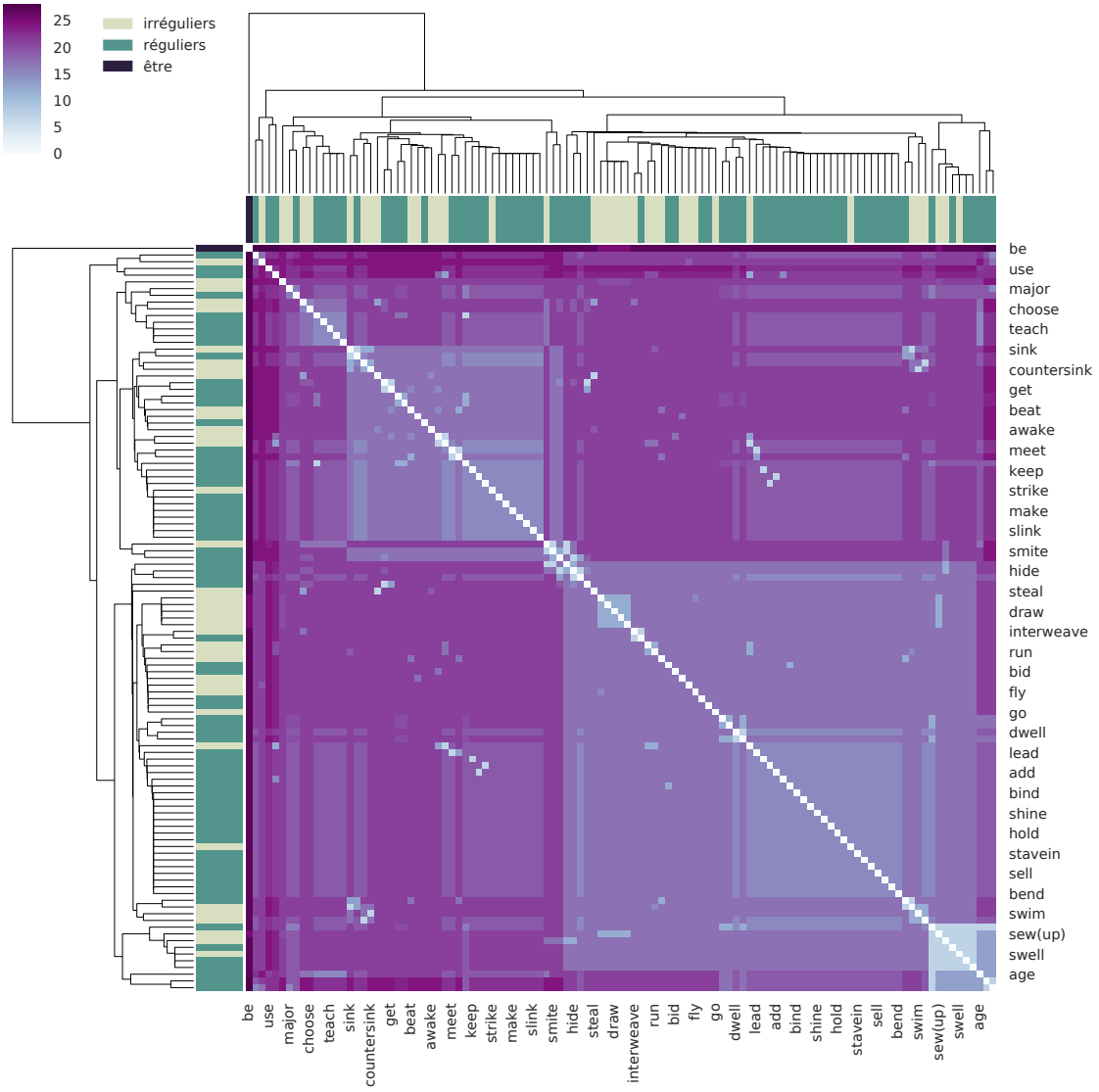


FIGURE B.4 – Classification hiérarchique des verbes de l'anglais par l'algorithme UPGMA.

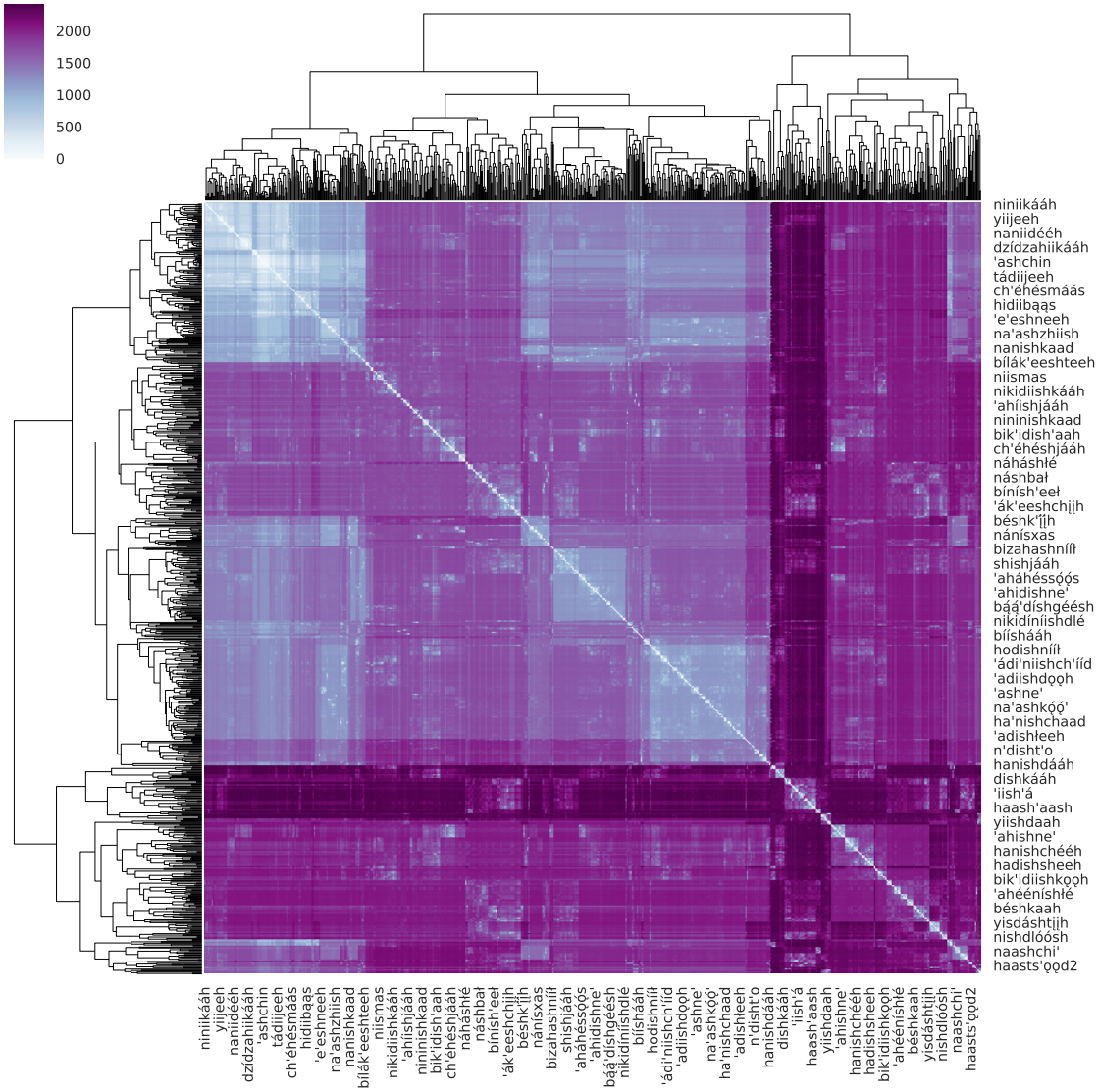


FIGURE B.5 – Classification hiérarchique des verbes du navajo par l'algorithme UPGMA ($bases_1$).

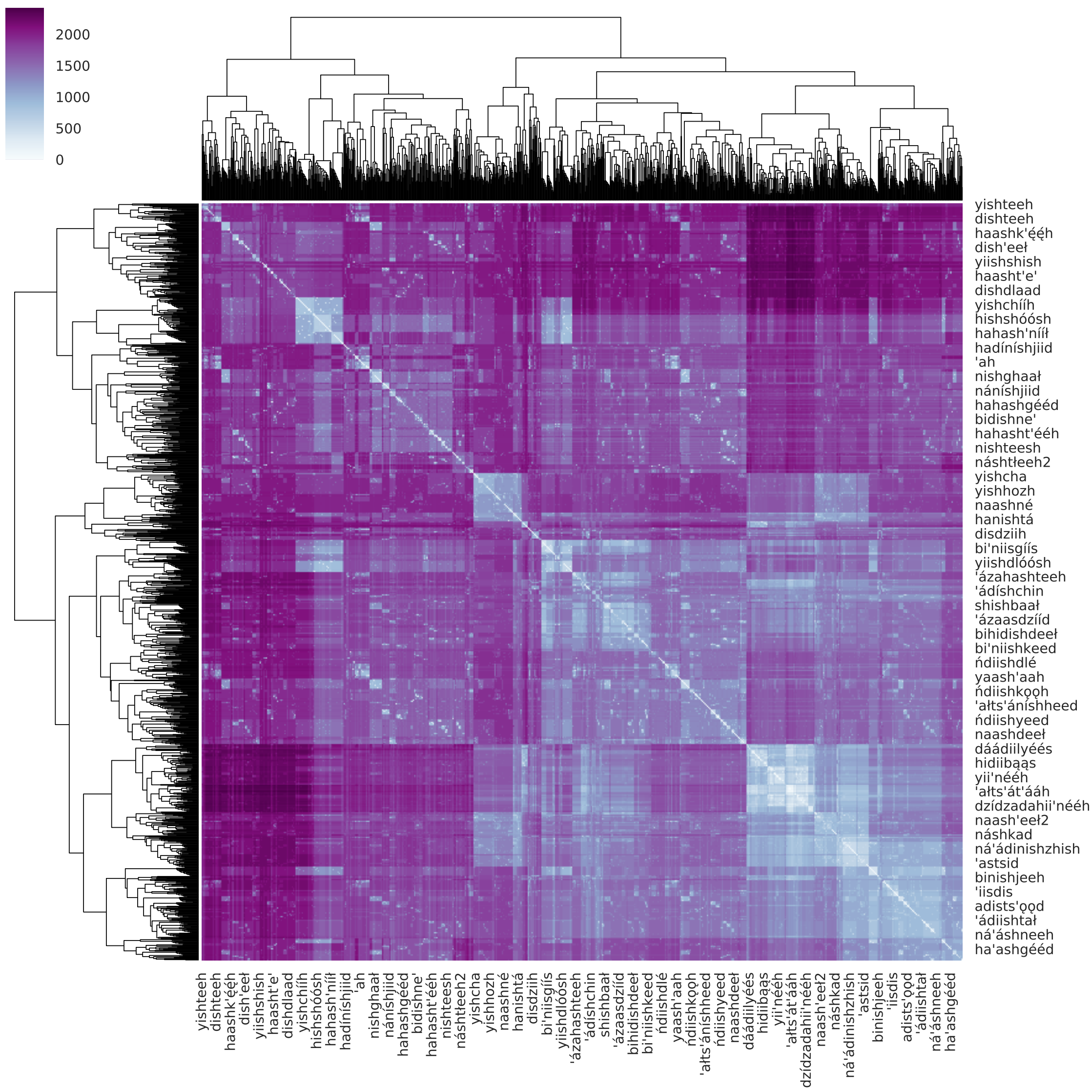


Figure B.6 – Classification hiérarchique des verbes du navajo par l'algorithme UPGMA (bases$_2$).

Bibliographie

Ackerman, Farrell, James P. Blevins et Robert Malouf (2009). « Parts and wholes : implicative patterns in inflectional paradigms ». In : *Analogy in Grammar*. Sous la dir. de James P. Blevins et Juliette Blevins. Oxford : Oxford University Press, p. 54–82.

Ackerman, Farrell et Robert Malouf (2013). « Morphological organization : the low conditional entropy conjecture ». In : *Language* 89, p. 429–464.

— (2015). « The No Blur Principle Effects as an Emergent Property of Language Systems ». In : *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*. T. 41. DOI : 10.20354/B4414110014.

Ahlberg, Malin, Markus Forsberg et Mans Hulden (2015). « Paradigm classification in supervised learning of morphology ». In : p. 1024–1029. DOI : 10.3115/v1/N15-1107. URL : http://www.aclweb.org/anthology/N15-1107.

Ahlberg, Malin, Markus Forsberg et Manstio Hulden (2014). « Semi-supervised learning of morphological paradigms and lexicons ». In : *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden, p. 569–578. ISBN : 978-1-937284-78-7. DOI : 10.3115/v1/E14-1060.

Albright, Adam C. (2002). « The Identification of Bases in Morphological Paradigms ». Thèse de doct. University of California, Los Angeles.

Albright, Adam C. et Bruce P. Hayes (2003). « Rules vs. Analogy in English Past Tenses : A Computational/Experimental Study ». In : *Cognition* 90, p. 119–161.

Albright, Adam et Bruce Hayes (1999). « An automated learner for phonology and morphology ».

Albright, Adam et Bruce Hayes (2002). « Modeling English Past Tense Intuitions with Minimal Generalization ». In : *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning - Volume 6*. Stroudsburg, PA, USA : Association for Computational Linguistics, p. 58–69. DOI : 10.3115/1118647.1118654.

— (2006). « Modeling productivity with the Gradual Learning Algorithm : the problem of accidentally exceptionless generalizations ». In : *Gradience in Grammar : Generative Perspectives*. Sous la dir. de Gisbert Fanselow et al. Oxford : Oxford University Press, p. 185–204.

Aronoff, Mark (1994). *Morphology by itself*. Cambridge : MIT Press.

Arrivé, Michel, éd. (2012). *Bescherelle : La conjugaison pour tous*. nouvelle édition. Hatier.

Baayen, R, R Piepenbrock et L Gulikers (1995). *CELEX2 LDC96L14*. Philadelphia : Linguistic Data Consortium.

Bahdanau, Dzmitry, Kyunghyun Cho et Yoshua Bengio (2014). « Neural Machine Translation by Jointly Learning to Align and Translate ». In : *CoRR* abs/1409.0473. arXiv : 1409.0473. URL : http://arxiv.org/abs/1409.0473.

Bank, Sebastian (2016). « Assessing the typology of person portmanteaus ». In : *Under review*.

Baroni, Marco et al. (2009). « The WaCky Wide Web : A Collection of Very Large Linguistically Processed Web-Crawled Corpora ». In : 43, p. 209–226.

Bělohlávek, Radim (2009). « Introduction to formal concept analysis ». Olomouc. URL : https://phoenix.inf.upol.cz/esf/ucebni/formal.pdf.

Beniamine, Sacha (2017). « Un algorithme universel pour l'abstraction automatique d'alternances morphophonologiques ». In : *Actes de TALN 2017*, p. 77–85.

Beniamine, Sacha, Olivier Bonami et Joyce McDonough (2017). « When segmentation helps. Implicative structure and morph boundaries in the Navajo verb ». In : *First International Symposium on Morphology*. Lille.

Beniamine, Sacha, Olivier Bonami et Benoît Sagot (2017). « Inferring Inflection Classes with Description Length ». In : *Journal of Language Modelling* 5.3. DOI : 10.15398/jlm.v5i3.184.

Berko, Jean (1958). « The Child's Learning of English Morphology ». In : *Word* 14, p. 150–177.

Blevins, James P. (2006). « Word-based morphology ». In : *Journal of Linguistics* 42, p. 531–573.

— (2016). *Word and Paradigm Morphology*. Oxford : Oxford University Press.

Blevins, James P., Petar Milin et Michael Ramscar (2017). « The Zipfian Paradigm Cell Filling Problem ». In : *Morphological paradigms and functions*. Sous la dir. de Ferenc Kiefer, James P. Blevins et Huba Bartos. Leiden : Brill.

Blevins, Jim (2004). « Inflection classes and economy ». In : *Explorations in Nominal Inflection*. Sous la dir. de L. Gunkel, G. Müller et G. Zifonun. Berlin, Boston : De Gruyter, p. 41–85. DOI : 10.1515/9783110197501.51.

Bonami, Olivier (2014). « La structure fine des paradigmes de flexion ». Mémoire d'habilitation, Université Paris Diderot.

Bonami, Olivier et Sacha Beniamine (2016). « Joint predictiveness in inflectional paradigms ». In : *Word Structure* 9.2, p. 156–182. DOI : https://doi.org/10.3366/word.2016.0092.

Bonami, Olivier et Gilles Boyé (2003a). « La construction des paradigmes ». In : *Deuxièmes Décembrettes*. Toulouse.

— (2003b). « Supplétion et classes flexionnelles dans la conjugaison du français ». In : *Langages* 152, p. 102–126. ISSN : 0458-726X. DOI : 10.3406/lgge.2003.2441.

— (2014). « De formes en thèmes ». In : *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux*. Sous la dir. de Florence Villoing, Sarah Leroy et Sophie David. Presses Universitaires de Paris Ouest, p. 17–45.

Bonami, Olivier, Gilles Boyé, Hélène Giraudo et al. (2008). « Quels verbes sont réguliers en français ? » In : *Actes du premier Congrès Mondial de Linguistique Française*, p. 1511–1523.

Bonami, Olivier, Gilles Boyé et Fabiola Henri (2011). « Measuring inflectional complexity : French and Mauritian ». In : *Workshop on Quantitative Measures in Morphology and Morphological Development*. San Diego.

Bonami, Olivier, Gauthier Caron et Clément Plancq (2014). « Construction d'un lexique flexionnel phonétisé libre du français ». In : *Actes du quatrième Congrès Mondial de Linguistique Française*. Sous la dir. de Franck Neveu et al., p. 2583–2596.

Bonami, Olivier et Ana R. Luís (2014). « Sur la morphologie implicative dans la conjugaison du portugais : une étude quantitative ». In : *Morphologie flexionnelle et dialectologie romane. Typologie(s) et modélisation(s)*. Sous la dir. de Jean-Léonard Léonard. Mémoires de la Société de Linguistique de Paris 22. Leuven : Peeters, p. 111–151.

Boyé, Gilles (2000). « Problèmes de morpho-phonologie verbale en français, espagnol et italien ». Thèse de doct. Université Paris 7.

Brown, Dunstan (1998). « Defining 'subgender' : Virile and devirilized nouns in Polish ». In : *Lingua* 104, p. 187–233.

Brown, Dunstan et Roger Evans (2012). « Morphological complexity and unsupervised learning : validating Russian inflectional classes using high frequency data ». In : *Current Issues in Morphological Theory : (Ir)regularity, analogy and frequency*. Sous la dir. de F. Kiefer, M. Ladányi et P. Siptár. Amsterdam : John Benjamins, p. 135–162.

Brown, Dunstan et Andrew Hippisley (2012). *Network Morphology : a defaults based theory of word structure*. Cambridge : Cambridge University Press.

Campbell, Eric (2011). « Zenzontepec Chatino Aspect Morphology and Zapotecan Verb Classes ». In : *International Journal of American Linguistics* 77, p. 219–246.

— (2014). « Aspects of the phonology and morphology of Zenzontepec Chatino, a Zapotecan language of Oaxaca, Mexico ». Thèse de doct. University of Texas at Austin.

— (2016). « Tone and inflection in Zenzontepec Chatino ». In : *Tone and inflection*. Sous la dir. d'Enrique L. Palancar et Jean-Léonard Léonard. Berlin : Mouton de Gruyter, p. 141–162.

Carnie, Andrew (2008). *Irish nouns : a reference guide*. Oxford : Oxford University Press.

Carstairs, Andrew (1987). *Allomorphy in Inflection*. London : Croom Helm.

Carstairs-McCarthy, Andrew (1991). « Inflection classes : Two questions with one answer. » In : *Paradigms : The Economy of Inflection*. Sous la dir. de F. Plank. Empirical Approaches to Language Typology [EALT]. De Gruyter, p. 213–253. ISBN : 9783110889109. DOI : 10.1515/9783110889109.213.

— (1994). « Inflection Classes, Gender, and the Principle of Contrast ». In : *Language* 70, p. 737–788.

Chan, Erwin (2008). « Structures and distributions in morphological learning ». Thèse de doct. University of Pennsylvania.

Chomsky, Noam (1965). *Aspects of the theory of syntax*. Cambridge : MIT Press.

Chomsky, Noam et Morris Halle (1968). *The sound pattern of English*. Harper et Row.

Cilibrasi, R. et P.M.B. Vitanyi (2005). « Clustering by Compression ». In : *IEEE Transactions on Information Theory* 51.4, p. 1523–1545. DOI : 10.1109/tit.2005.844059.

Clahsen, Harald (2006). « Dual-mechanism morphology ». In : *Encyclopedia of Language and Linguistics*. Sous la dir. de Keith Brown. T. 4. Elsevier, p. 1–5.

Corbett, Greville G (1982). « Gender in Russian : An account of gender specification and its relationship to declension ». In : *Russian linguistics* 6 (2), p. 197–232.

Corbett, Greville G. (2009). « Canonical inflection classes ». In : *Selected Proceedings of the 6th Décembrettes : Morphology in Bordeaux*. Sous la dir. de Fabio Montermini, Gilles Boyé et Jesse Tseng. Somerville : Cascadilla Press, p. 1–11.

Corbett, Greville G. et Norman M. Fraser (1993). « Network Morphology : a DATR account of Russian nominal inflection ». In : *Journal of Linguistics* 29, p. 113–142.

Cotterell, Ryan et al. (2016). « The SIGMORPHON 2016 Shared Task—Morphological Reinflection ». In : *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Berlin, Germany : Association for Computational Linguistics, p. 10–22. DOI : 10.18653/v1/W16-2002. URL : http://www.aclweb.org/anthology/W16-2002.

Dell, François (1973). *Les règles et les sons : Introduction à la phonologie générative*. Collection Savoir. Paris : Hermann. ISBN : 9782705657680.

Dressler, Wolfgang U. (2012). « On the acquisition of inflectional morphology : introduction ». In : *Morphology* 22.1, p. 1–8. ISSN : 1871-5656. DOI : 10.1007/s11525-011-9198-1.

Dressler, Wolfgang U, Marianne Kilani-Schoch et al. (2008). « On the Typology of Inflection Class Systems ». In : *Folia Linguistica* 40.1-2 (Special Issue : Natural Morphology.), p. 51–74. DOI : 10.1515/flin.40.1-2.51.

Dressler, Wolfgang U., Willi Mayerthaler et al. (1987). *Leitmotifs in natural morphology*. T. 10. 10 t. Studies in Language Companion Series. Amsterdam : John Benjamins Publishing. DOI : 10.1075/slcs.10.

Dressler, Wolfgang U. et Anna M. Thornton (1996). « Italian Nominal Inflection ». In : *Wiener Linguistische Gazette* 55-57, p. 1–26.

Durrett, Greg et John DeNero (2013). « Supervised Learning of Complete Morphological Paradigms ». In : *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. Atlanta, Georgia : Association for Computational Linguistics, p. 1185–1195. URL : http://www.aclweb.org/anthology/N13-1138.

Embick, David et Morris Halle (2005). « On the status of stems in morphological theory ». In : *Romance languages and linguistic theory 2003*. Sous la dir. de Twan Geerts, Ivo van Ginneken et Haike Jacobs. Amsterdam : John Benjamins, p. 1–31.

Faaß, Gertrud et Kerstin Eckart (2013). « SdeWaC — a corpus of parsable sentences from the web ». In : *Language processing and knowledge in the Web*. Sous la dir. d'Irina Gurevych, Chris Biemann et Torsten Zesch. Heidelberg : Springer, p. 61–68.

Feist, Timothy et Enrique L. Palancar (2015). *Oto-Manguean Inflectional Class Database*. University of Surrey. DOI : 10.15126/SMG.28/1.

Finkel, Raphael et Gregory T. Stump (2007). « Principal parts and morphological typology ». In : *Morphology* 17, p. 39–75.

— (2009). « Principal parts and degrees of paradigmatic transparency ». In : *Analogy in Grammar*. Sous la dir. de James P. Blevins et Juliette Blevins. Cambridge : Cambridge University Press, p. 13–54.

Flickinger, Dan (1987). « Lexical rules in the hierarchical lexicon ». Thèse de doct. Stanford University.

Fred, Ana L.N et Anil K Jain (2003). « Robust data clustering ». In : *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*. T. 2. IEEE, p. 128–133. DOI : 10.1109/CVPR.2003.1211462.

Frisch, Stefan (1997). « Similarity and frequency in phonology ». Thèse de doct. Northwestern University.

Frisch, Stefan A., Janet B. Pierrehumbert et Michael B. Broe (2004). « Similarity avoidance and the OCP ». In : *Natural Language and Linguistic Theory* 22, p. 179–228.

Ganter, Bernhard et Rudolf Wille (1998). *Formal concept analysis : Mathematical foundations*. Springer. ISBN : 3540627715. DOI : 10.1007/978-3-642-59830-2.

Gilkerson, Jill et Jeffrey A. Richards (2009). *The Power of Talk, 2nd edition*. Rapp. tech. LENA Foundation. URL : https://www.lenafoundation.org/wp-content/uploads/2014/10/LTR-01-2_PowerOfTalk.pdf.

Ginzburg, Jonathan et Ivan A. Sag (2000). *Interrogative Investigations. The Form, Meaning, and Use of English Interrogatives*. Stanford : CSLI Publications.

Goldsmith, John et Jeremy O'Brien (2006). « Learning inflectional classes ». In : *Language Learning and Development*. T. 2. 4. Routledge, p. 219–250. DOI : 10.1207/s15473341lld0204_1.

Grönroos, Stig-Arne et al. (2014). « Morfessor FlatCat : An HMM-Based Method for Unsupervised and Semi-Supervised Learning of Morphology. » In : *Proceedings of the 25th International Conference on Computational Linguistics*. Dublin, Ireland : Association for Computational Linguistics, p. 1177–1185.

Grünwald, P.D. (2007). *Minimum Description Length Principle*. Cambridge : MIT press. ISBN : 978-0-262-07281-6.

Guzman Naranjo, Matías (2017). « Analogy in Formal Grammar ». Thèse de doct. University of Leipzig, Faculty of Philology.

Halle, M. et George N. Clements (1983). *Problem Book in Phonology : A Workbook for Introductory Courses in Linguistics and in Modern Phonology*. Bradford Books. MIT Press. ISBN : 9780262580595.

Hammarström, Harald et Lars Borin (2011). « Unsupervised Learning of Morphology ». In : *Computational Linguistics* 37.2. DOI : 10.1162/COLI_a_00050. URL : http://www.aclweb.org/anthology/J11-2002.

Hayes, Bruce (2012). *Spreadsheet with segments and their feature values*. Distributed as part of course material for Linguistics 120A : Phonology I at UCLA. URL : http://www.linguistics.ucla.edu/people/hayes/120a/index.htm.

Hockett, Charles F. (1954). « Two Models of Grammatical Description ». In : *Word* 10, p. 210–234.

— (1967). « The Yawelmani basic verb ». In : *Language* 43, p. 208–222.

— (1987). *Refurbishing our foundations*. Amsterdam : John Benjamins.

Kann, Katharina et Hinrich Schütze (2016). « MED : The LMU System for the SIGMORPHON 2016 Shared Task on Morphological Reinflection ». In : p. 62–70. DOI : 10.18653/v1/W16-2010. URL : http://www.aclweb.org/anthology/W16-2010.

Kari, James (1989). « Affix positions and zones in the Athapaskan verb complex : Ahtna and Navajo ». In : *International Journal of American Linguistics* 55, p. 424–454.

Kaufman, Leonard et Peter J. Rousseeuw (1990). *Finding Groups in Data : An Introduction to Cluster Analysis*. 1re éd. Wiley Series in Probability and Statistics. Wiley-Interscience. ISBN : 9780471735786,0-47-1-73578-7.

Kaufman, Terrence (1989). « The phonology and morphology of Zapotec verbs ».

Kilani-Schoch, Marianne et Wolfgang Dressler (2005). *Morphologie naturelle et flexion du verbe français*. Tübingen : Gunter Narr Verlag.

Kirov, Christo et al. (2016). « Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms ». In : *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Sous la dir. de Nicoletta Calzolari (Conference Chair) et al. Portorož, Slovenia : European Language Resources Association (ELRA). ISBN : 978-2-9517408-9-1. URL : http://ckirov.github.io/UniMorph/.

Kuryłowicz, Jerzy (1945). « La nature des procès dits «analogiques» ». In : *Acta linguistica* 5.1, p. 15–37. DOI : 10.1080/03740463.1945.10410880.

Lee, Jackson L. et John A. Goldsmith (2016). « Linguistica 5 : Unsupervised Learning of Linguistic Structure ». In : *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*. San Diego, California : Association for Computational Linguistics.

Lee, Jackson et John A. Goldsmith (2013). « Automatic morphological alignment and clustering ». Presented at the 2nd American International Morphology Meeting.

Levenshtein, Vladimir I (1966). « Binary codes capable of correcting deletions, insertions, and reversals ». In : *Soviet physics doklady* 10.8, p. 707–710.

Malouf, Robert (2017). « Abstractive morphological learning with a recurrent neural network ». In : *Morphology*. ISSN : 1871-5656. DOI : 10.1007/s11525-017-9307-x.

Mateus, Maria Helena et Ernesto d'Andrade (2000). *The Phonology of Portuguese*. Oxford : Oxford University Press.

Matthews, P. H. (1965). « The inflectional component of a word-and-paradigm grammar ». In : *Journal of Linguistics* 1, p. 139–171.

— (1972). *Inflectional Morphology. A Theoretical Study Based on Aspects of Latin Verb Conjugation*. Cambridge : Cambridge University Press.

— (1974). *Morphology*. Cambridge : Cambridge University Press.

— (1991). *Morphology*. 2nd. Cambridge : Cambridge University Press.

McDonough, J. (2000). « Athabaskan redux : Against the position class as a morphological category ». In : *Morphological Analysis in Comparison*. Sous la dir. de Dressler et al. Amsterdam : John Benjamins, p. 155–78.

McDonough, Joyce (1999). « On a bipartite model of the Athabaskan verb. » In : *The Athabaskan Languages : Perspectives on a Native American Language Family*. Sous la dir. de T. B. Fernald et P. R. Platero. Oxford Studies in Anthropological Linguistics. Oxford University Press, p. 139–166.

— (2003). *The Navajo Sound System*. T. 55. Studies in Natural Language and Linguistic Theory. Springer Netherlands. DOI : 10.1007/978-94-010-0207-3.

McDonough, Joyce M. (2014). « The Dene verb : how phonetics supports morphology ». In : *Proceedings of 18th Workshop on Structure and Constituency in the Languages of the Americas*. University of California. Berkeley.

McDonough, Joyce Mary (1990). « Topics in the phonology and morphology of Navajo verbs ». Thèse de doct. University of Massachusetts Amherst. URL : https://scholarworks.umass.edu/dissertations/AAI9110184/.

McDonough, Joyce et Valerie Wood (2008). « The stop contrasts of the Athabaskan languages ». In : *Journal of Phonetics* 36.3. Phonetic Studies of North American Indigenous Languages, p. 427–449. ISSN : 0095-4470. DOI : 10.1016/j.wocn.2007.11.001.

New, B., C. Pallier et al. (2001). « Une base de données lexicales du français contemporain sur internet : LEXIQUE ». In : *L'Année Psychologique* 101, p. 447–462.

New, Boris, Marc Brysbaert et al. (2007). « The use of film subtitles to estimate word frequencies ». In : *Applied Psycholinguistics* 28, p. 661–677.

Nicolai, Garrett, Colin Cherry et Grzegorz Kondrak (2015). « Inflection Generation as Discriminative String Transduction ». In : *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. Denver, Colorado : Association for Computational Linguistics, p. 922–931. DOI : 10.3115/v1/N15-1093. URL : http://www.aclweb.org/anthology/N15-1093.

Osswald, Rainer et Wiebke Petersen (2002). « Induction of classifications from linguistic data ». In : *Proc. of ECAI'02 Workshop on Advances in Formal Concept Analysis for Knowledge Discovery in Databases*.

— (2003). « A Logical Approach to Data-Driven Classification ». In : *Proceedings of the 26th Annual German Conference on Advances in Artificial Intelligence, KI 2003*. T. 2821. LNCS. Springer, p. 267–281.

Petersen, Wiebke (2001). « A Set-Theoretical Approach for the Induction of Inheritance Hierarchies ». In : *Proceedings of the Joint Conference on Formal Grammar and Mathematics of Language (FG/MOL-01)*. T. 53. Electronic Notes in Theoretical in Computer Science. Elsevier, p. 296–308.

Plénat, Marc (1987). « Morphologie du passé simple et du passé composé des verbes de l' "autre" conjugaison ». In : *ITL Review of Applied Linguistics* 77–78, p. 93–150.

Pollard, Carl et Ivan A. Sag (1994). *Head-driven Phrase Structure Grammar*. Stanford : CSLI Publications ; The University of Chicago Press.

Rasch, Jeffrey Walter (2002). « The Basic Morpho-syntax of Yaitepec Chatino ». under the supervision of Philip W. Davis. Thèse de doct. Rice University.

Rissanen, J. (1984). « Universal coding, information, prediction, and estimation ». In : *IEEE Transactions on Information Theory* 30.4, p. 629–636. DOI : 10.1109/TIT.1984.1056936.

Rissanen, Jorma (1978). « Modeling by shortest data description ». In : *Automatica* 14 (5), p. 465–658. DOI : 10.1016/0005-1098(78)90005-5.

Robins, R. H. (1959). « In defense of WP ». In : *Transactions of the Philological Society* 58, p. 116–144.

Rousseeuw, Peter J. (1987). « Silhouettes : A graphical aid to the interpretation and validation of cluster analysis ». In : *Journal of Computational and Applied Mathematics* 20, p. 53–65. ISSN : 0377-0427. DOI : doi.org/10.1016/0377-0427(87)90125-7. URL : http://www.sciencedirect.com/science/article/pii/0377042787901257.

Sagot, Benoît (2013). « Comparing Complexity Measures ». In : *Computational approaches to morphological complexity*. Paris, France. URL : https://hal.inria.fr/hal-00927276.

Sagot, Benoît et Géraldine Walther (2011). « Non-canonical inflection : data, formalisation and complexity measures. » In : *Systems and Frameworks in Computational Morphology*. Sous la dir. de Cerstin Mahlow et Michael Piotrowski. T. 100. Zurich, Switzerland : Springer-Verlag, p. 23–45.

Sagot, Benoît (2010). « The *Lefff*, a freely available and large-coverage morphological and syntactic lexicon for French ». In : *Proceedings of LREC 2010*.

Sagot, Benoît et Géraldine Walther (2013). « Implementing a formal model of inflectional morphology ». In : *Proceedings of Systems and Frameworks in Computational Morphology*, p. 115–134.

Santos, Diana et Paulo Rocha (2001). « Evaluating CETEMPúblico, a free resource for Portuguese ». In : *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, p. 442–449.

Sapir, Edward et Harry Hoijer (1967). *The phonology and morphology of the Navaho language*. T. 50. University of California publications in linguistics. Berkele : University of California Press.

Shannon, Claude E. (1948). « A mathematical theory of communication ». In : *Bell System Technical Journal* 27, p. 379–423, 623–656.

Sims, Andrea et Jeff Parker (2016). « How inflection classes work : On the informativity of implicative structure ». In : *Word Structure* 9.2, p. 215–239. DOI : 10.3366/word.2016.0094.

Sokal, Robert R. et Charles D. Michener (1958). « A Statistical Method for Evaluating Systematic Relationships ». In : *The University of Kansas Scientific Bulletin* 38.1409–1438.

Sorokin, Alexey (2016). « Using longest common subsequence and character models to predict word forms ». In : p. 54–61. DOI : 10.18653/v1/W16-2009. URL : http://www.aclweb.org/anthology/W16-2009.

Spencer, Andrew (2012). « Identifying stems ». In : *Word Structure* 5.1, p. 88–108. DOI : 10.3366/word.2012.0021.

Stump, Gregory T. (2001). *Inflectional Morphology. A Theory of Paradigm Structure*. Cambridge : Cambridge University Press.

Stump, Gregory T. et Raphael Finkel (2013). *Morphological Typology : From Word to Paradigm*. Cambridge : Cambridge University Press.

Taji, Dima et al. (2016). « The Columbia University - New York University Abu Dhabi SIGMORPHON 2016 Morphological Reinflection Shared Task Submission ». In : p. 71–75. DOI : 10.18653/v1/W16-2011. URL : http://www.aclweb.org/anthology/W16-2011.

Veiga, Arlindo, Sara Candeias et Fernando Perdigão (2013). « Generating a pronunciation dictionary for European Portuguese using a joint-sequence model with embedded stress assignment ». In : *Journal of the Brazilian Computer Society* 19.2, p. 127–134. ISSN : 0104-6500. DOI : 10.1007/s13173-012-0088-0.

Virpioja, Sami et al. (2013). *Morfessor 2.0 : Python Implementation and Extensions for Morfessor Baseline*. Rapp. tech., p. 38. URL : http://urn.fi/URN:ISBN:978-952-60-5501-5.

Wagner, Silke et Dorothea Wagner (2007). *Comparing clusterings : an overview*. Rapp. tech. URL : http://www.cs.ucsb.edu/~veronika/MAE/wagner07comparingclusterings.pdf.

Walther, Géraldine (2013). « On canonicity in morphology :an empirical, formal and computational approach ». Thèse de doct. Université Paris Diderot, École doctorale de sciences du langage 132, U.F.R. de linguistique.

— (2016). « Paradigm Realisation and the Lexicon ». In : *Morphological paradigms and functions*. Sous la dir. de Ferenc Kiefer, James P. Blevins et Huba Bartos. Leiden, Pays-Bas : Brill.

Walther, Géraldine et Benoît Sagot (2011). « Modélisation et implémentation de phénomènes flexionnels non-canoniques ». In : *Traitement Automatique des Langues* 52.2, p. 91–122.

Wurzel, Wolfgang Ulrich (1989). *Inflectional Morphology and Naturalness*. Dordrecht : Kluwer.

Young, Robert W. et William Morgan (1987). *The Navajo Language : A Grammar and Colloquial Dictionary*. Revised edition. Albuquerque : University of New Mexico Press.