Modélisation et implémentation de phénomènes flexionnels non canoniques

Géraldine Walther* — Benoît Sagot**

** Laboratoire de Linguistique Formelle, CNRS & Université Paris 7*
*** ALPAGE, INRIA Paris–Rocquencourt & Université Paris 7*
geraldine.walther@linguist.jussieu.fr, benoit.sagot@inria.fr

RÉSUMÉ. *Les phénomènes flexionnels non canoniques (déponence, hétéroclise…) font l'objet de nombreux travaux en morphologie théorique. Toutefois, ces travaux manquent souvent d'implémentations associées à des lexiques à grande échelle, pourtant nécessaires pour comparer objectivement la complexité de descriptions morphologiques. Nous montrons comment* ᚠᚨᚱᛋᛚ*, notre modèle de la morphologie flexionnelle, permet de représenter ces phénomènes non canoniques et de les formaliser en vue d'une implémentation. Nous l'illustrons au moyen de données de langues variées. Nous évaluons la complexité de quatre modélisations morphologiques concurrentes pour les verbes du français grâce à la notion informationnelle de* longueur de description *et montrons que les concepts nouveaux de* ᚠᚨᚱᛋᛚ *réduisent la complexité des modélisations morphologiques par rapport à des modèles traditionnels ou plus récents.*

ABSTRACT. *Non-canonical inflection (deponency, heteroclisis…) is extensively studied in theoretical morphology. However, these studies often lack practical implementations associated with large-scale lexica. Yet these are precisely the requirements for objective comparative studies on the complexity of morphological descriptions. We show how* ᚠᚨᚱᛋᛚ*, our model of inflectional morphology, manages to represent many non-canonical phenomena and to formalise them in way allowing for their subsequent implementation. We illustrate it with data about a variety of languages. We expose experiments conducted on the complexity of four competing descriptions of French verbal inflection, which is evaluated using the information-theoretic concept of* description length*. We show that the new concepts introduced in* ᚠᚨᚱᛋᛚ *reduce the complexity of morphological descriptions w.r.t. both traditional or more recent models.*

MOTS-CLÉS : ᚠᚨᚱᛋᛚ*, flexion, structure de paradigme, canonicité, zone flexionnelle, schème flexionnel, schème de radicaux, complexité de description, MDL.*

KEYWORDS: ᚠᚨᚱᛋᛚ*, Inflectional Morphology, Paradigm Shape, Canonicity, Inflection Zone, Stem Zone, Inflection Pattern, Stem Pattern, Description Complexity, MDL.*

1. Introduction

Construire automatiquement toutes les formes des paradigmes flexionnels d'une langue est souvent considéré comme une tâche simple, résolue depuis longtemps pour la plupart des langues qui intéressent le domaine du traitement automatique des langues (TAL). À l'inverse, dans le domaine de la morphologie théorique, nombreux sont les travaux dans le cadre des approches dites lexicalistes lexématiques qui cherchent à modéliser et expliquer les phénomènes flexionnels, et notamment ceux qui sont dits non canoniques. Ainsi, le Surrey Morphology Group a travaillé sur des projets portant sur les notions de *syncrétisme* (1999-2002), *supplétion* (2000-2003), *déponence* (2004-2006) et *défectivité* (2006-2009). En 2003, G. G. Corbett publie son premier article sur la *Typologie Canonique* (Corbett, 2003), définissant ainsi les fondements d'une approche théorique dont le but est de capter la différence entre régularité et irrégularité dans les paradigmes flexionnels. Toutefois, les travaux en morphologie (formelle) sont parfois limités par le manque de formalisation complète et d'implémentation à grande échelle des concepts qu'ils manipulent, tant en termes de couverture morphologique que lexicale. De telles ressources sont pourtant nécessaires pour évaluer quantitativement la pertinence d'une approche particulière ou comparer différents modèles décrivant tout ou partie de la morphologie d'une langue donnée. Cette voie de recherche rejoint des travaux récents dont le but est de mesurer la *complexité* linguistique, et plus spécifiquement la complexité morphologique (Bane, 2008), qui ont des objectifs typologiques mais permettent également de comparer des descriptions linguistiques concurrentes à l'aide de mesures objectives (*cf.* section 5).

Dans cet article, nous adoptons une approche *Word and Paradigm* (*Mot et Paradigme*) de la morphologie (Hockett, 1954). Nous présentons le modèle ᚹᚪᚱᛋᛚᛁ (Walther, 2011c) qui permet d'encoder formellement les informations contenues respectivement dans le lexique et la description des règles de réalisation des formes en mettant l'accent sur la forme des paradigmes des lexèmes (du fragment) de la langue décrite. En modélisant la forme des paradigmes, il permet notamment de rendre compte d'une multiplicité de phénomènes flexionnels non canoniques comme la défectivité, la supplétion, l'hétéroclise, etc. Nous montrons la pertinence de ce modèle en en présentant une implémentation pour des données non canoniques de plusieurs langues ainsi que des résultats comparatifs sur la complexité de descriptions morphologiques reposant sur ᚹᚪᚱᛋᛚᛁ. Il est à noter que ᚹᚪᚱᛋᛚᛁ n'est pas en concurrence avec les modèles réalisationnels de type PFM (*Paradigm Function Morphology*, (Stump, 2001)) ou *Network Morphology* (Corbett et Fraser, 1993) : contrairement à ce type de modèles, il se situe principalement en *amont* des questions de réalisation de formes de surface. Il se concentre sur la répartition des informations morphologiques entre le lexique morphologique et la *grammaire morphologique* (ensemble structuré de règles de réalisation), et la structure de la grammaire morphologique, ainsi que sur la représentation formelle de la structure des paradigmes morphologiques.

Après une rapide présentation de l'état de l'art en morphologie computationnelle et formelle (en section 2), nous rappelons en un premier temps les définitions de divers phénomènes flexionnels non canoniques. Ces définitions sont illustrées par des don-

nées empruntées au français, au latin, à l'italien, au kurde sorani, au persan, au croate et au slovaque (*cf.* section 3). En section 4, nous présentons alors les principes fondamentaux du modèle formel de la morphologie flexionnelle PARSLI et montrons dans quelle mesure ce modèle permet de représenter tous ces phénomènes flexionnels non canoniques. Enfin, en section 5, nous soulignons la pertinence de PARSLI pour formaliser des descriptions morphologiques concurrentes parfaitement implémentables. Nous y prenons comme exemple le cas de la flexion verbale du français, qui a récemment bénéficié d'un regain d'intérêt dans les travaux de morphologie théorique et formelle, et affiche à elle seule plusieurs de ces phénomènes. Nous montrons comment une implémentation de PARSLI dans le formalisme lexical Alexina (Sagot, 2010) permet de définir une notion de *la complexité de descriptions morphologiques* en termes de théorie de l'information. Cette mesure de complexité prend en compte la grammaire morphologique et le lexique associé et s'appuie sur le principe de *longueur de description (Minimum Description Length)* (Rissanen, 1984). Nous étudions la complexité de quatre descriptions concurrentes de la flexion verbale du français et montrons que PARSLI permet non seulement d'encoder des descriptions existantes (Bonami et Boyé, 2006 ; Sagot, 2010), mais également de composer une description de la flexion verbale française qui se traduit par une complexité moindre.

2. État de l'art

2.1. *État de l'art en morphologie computationnelle*

En morphologie computationnelle, la flexion est principalement abordée par des approches par règles ou par des méthodes par apprentissage supervisé ou non supervisé, ces différentes approches pouvant être combinées entre elles (Tepper et Xia, 2010). La rapide présentation que nous faisons ici de ces travaux résume essentiellement l'excellent tour d'horizon publié dans (Bernhard, 2010).

Parmi les méthodes à base de règles, on trouve (i) des outils de *racinisation* comme l'algorithme de désuffixation de Porter (1980) ; (ii) des transducteurs pour l'analyse et la génération de phénomènes flexionnels comme les méthodes issues de la *Morphologie à deux niveaux* (*Two-level morphology*, (Koskenniemi, 1984)) qui utilise des transducteurs mettant en correspondance un niveau lexical profond et des formes de surface en applicant des transformations phonotactiques spécifiques, et les autres approches de « *morphologie à état fini* », *Finite State Morphology* (Beesley et Karttunen, 2003) ; (iii) des approches morphosémantiques, essentiellement utilisées pour le traitement de corpus spécialisés, telles que (Pratt et Pacak, 1969 ; Hahn *et al.*, 2003 ; Lovis *et al.*, 1995), le système DériF (Namer, 2009) et ses adaptations plus récentes pour le vocabulaire médical anglais (Deléger *et al.*, 2009) ou des approches multilingues (Cartoni, 2009).

Les approches du deuxième type, les méthodes par apprentissage supervisé, s'appuient sur des corpus d'apprentissage préannotés indiquant les résultats à obtenir par les outils développés (van den Bosch et Daelemans, 1999 ; Stroppa et Yvon, 2006).

Existent enfin les approches non supervisées. Ces approches ont l'avantage de pouvoir être utilisées pour traiter même les langues pour lesquelles aucune description linguistique préalable n'est disponible. On peut en citer quatre types :

– comparaison des graphèmes dans un corpus donné, que ce soit par *distance orthographique* (Baroni *et al.*, 2002), reconnaissance du plus long affixe commun (Jacquemin, 1997 ; Gaussier, 1999 ; Zweigenbaum et Grabar, 2000), test d'insertion de mots (Keshava, 2006 ; Demberg, 2007 ; Bernhard, 2007), ou par analogie (Lepage, 1998 ; Hathout, 2002 ; Moreau *et al.*, 2007) ;

– modèles de maximum d'entropie fondée sur l'hypothèse de Harris (Harris, 1955) ; ces méthodes sont essentiellement employées pour déterminer les frontières de morph(ème)es (Keshava, 2006 ; Bernhard, 2007 ; Spiegler *et al.*, 2009) ;

– méthodes probabilistes comme l'inférence bayésienne (Creutz et Lagus, 2002), les modèles bayésiens hiérarchiques (Snyder et Barzilay, 2008) ou encore les modèles probabilistes génératifs (Spiegler *et al.*, 2009) ;

– méthodes de segmentation par compression de données (Goldsmith, 2001 ; Creutz et Lagus, 2002) s'appuyant sur le principe de la longueur minimale de description, *Minimum Description Length (MDL)* (Rissanen, 1984), l'idée étant que la morphologie tend toujours à utiliser l'encodage le plus compact en s'appuyant sur les phénomènes réguliers de flexion (*cf.* section 5 pour plus de détails) ; les auteurs considèrent que cette régularité se répercute dans la distribution des radicaux par rapport à leurs exposants (Matthews, 1974), permettant ainsi une segmentation de corpus en des unités morph(ém)iques minimales.

Malgré l'existence de nombreuses approches en morphologie computationnelle, aucune d'entre elles ne semble explicitement traiter de la question de la complexité de descriptions morphologiques particulières.

2.2. *État de l'art en morphologie formelle*

Parmi les approches théoriques de la linguistique, certaines accordent un statut indépendant à la morphologie, d'autres non. Ces dernières, représentées notamment par (Chomsky et Halle, 1968 ; Lieber, 1992) et plus précisément par la *Morphologie Distribuée* (Halle et Marantz, 1993) s'opposent ainsi aux approches de type lexématiques et en particulier aux approches *Word and Paradigm* (*Mot et Paradigme*) (Hockett, 1954 ; Robins, 1959 ; Matthews, 1974 ; Aronoff, 1994 ; Stump, 2001). Les questions portant sur les régularités et irrégularités dans les paradigmes n'ont de sens que pour les approches apparentées à ces dernières. Ce travail adopte pas conséquent une approche lexicaliste de type *Word and Paradigm* et s'appuie notamment largement sur les notions de *radical* et d'*exposant* définies dans (Robins, 1959) et (Matthews, 1974)[1].

1. Le lecteur trouvera une synthèse des arguments en faveur d'une approche lexématique dans (Fradin, 2003).

Les questions portant sur la régularité en morphologie ne sont pas toujours traitées en tant que telles, explicitement. Même au sein des approches lexématiques, la notion de *régularité* ne bénéficie pas systématiquement d'un statut théorique à part entière (Stump, 2001). Néanmoins, certaines approches contemporaines, à l'instar d'observations psycholinguistiques telles que, par exemple, décrites dans (Pinker, 1999), s'intéressent explicitement aux questions de régularité dans les paradigmes. Ainsi Bonami et Boyé (2006) affirment que l'irrégularité est un phénomène grammatical à part entière qui se manifeste non seulement dans les comportements psycholinguistiques, mais également en diachronie ou en grammaire synchronique.

Les travaux sur la régularité et l'irrégularité des paradigmes[2] se retrouvent également dans le domaine de la Typologie Canonique introduite par Corbett (Corbett, 2003 ; Corbett, 2007b). Les phénomènes non canoniques tels que la *supplétion* (Boyé, 2006) (et les références qu'il cite), la *déponence* (Baerman *et al.*, 2007), l'*hétéroclise* (Stump, 2006), la *défectivité* (Baerman *et al.*, 2010) et plus récemment la *surabondance* (Thornton, 2010, sous le nom anglais de *overabundance*) ont été étudiés dans le cadre de cette approche et sont à l'origine de bon nombre de publications récentes[3].

Néanmoins, ces travaux visent rarement le développement de descriptions optimales en termes de compacité. Cela s'explique notamment par le fait qu'une évaluation de la compacité d'une description n'est possible que si la description s'accompagne d'une implémentation complète de cette description liée à un lexique à grande échelle. Il est nécessaire d'évaluer la complexité de la description dans son application à aux unités lexicales de la langue décrite, dans leur (quasi-)totalité. Une rapide présentation de l'état de l'art spécifique à la mesure de complexité d'une description est donnée en introduction de la section 5.

3. Données sur les phénomènes flexionnels non canoniques

Dans l'approche de type *Word and Paradigm* que nous adoptons, nous définissons une *description morphologique* comme la combinaison d'un ensemble d'entrées lexicales fléchissables et des règles de réalisation de formes exprimant chacune un ensemble particulier de traits morphosyntaxiques. Le résultat de l'ensemble des règles de réalisation d'une langue donnée correspond aux paradigmes des lexèmes de cette langue.

2. Régularité et irréguliarité y sont présentées comme relevant d'une forme de *complexité qualitative*.

3. Ces travaux prennent essentiellement, sinon exclusivement, en compte des données orales. Notre travail se situe ici sur des données de l'écrit. Nous projetons, dès que nous aurons une quantité de données suffisante à disposition, de prolonger ce travail par une étude de données similaires pour l'oral, afin notamment de comparer la complexité morphologique à l'oral et à l'écrit.

Dans le but de pouvoir mesurer la complexité d'une description morphologique particulière, nous commençons par identifier les phénomènes susceptibles d'augmenter la complexité des paradigmes. Ces phénomènes sont les phénomènes irréguliers, les cas ne pouvant pas être définis par défaut. Dans une approche de *Typologie Canonique* (Corbett, 2003), il s'agit des phénomènes non canoniques.

3.1. *Flexion canonique*

Le concept de *typologie canonique* apparaît avec les travaux de Corbett (2003) et la volonté de permettre une meilleure compréhension de ce qui, dans la constitution d'un paradigme, différencie un idéal hypothétique, canonique, des différentes réalisations concrètes de phénomènes moins canoniques. Dans une telle approche, il convient de ne pas confondre *flexion canonique* et *flexion prototypique*. La flexion canonique est rare, si ce n'est inexistante. Elle constitue un étalon théorique qui a pour seul objectif de servir de frontière pour en différencier les phénomènes non canoniques réels (Corbett, 2007a). Dans ce travail, nous considérons qu'un paradigme est canonique s'il satisfait les critères du tableau 1 indiqués dans (Corbett, 2007a). Nous ajoutons aux critères de Corbett les critères donnés dans le tableau 2[4] permettant de définir plus précisément la structure des paradigmes concernés. Tout écart par rapport à ces critères est considéré comme une instance de non-canonicité (de la structure) des paradigmes.

Dans la suite de ce travail, nous donnons une représentation formelle de cinq types de phénomènes flexionnels non canoniques : *supplétion, déponence, hétéroclise, défectivité* et *surabondance* dans le modèle ᚹᚪᚱᛋᛚᛁ (Walther, 2011c). À la section 5, nous montrons ensuite quelle peut être l'influence de ce type de phénomènes sur la complexité d'une description morphologique.

3.2. *Alternance de radicaux et supplétion*

Parmi les phénomènes qualifiés de supplétion, on différencie deux types : la supplétion de radicaux et la supplétion de formes (Boyé, 2006). La supplétion de radicaux concerne les cas où, à l'intérieur d'un paradigme donné, seuls les radicaux varient d'une manière non régulière. L'alternance des exposants[5] reste celle attendue dans le paradigme en question. Un exemple évident de supplétion de radicaux est fourni par le verbe *aller* pour lequel les descriptions indiquent jusqu'à quatre radicaux différents : *all-, v-, i-* et *aill-*. La supplétion de formes concerne des cas où une forme complète d'un paradigme est remplacée par une forme inattendue. Un exemple de supplétion de formes en français est donné dans (Bonami et Boyé, 2002) pour le verbe *être* au

4. Parmi ces critères additionnels, le critère 3 découle directement du critère 3 dans (Corbett, 2007a).

5. Au sens de Matthews (1974) : partie de la forme qui n'est pas le radical.

	COMPARAISON ENTRE CASES D'UN PARADIGME	COMPARAISON ENTRE LEXÈMES
COMPOSITION/STRUCTURE	*identiques*	*identiques*
MATÉRIAU LEXICAL (≈ radicaux)	*identiques*	*différents*
MATÉRIAU FLEXIONNEL (≈ flexions)	*différents*	*identiques*
RÉSULTAT (≈ formes fléchies)	*différents*	*différents*

Tableau 1. *Critères pour la flexion canonique selon Corbett*

	FLEXION CANONIQUE
RADICAL ET TRAITS	*Il n'y a pas de « décalage entre forme et fonction. » (Baerman, 2007)* *Chaque lexème a exactement un radical qui se combine avec un ensemble d'exposants.*
COMPLÉTUDE	*Chaque lexème dispose d'exactement une forme pour exprimer chacune des structures de traits morphosyntaxiques pertinentes pour sa catégorie.*
CLASSE FLEXIONNELLE	*Les formes d'un même lexème sont construites selon une classe flexionnelle unique.*

Tableau 2. *Critères supplémentaires pour la flexion canonique*

présent de l'indicatif (*cf.* tableau 3). Ici, la forme de la première personne du pluriel, par exemple, n'affiche pas l'exposant attendu que serait le *–ons*.

	SINGULIER	PLURIEL
P1	***suis***	***sommes***
P2	*es*	*êtes*
P3	*est*	***sont***

Tableau 3. *Supplétion de formes dans le paradigme du présent de l'indicatif du verbe français* être

LEXÈME	TRAD.	RAD1	RAD2
ÂRÂSTAN	'orner'	*ârâst*	*ârâ*
ÂMUXTAN	'apprendre'	*âmuxt*	*âmuz*
RAQSIDAN	'danser'	*raqsid*	*raqs*

Tableau 4. *Radicaux du présent (*RAD1*) et du passé (*RAD2*) en persan*

Par ailleurs, la supplétion peut se distribuer de manière plus ou moins transparente : elle peut plus ou moins fortement refléter des variations dans les traits morphosyntaxiques exprimés par des formes données. Ainsi, les langues iraniennes font usage d'alternance de radicaux pour exprimer une opposition entre les temps du passé et les temps du présent. Le persan utilise le radical RAD1 pour les temps du présent et le radical RAD2 pour les temps du passé. En plus de cette opposition claire, ces radicaux ont également d'autres fonctions : le RAD2 est employé pour former les infinitifs, le RAD1 pour les impératifs.

De même, les descriptions traditionnelles du latin (Ernout et Thomas, 1953) font état de trois radicaux verbaux distincts conditionnés par des traits morphosyntaxiques liés à des sous-paradigmes de leur système flexionnel. Il s'agit des radicaux du pré-

FORMATION DU RADICAL DU PASSIF	LEXÈME	RAD1	RADICAL DU PRÉSENT PASSIF	TRADUCTION
RAD1	KUŠTIN	*kuš*	*kuš–rê*	'tuer'
RAD2	ÛTIN	*l'ĕ*	*ût–rê*	'dire'
RAD2	BISTIN	*bîe*	*bist–rê*	'entendre'
RAD1 MOINS VOY. FIN.	KIRDIN	*ke*	*k–rê*	'faire'
RAD1 MOINS VOY. FIN.	DAN	*de*	*d–rê*	'donner'
AUTRE	XWARDIN	*xo*	*xû–rê*	'manger'
AUTRE	GIRTIN	*gir*	*gîr–rê*	'prendre'

Tableau 5. *Radicaux irréguliers du passif en kurde sorani*

sent, du passé et du supin : *amo* 'j'aime', *amāvī* 'j'ai aimé', *amātum* 'aimé'. Ces radicaux n'expriment pas dans leur répartition une réalisation morphologique de traits morphosyntaxiques régulière. Ainsi, le RAD3 peut également servir pour les formes du participe passé passif, le participe futur actif et pour les formes finies perfectives du passif. Il ne semble pas y avoir de trait morphosyntaxique unique qui implique nécessairement l'emploi du RAD3. Cette distribution synchroniquement immotivée, mais systématique d'un paradigme à l'autre, des radicaux du latin permet de dire qu'ils sont *morphomiques* au sens de (Aronoff, 1994).

La supplétion de radicaux peut également se montrer plus extensive. Ainsi, dans (Bonami et Boyé, 2003), les auteurs montrent pour la flexion verbale du français qu'il existe jusqu'à douze ensembles de combinaisons de traits morphosyntaxiques possibles qui se réalisent au moyen de radicaux spécifiques. L'inventaire de ces douze radicaux correspond à ce que les auteurs qualifient d'*espaces thématiques* ; les espaces thématiques sont reliés entre eux par des règles de *dépendance thématique*.

Même au sein des langues qui emploient les variations de radicaux pour exprimer des traits morphosyntaxiques spécifiques, des irrégularités locales peuvent encore apparaître. Ainsi, en kurde sorani, langue iranienne de l'ouest qui oppose le RAD1 du présent au RAD2 du passé, des irrégularités de selection de radicaux apparaissent pour la formation des radicaux du passif. Comme le montrent les grammaires de références (McCarus, 1958 ; Thackston, 2006) les formes du passifs sont généralement dérivées du RAD1. Cependant, il existe des verbes pour lesquels le RAD2 sert de base à la formation des passifs, tandis que d'autres verbes encore ont des radicaux du passif complètement irréguliers (*cf.* tableau 5) (Walther, 2011a).

3.3. *Déponence*

Les noms du croate emploient parfois des formes du singulier pour exprimer le pluriel (Baerman, 2006). Ce décalage entre la forme et la fonction est ce qu'en suivant Baerman (2007), nous appelons *déponence*. Les noms croates se fléchissent en nombre en suivant plusieurs classes flexionnelles distinctes. Parmi ces classes, nous nous intéressons ici plus précisément à celles présentées dans le tableau 6 :

	FÉMININ À RADICAL EN –A *žena* 'femme'		FÉMININ À RADICAL EN –I *stvar* 'chose'	
	SINGULIER	PLURIEL	SINGULIER	PLURIEL
NOM	*žen-a*	*žen-e*	*stvar*	*stvar-i*
ACC	*žen-u*	*žen-e*	*stvar*	*stvar-i*
GEN	*žen-e*	*žen-a*	*stvar-i*	*stvar-i*
DAT	*žen-i*	*žen-ama*	*stvar-i*	*stvar-ima*
INS	*žen-om*	*žen-ama*	*stvar-i*	*stvar-im*

Tableau 6. *Déclinaison de noms croates*

	NEUT. -ET∼A-STEM *dete* 'enfant'		NEUT. -ET∼I-STEM *tele* 'veau'	
	SINGULIER	PLURIEL	SINGULIER	PLURIEL
NOM	*dete*	*deca*	*tele*	*telad*
ACC	*dete*	*decu*	*tele*	*telad*
GEN	*deteta*	*dece*	*teleta*	*telad*
DAT	*detetu*	*deci*	*teletu*	*teladi(ma)*
INS	*detetom*	*decom*	*teletom*	*teladi(ma)*

Tableau 7. *Déclinaison de déponents croates*

les noms *dete* 'enfant' et *tele* 'veau' se fléchissent au pluriel respectivement selon les patrons du singulier des noms à RADICAUX EN –A et RADICAUX EN –I. Le fait d'employer le singulier pour exprimer le pluriel constitue un décalage entre forme et fonction (Baerman, 2007).

Le Surrey Morphology Group a mis en ligne une base de données étendue de phénomènes de déponence[6]. Nous ne retenons pas tous les phénomènes qui y sont référencés comme des phénomènes de déponence dans le modèle que nous présentons en section 4. Cette base de données constitue néanmoins une excellente présentation de divers phénomènes de déponence observables dans les langues du monde.

L'exemple de déponence qui a probablement été le plus décrit et discuté est sans doute le cas des verbes déponents latins. Ils sont habituellement considérés comme des verbes employant une forme passive pour exprimer un sens ou une valeur syntaxique actifs (Kiparsky, 2005 ; Hippisley, 2007 ; Baerman, 2007 ; Corbett, 2007b). Dans le cas particulier des « déponents latins », nous donnons une analyse concurrente de ces verbes à la section 4 dans laquelle nous montrons que ces verbes ne sont justement pas des instances de déponence mais plutôt un exemple type de ce que l'on appelle l'hétéroclise (Walther, 2011b).

3.4. *Hétéroclise*

L'hétéroclise désigne le phénomène où le paradigme d'un lexème se compose de parties d'au moins deux classes flexionnelles distinctes.

6. `http://www.smg.surrey.ac.uk/deponency`.

	MASCULIN ANIMÉ *chlap* 'garçon'		MASCULIN INANIMÉ *dub* 'chêne'		MASCULIN HÉTÉROCLITE *orol* 'aigle'	
	SINGULIER	PLURIEL	SINGULIER	PLURIEL	SINGULIER	PLURIEL
NOM	*chlap*	*chlap-i*	*dub*	*dub-**y***	*orol*	*orl-**y***
GEN	*chlap-a*	*chlap-ov*	*dub-a*	*dub-ov*	*orl-a*	*orl-ov*
DAT	*chlap-**ovi***	*chlap-om*	*dub-u*	*dub-om*	*orl-**ovi***	*orl-om*
ACC	*chlap-a*	*chlap-ov*	*dub*	*dub-**y***	*orl-a*	*orl-**y***
LOC	*chlap-**ovi***	*chlap-och*	*dub-e*	*dub-och*	*orl-**ovi***	*orl-och*
INS	*chlap-om*	*chlap-mi*	*dub-om*	*dub-mi*	*orl-om*	*orl-ami*

Tableau 8. *Hétéroclise dans les paradigmes des noms d'animaux slovaques*

Un exemple d'hétéroclise est fourni par certains des noms d'animaux du slovaque. En slovaque, la plupart des noms d'animaux masculins se fléchissent comme des noms masculins animés au singulier, mais peuvent (voire doivent, pour certains lexèmes) être fléchis comme des inanimés au pluriel (sauf dans des cas particuliers de personnification explicite) (Zauner, 1973). On peut comparer, par exemple, la flexion de *chlap* 'garçon', *dub* 'chêne' et *orol* 'aigle' (*cf.* tableau 8)[7].

3.5. *Défectivité*

La défectivité désigne les cas où un lexème possède dans son paradigme des cases vides (ou manquantes). Parfois, les langues possèdent des lexèmes pour lesquels des formes pourtant attendues n'existent pas ; les locuteurs sont dans l'incapacité de les produire. Quand, pour autant, ces formes manquantes deviennent nécessaires pour produire un énoncé, les locuteurs doivent avoir recours à des stratégies de réparation par l'emploi de formes synonymes qui prennent alors le relais. Cette stratégie s'applique notamment aux cas des verbes latins dits *activa tantum*, à savoir des verbes transitifs pour lesquels il n'existe pas de formes passives. Le latin emprunte donc des formes synonymes à d'autres lexèmes pour exprimer la valeur du passif. Parmi ces verbes figurent notamment *facere* 'faire' et *perdire* 'détruire' qui ne possèdent pas de morphologie passive au présent. Les formes manquantes sont puisées dans les paradigmes des verbes transitifs actifs *fieri* 'devenir' et *perire* 'périr' respectivement (Kiparsky, 2005)[8]. Un autre exemple de défectivité est celui des *pluralia tantum* pour lesquels il n'existe que des formes du pluriel, *cf. vivres* en français, *trousers* 'pantalon' en anglais, ou encore *Vianoce* 'Noël' en slovaque.

7. *Chlap* et *dub* se fléchissent de façon régulière : *chlap* appartient à la classe de flexion par défaut des noms masculins animés qui se terminent par une consonne, tandis que *dub* appartient à la classe standard pour les inanimés masculins se terminant par ce qu'on désigne comme des consonnes *dures* ou *neutres* dans la tradition linguistique slave.

8. Ces verbes de remplacement ne fonctionnent pas uniquement comme formes passives de *facere* et *perdire*. Ils sont également des verbes transitifs indépendants à part entière et ne peuvent pas, par conséquence, être simplement traités comme des formes supplétives dans les paradigmes de *facere* et *perdire*. Ils constituent des entrées lexicales indépendantes.

LEXÈME	TRAITS MORPHOSYNTAXIQUES	COMPAGNON 1	COMPAGNON 2
'languir'	3PL.PRS.SUBJ	*languano*	*languiscano*
'posséder'	3PL.PRS.SUBJ	*possiedano*	*posseggano*
'posséder'	3SG.PRS.SUBJ	*possieda*	*possegga*
'posséder'	1SG.PRS.SUBJ	*possiedo*	*posseggo*

Tableau 9. *Surabondance en italien*

3.6. *Surabondance*

La contrepartie à la défectivité est la notion de *surabondance*. On parle de *surabondance* à chaque fois qu'une case d'un paradigme donné contient plus d'une forme. La notion de surabondance a été introduite par Thornton (2010) pour l'italien. La surabondance concerne canoniquement les cas où deux éléments d'une même case (*cell mates*) sont en compétition pour la réalisation des traits associés, sans qu'aucun trait particulier ne permette de choisir l'un plutôt que l'autre. Le tableau 9 montre des exemples de surabondance présentés par Thornton pour l'italien.

Un exemple en français serait celui du verbe *asseoir* qui possède deux formes distinctes dans la majorité de ses cases[9] (*cf.* le tableau 10). De même, tous les verbes français en *-ayer* affichent une surabondance systématique (*cf.* le tableau 11). En effet, dans certaines cases, ces verbes affichent deux radicaux concurrents, à savoir l'un en *-ay-* et l'autre en *-ai-*. Ils ont donc à chaque fois deux formes fléchies concurrentes, équivalentes morphologiquement (même s'il peut y avoir dans ces cas des différences d'ordre sémantique, pragmatique, sociolectales ou autres).

IND.PRES	SINGULIER	PLURIEL
P1	*assois* *assieds*	*assoyons* *asseyons*
P2	*assois* *assieds*	*assoyez* *asseyez*
P3	*assoit* *assied*	*assoient* *asseyent*

Tableau 10. *Surabondance en français pour le verbe* asseoir

IND.PRES	SINGULIER	PLURIEL
P1	*balaye* *balaie*	*balayons*
P2	*balayes* *balaies*	*balayez*
P3	*balaye* *balaie*	*balayent* *balaient*

Tableau 11. *Surabondance en français pour le verbe* balayer

4. Modèle formel pour la morphologie flexionnelle

4.1. *Définition des notions principales*

Dans la suite, nous nous intéressons à la formalisation (implémentable) des phénomènes décrits ci-dessus. Dans la mesure où ce sont les phénomènes flexionnels

9. Voir (Bonami et Boyé, 2010) pour plus de détails à ce sujet.

irréguliers comme ceux décrits en 3 qui sont le plus susceptibles de faire varier la complexité de descriptions morphologiques, nous avons besoin, pour mesurer et comparer la complexité de descriptions concurrentes, d'avoir recours à un modèle capable de capter entièrement ces irrégularités. Un tel modèle formel de la morphologie est le modèle ℙ𝔸ℝ𝕊𝕃, décrit en détail dans (Walther, 2011c). Il permet d'encoder à la fois les entrées lexicales et leurs spécifications lexicalisées (*cf.* la présence de radicaux supplétifs) et les règles de réalisation de formes proprement dites, en passant par la structure des paradigmes.

Lorsque nous parlons de *description morphologique* nous appelons *lexique morphologique* l'ensemble des entrées lexicales contenant les informations morphologiques qui leur sont considérées comme spécifiques. Nous appelons *grammaire morphologique* l'ensemble des informations morphologiques indépendantes de l'entrée lexicale considérée (dans un modèle simple, il s'agirait notamment des définitions des classes flexionnelles). Une *description morphologique*, qui doit permettre la production effective de l'ensemble des formes fléchies des lexèmes considérés, est alors la combinaison d'un lexique et d'une grammaire morphologiques.

ℙ𝔸ℝ𝕊𝕃 ne s'intéresse aux lexèmes qu'en tant qu'ils participent au processus de formation de l'ensemble des formes données d'une langue. Il ne s'agit donc pas d'encoder les différences sémantiques, syntaxiques ou de propriétés en termes de morphologie constructionnelle pouvant intervenir entre deux lexèmes homonymes. Autrement dit, nous nous intéressons ici à la notion de *flexème* telle que définie dans (Fradin et Kerleroux, 2003). Cette notion se différencie de la notion de *lexème* ; un flexème pourrait être grossièrement défini comme *un lexème privé de ses dimensions sémantiques et de sa spécification en termes de structure argumentale*.

ℙ𝔸ℝ𝕊𝕃 laisse (pour l'instant) sous-spécifié le dispositif utilisé pour la réalisation proprement dite des formes. Il se concentre sur la répartition des informations morphologiques entre le lexique morphologique et la grammaire morphologique, et la structure de la grammaire morphologique. La réalisation concrète des formes est prévue pour être représentée par des règles de réalisation qui peuvent tout à fait être exprimées par des formalismes indépendants de morphologie réalisationnelle. Ce modèle se situe donc essentiellement en amont du cœur de théories comme PFM (Stump, 2001) ou *Network Morphology* (Corbett et Fraser, 1993).

4.2. *Présentation de* ℙ𝔸ℝ𝕊𝕃

ℙ𝔸ℝ𝕊𝕃 (Walther, 2011c) tient son nom de « ℙ𝔸ℝ*adigm* 𝕊*hape and* 𝕃*exicon* 𝕀*nterface* ». Il s'agit d'un modèle formalisé[10] de la morphologie flexionnelle qui permet de structurer conjointement la grammaire et le lexique morphologiques d'une façon qui rende compte du degré de canonicité des paradigmes que comporte la langue

10. Dans cette section, les définitions formelles sont indiquées typographiquement par des boîtes. Le lecteur moins à l'aise avec la formalisation pourra sauter les passages formels sans pour autant perdre de vue le sens général de l'article.

décrite : le modèle modélise spécifiquement le flexème dans sa relation avec le paradigme auquel il est associé.

Dans PARSLI, une classe flexionnelle est définie comme étant une fonction qui associe à des structures de traits morphosyntaxiques des règles de réalisation, c'est-à-dire une façon d'appliquer les exposants correspondant à chaque structure de traits donnée. De façon informelle[11], on peut dire qu'une classe flexionnelle est partitionnée en une ou plusieurs *zones flexionnelles*[12], qui sont utilisées par PARSLI pour modéliser la flexion.

De même, PARSLI définit la notion de *classe de radicaux*, qui est une fonction qui associe à des structures de traits des règles de formation de radicaux. Une classe de radicaux est elle aussi partitionnée en *zones de radicaux*.

La construction d'une forme correspondant à une structure de traits particulière pour un flexème donné peut alors se décrire comme suit. Les informations associées au flexème indiquent tout d'abord, pour cette structure de traits, quelle est la zone de radicaux à utiliser parmi l'ensemble des zones incluses dans l'ensemble des classes de radicaux. Cette zone de radicaux indique la façon dont le radical de la forme recherchée doit être construit. Parallèlement, les informations associées au flexème indiquent, pour la structure de traits étudiée, quelle est la zone flexionnelle à utiliser. Intervient alors une *règle de transfert*, qui fait partie de la définition du flexème : elle prend en entrée la structure de traits étudiée, et donne en sortie la structure de traits à prendre en considération pour obtenir la règle de réalisation correcte au sein de la zone flexionnelle. En général, cette règle de transfert est la fonction identité, mais nous verrons que ça n'est pas toujours le cas, et qu'il peut y avoir décalage entre la forme et la fonction (Baerman, 2007). Enfin, on applique alors sur le radical construit préalablement les règles de réalisation ainsi obtenues.

En réalité, PARSLI prévoit également une règle de transfert (appartenant également à la définition des flexèmes) pour la sélection des radicaux. Cette règle permet de modéliser des cas où un radical, habituellement prévu pour la réalisation de certaines formes d'un paradigme, se trouve être utilisé dans d'autres formes, habituellement construites à partir d'un autre radical. Cette règle de transfert est le plus souvent la fonction identité[13].

Ainsi, outre sa catégorie et l'ensemble des traits morphosyntaxiques qu'il peut exprimer par l'une de ses formes, PARSLI associe à un flexème $\mathfrak{I}$ les informations suivantes :

11. Une définition formelle de la notion de *zone flexionnelle* est donnée dans le second encart de la section 4.2.

12. Une fonction étant un ensemble de couples, la notion de partition peut s'y appliquer comme à tout type d'ensemble.

13. Nous ne détaillons pas dans la suite de cet article les cas de décalages entre forme et fonction dans la formation des radicaux. La règle de transfert pour les radicaux est néanmoins explicitée dans (Walther, 2011c). Des travaux empiriques sont également en cours (Walther, en préparation).

– une *schème radical*, relation binaire qui associe à chaque ensemble pertinent de traits morphosyntaxiques une zone de radicaux ;

– une *schème flexionnel*, relation binaire qui associe à chaque ensemble pertinent de traits morphosyntaxiques une zone flexionnelle ;

– une première *règle de transfert* qui associe à l'ensemble de traits morphosyntaxiques exprimés par la forme à construire un ensemble de traits morphosyntaxiques souvent identique qui sera à utiliser pour la construction effective de la forme ;

– une seconde *règle de transfert* qui associe à l'ensemble de traits morphosyntaxiques exprimés par le radical à construire un ensemble de traits morphosyntaxiques souvent identique qui sera à utiliser pour la formation effective du radical.

Les classes flexionnelles sont les ensembles de zones flexionnelles les plus naturels : elles correspondent à des schèmes flexionnels par défaut, c'est-à-dire qui concernent une majorité de flexèmes. Il en va de même pour les classes de radicaux.

Il se peut, et nous en verrons des exemples, qu'une même structure de traits soit associée par un schème radical à plus d'une zone de radicaux différentes, et par un schème flexionnel à plus d'une zone flexionnelle différentes. Or, dans ce cas, rien ne garantit que chacune des zones de radicaux puisse être utilisée avec chacune des zones flexionnelles. La situation se complique dès lors que l'on introduit en plus les règles de transfert. Il faut donc expliciter les couplages possibles. Nous appelons donc *sous-schème* un quadruplet formé d'une zone de radicaux, d'une zone flexionnelle, et de deux règles de transfert. Un sous-schème indique donc un couplage possible entre zone flexionnelle et zone de radicaux ; il n'est cohérent que si les traits morphosyntaxiques auxquels s'appliquent les deux types de zones ont une intersection non vide. L'ensemble des sous-schèmes d'un flexème est appelé son *schème*. Nous verrons des exemples concrets de sous-schèmes et de schèmes à la section 5.

L'ensemble des formes construites par les règles de formation de radicaux et les règles de réalisation obtenues pour un flexème donné à partir de son schème pour tous les traits morphosyntaxiques qu'il est capable d'exprimer constitue le paradigme de ce flexème.

Nous notons $\varkappa$ une structure de traits morphosyntaxiques, f un schème flexionnel, $\digamma$ une classe flexionnelle, ξ une zone flexionnelle. Un schème radical est notée s, les classes de radicaux Γ et une zone de radicaux ζ. Une règle de transfert est souvent notée $\mathscr{T}$. Enfin, un schème est généralement noté P.

Définition formelle d'un flexème
Un flexème $\mathfrak{I}$ est un septuplet ($\mathscr{K}_\mathfrak{I}$, $\mathcal{C}_\mathfrak{I}$, $\mathrm{s}_\mathfrak{I}$, $\mathrm{f}_\mathfrak{I}$, $\mathscr{T}_{\psi_\mathfrak{I}}$, $\mathscr{T}_{\chi_\mathfrak{I}}$, $P_\mathfrak{I}$), où

• $\mathscr{K}_\mathfrak{I}$ est l'ensemble des traits morphosyntaxiques $\varkappa$ pertinents pour $\mathfrak{I}$;

• $\mathcal{C}_\mathfrak{I}$ est la catégorie morphosyntaxique de $\mathfrak{I}$, et $\mathcal{C}_\mathfrak{I} \in \mathscr{C}$, l'ensemble des catégories existantes dans une langue donnée, noté $\mathscr{C}$;

• $\mathrm{s}_\mathfrak{I}$ est un *schème radical*, relation binaire de $\mathscr{K}_\mathfrak{I}$ vers $\mathcal{Z}_{\mathrm{s}_\mathfrak{I}}$, l'ensemble des *zones de radicaux* compatibles avec $\mathfrak{I}$; $\mathcal{Z}_{\mathrm{s}_\mathfrak{I}} \subset \mathscr{Z}$, l'ensemble des zones de radicaux d'une langue, noté $\mathscr{Z}$;

• $\mathrm{f}_\mathfrak{I}$ est un *schème flexionnel*, relation binaire de $\mathscr{K}_\mathfrak{I}$ vers $\mathcal{X}_{\mathrm{f}_\mathfrak{I}}$, l'ensemble des *zones flexionnelles* selon lesquelles est fléchi un flexème donné ($\mathcal{X}_{\mathrm{f}_\mathfrak{I}} \subset \mathscr{X}$, l'ensemble des zones flexionnelles d'une langue, noté $\mathscr{X}$) ;

• $\mathscr{T}_{\psi_\mathfrak{I}}$ est une *règle de transfert* telle que $\forall \varkappa \in \mathscr{K}_\mathfrak{I}$, où $\mathscr{T}_{\psi_\mathfrak{I}}(\varkappa)$ appartient à l'ensemble des traits morphosyntaxiques réalisés dans les zones de radicaux définies pour $\mathscr{K}_\mathfrak{I}$;

• $\mathscr{T}_{\chi_\mathfrak{I}}$ est une *règle de transfert* telle que $\forall \varkappa \in \mathscr{K}_\mathfrak{I}$, où $\mathscr{T}_{\chi_\mathfrak{I}}(\varkappa)$ appartient à l'ensemble des traits morphosyntaxiques réalisés dans les zones flexionnelles définies pour $\mathscr{K}_\mathfrak{I}$;

• $P_\mathfrak{I}$ est un schème, c'est-à-dire un ensemble de sous-schèmes définis comme ci-dessus.

Une zone flexionnelle étant définie comme étant l'un des éléments d'une partition d'une classe flexionnelle, on peut définir l'espace de définition d'une zone flexionnelle ξ, noté $\hat{\xi}$. Par construction, l'ensemble des $\hat{\xi}$ des zones flexionnelles d'une même classe flexionnelle forme lui-même une partition de l'ensemble de définition de la classe flexionnelle, c'est-à-dire de l'ensemble des traits morphosyntaxiques pour lesquels cette classe est définie.
Enfin, pour une zone flexionnelle ξ, nous notons $\tilde{\xi}$ la classe flexionnelle dont elle est l'un des éléments de la partition.

4.3. *Modélisation de cas spécifiques de flexion non canonique*

Le modèle précédemment décrit nous permet de formaliser les données de phénomènes non canoniques présentées en section 3.

4.3.1. *Supplétion et sélection de radicaux*

Étant donné un flexème $\mathfrak{I}$, son schème radical $\mathrm{s}_\mathfrak{I}$ associe à une structure de traits morphosyntaxiques $\varkappa$ (au moins) une zone de radicaux $\zeta_{\mathrm{s}_\mathfrak{I},\varkappa}$. Pour l'instant, nous n'avons pas rencontré de situation où une zone de radicaux n'était pas une fonction constante par rapport à la structure de traits qu'on lui donne en entrée. Autrement dit, pour un flexème donné, nous pouvons simplifier les notations et assimiler $\zeta_{\mathrm{s}_\mathfrak{I},\varkappa}$ à la valeur constante qu'elle donne en sortie.

Ainsi, les radicaux du passif en sorani (tableau 5) sont sélectionnés comme suit :

$$\text{S}_{\text{KUŠTIN}}(\{\text{MODE } \textit{passif}, \text{TEMPS } \textit{présent}\}, \text{kuš–rê})$$

$$\text{S}_{\text{KIRDIN}}(\{\text{MODE } \textit{passif}, \text{TEMPS } \textit{présent}\}, \text{ke–rê})$$

$$\text{S}_{\text{XWARDIN}}(\{\text{MODE } \textit{passif}, \text{TEMPS } \textit{présent}\}, \text{xû–rê})$$

Le schème radical permet l'expression de cas de répartition morphomique des radicaux, comme en latin (Aronoff, 1994). Dans ce cas, on est en présence d'une répartition régulière, ce qui correspond au fait que les différents lexèmes utilisent le même schème radical partitionné en trois zones de radicaux. Ces trois zones sont représentées dans les tableaux 12 et 13.

Ces zones de radicaux sont par ailleurs liées à des zones flexionnelles au sein desquelles la répartition des exposants sur les lexèmes concernés est homogène (*cf.* 16 et 17).

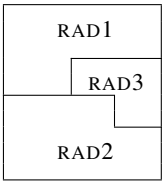

Tableau 12. *Zones de radicaux dans le (sous)paradigme actif du latin*

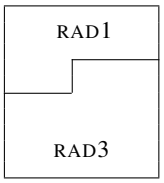

Tableau 13. *Zones de radicaux dans le (sous)paradigme passif du latin*

RAD	SOUSPARA. ACTIF	SOUSPARA. PASSIF
RAD1	*imperf. finite*	*imperf.finite*
RAD2	*perf. finite*	
RAD3	*active future part.*	*passive past part.* *perf. finite (periphr.)*

Tableau 14. *Association morphomique entre les traits morphosyntaxiques et les radicaux verbaux du latin*

4.3.2. *Déponence*

Rappelons qu'à l'instar de Baerman (2007), nous définissons la déponence comme « un décalage entre la forme et la fonction ». Ce décalage s'observe dès lors que les traits exprimés par un flexème donné ne correspondent pas aux traits habituellement exprimés par la règle de réalisation qui lui est associée pour ces traits, c'est-à-dire dès que la règle de transfert $\mathscr{T}_{\chi_{\mathfrak{I}}}$ n'est pas la fonction identité. Autrement dit, la structure de traits $\mathscr{T}_{\chi_{\mathfrak{I}}}(\varkappa)$ exprimée par la forme d'un flexème $\mathfrak{I}$ ne correspond pas à la structure de traits $\varkappa$ qui a servie d'entrée à la règle de flexion $\text{f}_{\mathfrak{I}}$ de ce flexème.

Un flexème $\mathfrak{I}$ donné est *déponent* si et seulement si il existe au moins une forme $\mathfrak{f}$ dans son paradigme $\mathfrak{P}_\mathfrak{I}$ telle que la structure de traits morphosyntaxiques exprimée par $\mathfrak{f}$ est telle que $\varkappa \neq \mathscr{T}_{\chi_\mathfrak{I}}(\varkappa)$.
Soit $\mathscr{P}_{\mathscr{K}_\mathfrak{I}} = \{\mathscr{K}_{\mathfrak{I},1}, ..., \mathscr{K}_{\mathfrak{I},n}\}$ la plus petite partition de $\mathscr{K}_\mathfrak{I}$ telle que $\mathrm{f}_\mathfrak{I}$ associe à chaque $\varkappa \in \mathscr{K}_{\mathfrak{I},i}$ la même zone sur $\mathscr{K}_{\mathfrak{I},i}$. (Pour les flexèmes surabondants, il ne s'agit pas d'une même zone unique, mais du même ensemble de zones.)
Un flexème est dit *semi-déponent* si et seulement si pour au moins un élément $\mathscr{K}_{\mathfrak{I},i}$ de $\mathscr{P}_{\mathscr{K}_\mathfrak{I}}$, mais pas tous, la restriction $\mathscr{T}_{\chi_\mathfrak{I}|\mathscr{K}_{\mathfrak{I}_i}} = \mathrm{id}$ de $\mathscr{T}_{\chi_\mathfrak{I}}$ sur $\mathscr{K}_{\mathfrak{I}_i}$ est la fonction identité.

Nous définissons l'*indice de déponence* $\mathcal{D}_\mathfrak{I}$ d'un flexème $\mathfrak{I}$ comme le nombre de $\mathscr{K}_{\mathfrak{I},i}$ tel que $\mathscr{T}_{|\mathscr{K}_\mathfrak{I},i} \neq \mathrm{id}$:

$$\mathcal{D}_\mathfrak{I} = |\{\mathscr{K}_{\mathfrak{I},i} \in \mathscr{P}_{\mathscr{K}_\mathfrak{I}} | \mathscr{T}_{|\mathscr{K}_\mathfrak{I},i} \neq \mathrm{id}\}|$$

Ainsi, on dit qu'un flexème est déponent si et seulement si $\mathcal{D}_\mathfrak{I} > 0$; un flexème est semi-déponent si et seulement si $|\mathscr{P}_{\mathscr{K}_\mathfrak{I}}| > \mathcal{D}_\mathfrak{I} > 0$; inversement, un flexème non déponent vérifié $\mathcal{D}_\mathfrak{I} = 0$.

Les données du croate du tableau 6 peuvent être modélisées à l'aide de la règle de transfert. Rappelons que le croate emploie parfois des formes du singulier pour exprimer le pluriel (Baerman, 2006). Dans notre modèle, un flexème $\mathfrak{I}$ fonctionnant de la sorte est associé à une règle de transfert $\mathscr{T}_{\chi_\mathfrak{I}}(\{\text{NOMBRE } \textit{pluriel}\})=\{\text{NOMBRE } \textit{singulier}\}$. Il est semi-déponent, la fonction de transfert ne différant de la fonction identité que pour les structures de traits morphosyntaxiques comportant le nombre pluriel.

4.3.3. *Hétéroclise*

Si l'ensemble $\mathcal{X}_{\mathrm{f}_\mathfrak{I}}$ des zones flexionnelles d'un flexème $\mathfrak{I}$ est entièrement inclus dans une classe flexionnelle $\digamma$ unique, le flexème est considéré comme tendant vers la canonicité. À l'inverse, si ses zones flexionnelles sont incluses dans au moins deux classes flexionnelles distinctes, alors le flexème est *hétéroclite*. Nous pouvons définir le degré d'hétéroclicité d'un lexème en calculant son *indice d'hétéroclicité* $\mathcal{H}_\mathfrak{I}$.

$$\mathcal{H}_\mathfrak{I} = |\{\tilde{\xi} | \xi \in \mathsf{X}_\mathfrak{I}\}| - 1$$

où $\tilde{\xi}$, comme défini ci-dessus, indique la classe flexionnelle dont ξ est un des éléments de la partition.
Ainsi, $\mathfrak{I}$ est hétéroclite si et seulement si $\mathcal{H}_\mathfrak{I} > 0$.

Dans le cas des noms d'animaux slovaques présentés au tableau 8, la zone flexionnelle permettant la formation des singuliers du nom *orol* 'aigle' est incluse dans la classe flexionnelle des noms animés comme *chlap* 'garçon', tandis que la zone flexionnelle permettant la formation du pluriel de ces noms d'animaux est incluse dans la classe flexionnelle des inanimés comme *dub* 'chêne'. Leur indice d'hétéroclicité est donc de 1.

CLASSE FLEXIONNELLE	A : NEUTRE RADICAL EN –ET	B : (FÉMININ) RADICAL EN –A	C : (FÉMININ) RADICAL EN –I
dete 'enfant'	SG : $\xi_{A,sg}$	PL : $\xi_{B,sg}$	
tele 'veau'	SG : $\xi_{A,sg}$		PL : $\xi_{C,sg}$

Tableau 15. *Flexion nominale de flexèmes déponents en croate*

De même, pour les noms irréguliers du croate du tableau 7, leurs schèmes flexionnels produisent les zones flexionnelles listées dans le tableau 15. Ces tableaux montrent qu'en plus d'être déponents, ces flexèmes sont également hétéroclites. Les comportements non canoniques de flexion peuvent par conséquent s'accumuler au sein d'un même paradigme.

Les verbes dits « déponents latins » présentent une forme morphologique passive (« m-passif ») tout en exprimant une valeur syntaxique active (« s-actif »). Pour cette raison, ils sont souvent considérés comme des cas de déponence. Cependant, Kiparsky (2005) montre que la morphologie passive du latin induit divers changements sémantiques, dont certains sont imprévisibles. Il s'agit là d'une propriété de la morphologie dérivationnelle, et non de la morphologie flexionnelle qui est généralement considérée comme étant prévisible d'un point de vue sémantique (Boyé, 2011). Nous considérons donc ici le m-actif et le m-passif comme deux classes flexionnelles distinctes qui produisent respectivement les formes morphologiquement actives et passives (Walther, 2011b). Dans la mesure où il existe pour le m-actif et le m-passif des désinences différentes, nous proposons de poser des schèmes flexionnels f différents pour les lexèmes concernés. Autrement dit, chaque flexème associe différemment ses traits morphosyntaxiques et ses zones flexionnelles appartenant à certaines classes flexionnelles (*cf.* tableaux 16 et 17). D'après notre définition formelle des zones flexionnelles, les « verbes déponents latins » peuvent donc être analysés comme des hétéroclites dont la plupart des formes sont constituées à partir de zones appartenant à la classe flexionnelle B[14], tandis que quelques formes supplémentaires puisent dans les zones de la classe flexionnelle A[15] (à savoir A3 pour les participes actifs et A4 pour les gérondifs)[16].

4.3.4. *Défectivité et surabondance*

Dans notre modèle, les flexèmes sont associés à des catégories morphosyntaxiques dont les membres, dans le cas canonique, partagent tous le même ensemble de traits morphosyntaxiques à réaliser. Le fait qu'un flexème appartienne à une catégorie don-

14. Classe habituelle des passifs.

15. Classe habituelle des actifs.

16. Les zones dessinées dans les schémas des tableaux 16 et 17 correspondent à un découpage des paradigmes selon des exposants utilisés par les trois types de lexèmes, d'une part, et une superposition avec les tableaux 12 et 13 d'autre part. Cette superposition s'explique par le fait que les paradigmes doivent à la fois prendre en compte la sélection des radicaux et celle des exposants.

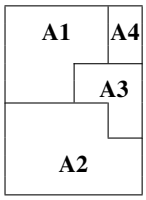

Tableau 16. *Zones flexionnelles de la classe A*

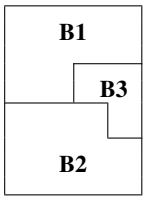

Tableau 17. *Zones flexionnelles de la classe B*

TYPE DU LEXÈME	M-ACTIF	M-PASSIF
ACTIFS	*A1, A2, A3, A4*	
PASSIFS		*B1, B2, B3*
DÉPONENTS	*A3, A4*	*B1, B2, B3*
SEMI-DEP. T1	*A1, A3, A4*	*B2, B3*
SEMI-DEP. T2	*A2, A3, A4*	*B1, B3*

Tableau 18. *Distribution des zones flexionnelles pour la flexion des verbes latins*

née crée donc des attentes morphologiques au sens de Brown *et al.* (2012) quant aux traits qui devront être réalisés par les formes de ce flexème. Si ces attentes ne sont pas remplies, le flexème en question est considéré comme défectif. Ainsi, nous définissons la défectivité comme le fait de ne pas remplir les attentes morphologiques provenant de l'appartenance à une catégorie donnée.

Inversement, lorsque plus de formes sont produites que ce qui serait *a priori* attendu d'un flexème donné (étant donné son appartenance à une certaine catégorie), ce flexème est considéré comme étant surabondant.

On considère qu'un paradigme est défectif si et seulement si il existe au moins une structure de traits morphosyntaxiques $\varkappa$ appartenant à l'ensemble $\mathscr{K}_{\mathcal{C}_\mathfrak{I}}$ des traits morphosyntaxiques de la catégorie à laquelle appartient un flexème $\mathfrak{I}$ que $\mathrm{f}_\mathfrak{I}$ n'associe à aucune zone flexionnelle ξ. Autrement dit, un flexème $\mathfrak{I}$ est défectif si et seulement si l'ensemble $\mathscr{K}_\mathfrak{I}$ de ses traits morphosyntaxiques ne couvre pas l'intégralité de l'ensemble $\mathscr{K}_{\mathcal{C}_\mathfrak{I}}$ des traits morphosyntaxiques de sa catégorie : $\mathscr{K}_\mathfrak{I} \subsetneq \mathscr{K}_{\mathcal{C}_\mathfrak{I}}$. Prenons par exemple le flexème $\mathfrak{I}$ *vivres* du français : $\mathscr{K}_\mathfrak{I} = \{\text{NOMBRE } \textit{pluriel}\}$ tandis que $\mathscr{K}_{\mathcal{C}_\mathfrak{I}} = \mathscr{K}_{\mathrm{nom}} = \{\text{NOMBRE } \textit{singulier}, \text{NOMBRE } \textit{pluriel}\}$.

On considère qu'un paradigme est *surabondant* si et seulement si il existe au moins une structure de traits morphosyntaxiques $\varkappa$ appartenant à l'ensemble $\mathscr{K}_{\mathcal{C}_\mathfrak{I}}$ des traits morphosyntaxiques de la catégorie d'un flexème $\mathfrak{I}$ à qui $\mathrm{f}_\mathfrak{I}$ associe plus d'une zone flexionnelle : dans ce cas $\mathrm{f}_\mathfrak{I}$ est bien une relation binaire générique et pas une fonction. Les données de l'italien dans le tableau 9 en sont un exemple : le flexème *languire* est tel que $\mathrm{f}_{languire}$ associe à la structure de traits $\{3\text{PL.PRS.SUBJ}\}$ deux zones permettant chacune la production d'une des formes respectives *languano* et *languiscano*.

4.3.5. *Flexion canonique*

Il suit de ces définitions que la *flexion canonique*, si elle existe, correspond au cas où le schème flexionnel $\mathrm{f}_\mathfrak{I}$ d'un flexème $\mathfrak{I}$ a pour image un ensemble de zones reconstituant exactement une classe flexionnelle unique $\digamma$. En particulier, quelle que soit la structure de traits $\varkappa$, le schème flexionnel $\mathrm{f}_\mathfrak{I}$ associe $\varkappa$ à un des éléments de la partition de $\digamma$.

De plus, le schème radical doit associer à tout $\varkappa$ une zone de radicaux constante, appartenant à une classe de radicaux dont la partition ne contient qu'un seul élément, et qui produit les mêmes règles de formation de radicaux quel que soit $\varkappa$. Autrement dit, le radical ne change pas.

Enfin, la règle de transfert $\mathscr{T}_{\chi_\mathfrak{I}}$ est la fonction identité et l'ensemble des traits morphosyntaxiques $\mathscr{K}_\mathfrak{I}$ de $\mathfrak{I}$ est égal à l'ensemble des traits morphosyntaxiques $\mathscr{K}_{\mathcal{C}_\mathfrak{I}}$ de sa catégorie $\mathcal{C}_\mathfrak{I} \in \mathscr{C}$.

Il en va de même pour la règle de transfert $\mathscr{T}_{\psi_\mathfrak{I}}$.

Flexion canonique :

$\exists \digamma,$ tel que $\forall \varkappa \in \mathscr{K}_\mathfrak{I}, \widetilde{\mathrm{f}_\mathfrak{I}}(\varkappa, \digamma)$ ce qui signifie que $|\tilde{\xi}|\xi \in \mathcal{X}_{\mathrm{f}_\mathfrak{I}}\}| - 1 = 0$

$\exists \Gamma,$ tel que $\forall \varkappa \in \mathscr{K}_\mathfrak{I}, \widetilde{\mathrm{s}_\mathfrak{I}}(\varkappa, \Gamma)$, où Γ est une fonction indépendante de $\varkappa$

et $\mathscr{T}_{\chi_\mathfrak{I}} = \mathrm{id}$

et $\mathscr{T}_{\psi_\mathfrak{I}} = \mathrm{id}$

et $\mathscr{K}_\mathfrak{I} = \mathscr{K}_{\mathcal{C}_\mathfrak{I}}$.

5. Mesurer la complexité de différentes descriptions de la flexion verbale du français

Nous avons montré comment notre modélisation de la morphologie flexionnelle PARSEMOR permet de représenter les phénomènes flexionnels non canoniques, au moyen de notions nouvelles comme celle de zone flexionnelle, que l'on peut considérer comme une généralisation de la notion d'espace thématique (*stem space*) de Bonami et Boyé (2003) (et avant eux Pirelli et Battista (2000)). PARSEMOR représente ces espaces thématiques comme des schèmes radicaux. Différentes analyses concurrentes des mêmes données peuvent être développées au sein de PARSEMOR, implémentées et ainsi évaluées quantitativement au moyen de métriques de complexité. Il s'agit là d'un moyen de comparer de telles analyses entre elles en termes de complexité, mais également d'examiner la pertinence des notions nouvelles introduites dans notre modèle, notamment en fonction de leur utilité pour obtenir les descriptions de moindre complexité.

Pour répondre à ces questions, nous avons développé et implémenté un formalisme capable de représenter le modèle décrit à la section 4. Nous sommes partis

de la couche morphologique du formalisme lexical Alexina (Sagot, 2010), sur lequel reposent un certain nombre de ressources lexicales morphologiques (et, pour certaines, syntaxiques). Nous avons étendu ce formalisme pour y intégrer les notions de zones flexionnelles et de radicaux, de règles de transfert, de schèmes flexionnels et de schèmes radicaux.

Nous avons alors encodé dans ce formalisme différentes descriptions morphologiques concurrentes de la flexion verbale du français, afin d'étudier la pertinence de ces nouvelles notions en termes de complexité.

5.1. *Analyses de la flexion verbale du français*

La flexion verbale du français est intéressante par plusieurs aspects, dont certains ont été mentionnés précédemment. Tout d'abord, il s'agit d'un système riche qui génère des formes pour une cinquantaine de structures de traits morphosyntaxiques. Par ailleurs, et c'est tout particulièrement pertinent dans le cadre d'analyses de complexité, ce système est traditionnellement décrit comme comportant une classe flexionnelle régulière et productive, celle des verbes du premier groupe (verbes en *-er*), une classe flexionnelle irrégulière (verbes du troisième groupe), et la classe flexionnelle des verbes du deuxième groupe (verbes comme *finir*), parfois considérée comme régulière et parfois comme irrégulière. Les analyses diffèrent sur la productivité réelle de cette dernière classe (Boyé, 2000 ; Kilani-Schoch et Dressler, 2005 ; Bonami *et al.*, 2008), ce qui constitue un premier point possible de divergence entre analyses différentes de la flexion verbale du français.

Parmi les verbes du premier groupe, et comme décrit à la section 3.6, les verbes en *-ayer* manifestent une surabondance régulière. Dans (Bonami et Boyé, 2010), les auteurs considèrent ces verbes comme polyparadigmatiques. Ceci n'est pas pleinement satisfaisant dans la mesure où les deux paradigmes ainsi supposés partageraient les mêmes formes pour la moitié des cases. Une autre façon de représenter cette situation est de définir deux radicaux, l'un en *-ay-* et l'autre en *-ai-*, et deux zones flexionnelles : l'une, ξ_1, qui est utilisée seulement par le radical en *-ay-*, et l'autre, ξ_2, qui est utilisée par les deux radicaux. Ainsi, il y a trois sous-schèmes au sein du schème spécifique concernant les verbes en *-ayer* : ξ_1+*-ay-*, ξ_2+*-ai-* et ξ_2+*-ay-*.

La modélisation des verbes du deuxième groupe peut également être réalisée de différentes façons. Si l'on utilise l'approche à douze espaces thématiques de Bonami et Boyé (2003), ces verbes peuvent définir explicitement un second radical en *-iss* dans le lexique, à côté du radical de base en *-i* (par exemple, *fini-* et *finiss-* pour *finir*). La méthode traditionnelle pour représenter ces verbes, qui est également la plus répandue, consiste à leur attribuer une classe flexionnelle faisant usage de suffixes commençant par *-ss-* pour un certain nombre de cases. Une autre possibilité consiste à diviser la classe flexionnelle des verbes du deuxième groupe en deux zones : la zone ζ'_1 pour les cases pour lesquelles les verbes du deuxième groupe utilisent un radical secondaire en

-iss, et la zone ζ'_2 pour les autres cases utilisant le radical en *-i*. Le schème est défini au moyen de deux sous-schèmes, à savoir ξ'_1+*-iss-* et ξ'_2+*-i-*.

Quant aux verbes du troisième groupe, les deux seules approches que nous avons prises en compte sont l'approche traditionnelle, qui fait usage de nombreuses classes flexionnelles, et l'approche à douze radicaux de Bonami et Boyé (2003). La représentation de cette dernière approche dans notre modèle se fait simplement, en représentant les dépendances (par défaut) entre radicaux au moyen d'un schème radical, et en spécifiant pour chaque verbe (seulement) ceux de ses radicaux qui ne peuvent être obtenus de façon régulière au moyen du schème radical et des radicaux déjà connus (calculés au moyen du schème radical ou eux aussi spécifiés dans le lexique).

À partir de là, nous avons développé quatre descriptions différentes de la morphologie verbale du français dans la nouvelle version du formalisme Alexina qui implémente notre modèle morphologique, afin d'en mesurer les complexités respectives.

5.2. *Quantifier et mesurer la complexité morphologique*

Ces dernières années, le développement de méthodes de mesure de la complexité linguistique est devenu un domaine de recherche actif. Par ordre décroissant de généralité, des travaux ont été réalisés dans ce domaine sur les langues prises de façon globale (McWhorter, 2001 ; Juola, 2008), en restreignant le travail à un niveau d'analyse particulier tel que la morphologie (Bane, 2008) et en mesurant la complexité de descriptions morphologiques particulières, notamment dans le contexte de l'acquisition automatique non supervisée ou faiblement supervisée de la morphologie (Goldsmith, 2001 ; Xanthos, 2008).

Différentes métriques pour mesurer la complexité linguistique sont décrites dans la littérature. Les plus simples se contentent de compter les occurrences d'un ensemble de propriétés linguistiques définies manuellement : taille de divers inventaires (par exemple, inventaires des phonèmes, des catégories, des types de morph(ème)es…) (McWhorter, 2001). Cependant, de telles approches sont intrinsèquement biaisées : tant l'ensemble de propriétés choisies que les critères selon lesquels ces propriétés sont décrites sont très difficiles à définir selon des principes clairs, objectifs et reproductibles (par exemple, comment construire d'une façon objective et indépendante de la langue un inventaire de parties du discours ?). Des approches alternatives de mesure de la complexité existent, dont la plupart reposent sur des définitions de la complexité qui sont issues de la théorie de l'information. Dans ce cadre, deux définitions différentes ont été utilisées dans les travaux récents, qui s'appliquent toutes deux à n'importe quel type de message et pas seulement aux descriptions ou aux modèles linguistiques : (i) l'*entropie informationnelle* (ou *complexité de Shannon*), dont le principal inconvénient est qu'il nécessite l'encodage du message sous la forme d'une séquence de variables aléatoires indépendantes et identiquement distribuées selon un certain modèle probabiliste, ce qui est difficile en pratique, et (ii) l'*entropie algorithmique* (ou *complexité de Kolmogoroff*), un moyen plus générique et plus objectif de mesure de

la quantité d'informations contenue dans un message, mais qui n'est pas directement calculable et pour lequel on doit donc se contenter d'approximations.

La complexité de Kolmogoroff, parce qu'elle est plus générale et ne nécessite pas de modèle probabiliste sous-jacent, est plus adaptée à nos besoins. Elle repose sur l'intuition suivante : un modèle est plus complexe qu'un autre s'il nécessite un message plus long pour être décrit. Cependant, puisque le calcul de cette complexité est impossible directement, ce calcul est souvent ramené à celui d'un certain type d'entropie au sein de l'espace formé par les descriptions possibles au sein d'un modèle particulier, au moyen d'une approximation de la complexité de Kolmogoroff définie sur ce modèle : le résultat est ce que l'on appelle la *longueur de description* par rapport au modèle. Cette notion est le fondement du paradigme d'apprentissage automatique appelé *longueur de description minimale* (*Minimum Description Length*, ou *MDL*) (Rissanen, 1984) : dans ce paradigme, l'apprentissage automatique se fait *via* l'identification de la description que l'on peut faire des données d'apprentissage qui ait une *longueur de description* minimale par rapport à un modèle choisi. Pour que la longueur de description soit une bonne approximation de la complexité de Kolmogoroff, il faut toutefois que l'on se dote d'une façon aussi efficace que possible de représenter une description, ici une description linguistique, sous la forme d'une chaîne de caractères (le code) (Bane, 2008). De plus, un modèle linguistique est le plus souvent structuré, ce qui n'est pas toujours exploité par les travaux faisant usage de la complexité morphologique. En particulier, l'évaluation de la complexité d'une représentation du lexique morphologique ne peut se réduire à mesurer la complexité d'un corpus dont les formes ont été segmentées en morph(ème)es. Cette approche est toutefois le fondement de travaux pionniers en acquisition automatique d'informations morphologiques (Goldsmith, 2001).

Dans notre cas, nous souhaitons mesurer la complexité de descriptions (d'une partie) de la morphologie d'une langue donnée. Il s'agit donc d'un objectif distinct de ceux des travaux constrastifs entre langues visant à comparer leur complexité morphologique (ou linguistique) de façon globale (McWhorter, 2001 ; Juola, 2008 ; Bane, 2008) : nous ne cherchons pas à estimer la complexité d'une langue, mais celle de descriptions particulières de son niveau morphologique, et, plus spécifiquement, de son système flexionnel lexical.

La longueur de description $DL(m)$ d'un message non structuré m au sein d'un modèle donné qui le code sous la forme d'une séquence de N symboles d'un alphabet $W = \{w_1, \ldots, w_n\}$ est définie par : $DL(m) = -\sum_i o(w_i) \log_2 (o(w_i)/N)$, où $o(w_i)$ est le nombre d'occurrences de w_i dans m. Cette longueur de description est égale à N fois l'entropie du message.

Dans notre cas, le codage ne peut être si simple, puisqu'une description morphologique est structurée. Tout d'abord, elle se décompose en un lexique et une grammaire morphologiques. Dans notre formalisme, nous définissons une entrée lexicale comme une forme canonique, un schème, un schème radical optionnel, si le schème radical par défaut du schème n'est pas le bon, et une liste optionnelle de radicaux non prédictibles, comme vu précédemment (les radicaux prédictibles n'ont pas à être

explicités). Quant à la grammaire morphologique, elle met en jeu des schèmes et sous-schèmes, des tables de flexion, des zones flexionnelles, des procédés de construction de formes (affixes...), des règles de sandhi[17] et d'autres mécanismes de factorisation. Nous avons défini un codage qui représente toute cette structure de façon bijective (il peut être décodé de façon non ambiguë) au moyen de symboles de seize alphabets différents (un pour les lettres des formes canoniques, un pour les étiquettes morpho-syntaxiques, un pour les identifiants de schèmes, un pour les informations structurelles au sein des tables, etc.). Des expériences préliminaires nous ont en effet montré que l'utilisation d'alphabets distincts conduit à des descriptions plus courtes en termes de longueur de description, laquelle est alors définie en généralisant la formule ci-dessus comme suit. Si un message m est codé comme une séquence de symboles pris parmi p alphabets $W^1, \ldots, W^p$, $W^1 = \{w_1^1, \ldots, w_{n_1}^1\}, \ldots, W^p = \{w_1^p, \ldots, w_{n_p}^p\}$ de telle façon que l'alphabet dont est sélectionné un symbole peut toujours être déterminé par son contexte gauche, alors nous définissons la longueur de description comme étant :

$$DL(m) = -\sum_{j=1}^{p} \sum_{i=1}^{n_p} o(w_i^p) \log_2 \frac{o(w_i^p)}{N_p},$$

où N_p est le nombre de symboles de l'alphabet W^p dans m. Une telle métrique permet d'approcher la complexité d'une description structurée, mais également de mesurer la contribution de chaque alphabet à cette complexité. C'est ainsi que nous avons pu calculer la complexité de différentes descriptions de la morphologie flexionnelle des verbes du français faites dans notre modèle, à la fois pour évaluer la pertinence des concepts nouveaux introduits dans ce modèle (zones flexionnelles, schèmes) et pour comparer ces descriptions morphologiques concurrentes.

5.3. *Mesurer la complexité de différentes descriptions morphologiques de la flexion verbale du français*

Nous avons mentionné précédemment différentes façons de décrire certains aspects de la flexion verbale du français, qui correspondent à des répartitions différentes entre les informations décrites dans le lexique et dans la grammaire morphologique. Nous avons utilisé l'inventaire de lexèmes verbaux du lexique Le*fff* (Sagot, 2010) pour nos expériences. Nous avons ignoré les distinctions entre entrées ne différant que par des propriétés non morphologiques (un même verbe peut avoir plusieurs entrées dans le Le*fff* correspondant à différents cadres de sous-catégorisation et/ou à différents sens). Nous avons ainsi obtenu un inventaire de flexèmes (ou lemmes). La version actuelle du Le*fff* contient 7 820 verbes parmi lesquels 6 966 verbes du premier groupe,

17. Ce que nous appelons ici *règles de sandhi*, sont des règles morphographémiques et/ou morphophonémiques, déjà implémentées dans le formalisme Alexina. Ce sont des transformations locales qui s'appliquent à l'interface entre deux morph(ème)es. Ainsi, dans la flexion verbale du français, un radical en *-g* suivi d'un suffixe en *[aou]-* est mis en correspondance avec une forme de surface où le radical est allongé d'un *e* : *mang_ons* ↔ *mange_ons*.

315 du deuxième groupe et 539 du troisième groupe. L'ensemble de ces verbes produisent 300 693 formes fléchies distinctes.

Dans notre nouvelle version de la couche morphologique d'Alexina, les informations morphologiques associées à une entrée lexicale sont les suivantes :

– une forme de citation (l'infinitif pour les verbes français) ;

– un schème suivi d'une variante de schème optionnelle : si deux schèmes ne diffèrent que pour quelques cases, ils peuvent être fusionnés, des règles alternatives de réalisation sont associées à ces cases, ces règles étant déclenchées lexicalement par ces variantes de schème ;

– facultativement, un schème radical (ne pas spécifier le schème radical indique que l'on doit utiliser le schème radical par défaut associé au schème) ;

– facultativement, une liste de radicaux (les radicaux non spécifiés sont à construire par l'application des règles de formation de radicaux associées au schème radical).

Par exemple, une entrée telle que « bouillir $v_{23r}/_{\mathrm{bouill,,bou}}$ » correspond à un flexème (ou lemme) dont la forme de citation est *bouillir*, le schème *v* (avec la variante de schème*23r*), le schème radical par défaut pour le schème *v* tel qu'indiqué dans la grammaire morphologique, et dont le radical 1 est *bouill*, le radical 3 *bou* (les autres radicaux étant à construire à partir de ces deux radicaux explicités et des règles de formation de radicaux du schème radical).

Nous allons désormais décrire brièvement les quatre modélisations différentes de la flexion verbale du français que nous avons développées. Pour illustrer notre propos, les entrées lexicales pour un petit ensemble de flexèmes dans chacune de ces descriptions sont données au tableau 19.

À l'une des extrémités du spectre de descriptions possibles, nous avons produit automatiquement une description « plate », nommée FLAT, qui n'utilise pas les notions de radicaux, de sandhi et de zones flexionnelles, de la façon suivante. Nous avons extrait pour chaque verbe la sous-chaîne commune la plus longue entre toutes ses formes. Dans chacune de ces formes, ce qui est à droite de cette partie commune est considéré comme le « suffixe ». Chaque verbe reçoit alors une *signature* qui est la liste des suffixes ainsi extraits, ordonnés par ordre alphabétique de l'étiquette morphosyntaxique correspondante. Ensuite, on attribue à tous les verbes ayant une même signature une même classe flexionnelle définie à partir de cette signature. Le résultat est une description comportant 139 classes flexionnelles. Sa longueur de description, mesurée de la façon expliquée à la section précédente, est d'environ 131 400 bits (9 200 bits dans le lexique[18] et 122 200 bits dans la grammaire morphologique).

À l'autre extrémité du spectre se trouve l'analyse de Bonami et Boyé (2003, 2008), qui utilise seulement une classe flexionnelle mais douze radicaux. Nous sommes partis

18. Ici et dans tous les résultats qui suivent, la contribution de la liste de formes canoniques est ignorée, puisqu'elle est commune à toutes les descriptions.

FORME DE CITATION	FLAT	ORIG	NEW	BOBO
aimer	v1	v-er$_{std}$	v-er	v_1
acheter	v18	v-er$_{std}$	v-er	v_1
jeter	v8	v-er$_{dbl}$	v-eter	$v_1/_{jett,,,,,,,jettə}$
balayer	v12	v-ayer	v-ayer	v-ayer$_1$
finir	v2	v-ir2	v-ir2	v_{23r}
requérir	v42	v-ir3	v-ir3	$v_{23r}/_{requér,requier,,,,,,,requer,requi,requis}$
cueillir	v51	v-assaillir	$v_{23r}/_{cueill,,,,,,,,cueillə}$	$v_{23r}/_{cueill,,,,,,,,cueillə}$
prendre	v24	v-prendre	v-prendre	v_{3re}
mettre	v17	v-mettre	v-mettre	$v_{3re}/_{„met,,,,,,,mi,mis}$

Tableau 19. *Entrées lexicales pour quelques flexèmes dans chacune de nos quatre représentations concurrentes de la flexion verbale du français*

d'une implémentation préliminaire en DATR de ce modèle (Bonami, c.p.). Parce que cette analyse était implémentée sur les phonèmes et non les graphèmes, nous avons dû appliquer un certain nombre de transformations pour nous permettre de représenter sans erreur la flexion graphémique, y compris par l'ajout de règles de sandhi. Pour générer correctement toutes les formes surabondantes, nous avons dû également étendre cette analyse de différentes façons. Le résultat est une description, appelée BOBO, qui ne contient qu'une classe flexionnelle, plusieurs schèmes (4 pour les verbes non défectifs, quelques autres pour les verbes défectifs) et 61 règles de sandhi. Cette description repose fortement sur une caractéristique importante de notre implémentation dans Alexina du modèle décrit dans cet article : toute information non explicitée est reconstruite par des mécanismes par défaut. Par exemple, ne pas spécifier un certain radical dans une entrée lexicale conduit à utiliser son schème radical pour le produire ; si le schème radical n'est pas spécifié, on utilise le schème radical par défaut associé au schème de l'entrée ; si ce schème ne définit pas de schème radical par défaut, on utilise le même radical que celui de la forme canonique. Le résultat est que BOBO a une longueur de description d'environ 51 400 bits qui se répartissent en 46 000 bits dans le lexique (ce qui est particulièrement élevé et provient du fait que de nombreux radicaux sont spécifiés explicitement, y compris pour les verbes du deuxième groupe) et 5 400 bits dans la grammaire morphologique, ce qui est très bas, comme attendu.

Entre ces deux extrêmes, la description ORIG utilisée actuellement par le Lefff, qui repose sur un nombre très important de règles de sandhis mais utilise un nombre assez élevé de classes flexionnelles pour les verbes du troisième groupe, a une longueur de description de 82 900 bits (8 100 bits dans le lexique, 74 800 bits dans la description). Mais c'est une autre description qui donne les résultats les plus intéressants. Comme indiqué plus haut, en utilisant la notion de zone flexionnelle conjointement à un ensemble raisonnable de règles de sandhi, nous avons pu développer une description morphologique plus satisfaisante des verbes français, nommée NEW, qui utilise 20 schèmes. Parmi eux, se trouvent notamment un schème pour les verbes du premier groupe sans surabondance, un schème pour les verbes en *-ayer* en utilisant les zones ξ_1 et ξ_2 décrites à la section 5.1 et un schème pour les verbes du deuxième groupe, analysés comme réguliers au moyen de deux sous-schèmes, comme indiqué à cette même

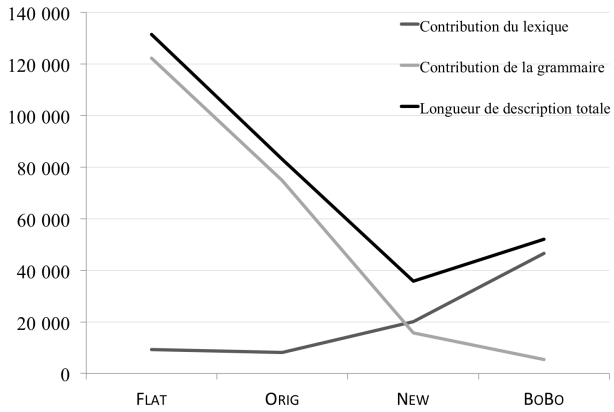

Figure 1. *Longueurs de description de différentes descriptions de la morphologie verbale du français*

.

section. Les schèmes définis pour les verbes du troisième groupe reprennent les classes flexionnelles les plus peuplées de ORIG (verbes en *-ir*, verbes en *-re* les « moins irréguliers », verbes en *-prendre, -mettre*, etc), et traitent les autres verbes comme dans BOBO. La longueur de description correspondante, seulement 35 800 bits, est inférieure à celle de BOBO. Elle se répartit en 20 100 bits dans le lexique (deux fois plus que pour FLAT, mais deux fois moins que pour BOBO) et 15 700 bits dans la grammaire morphologique (trois fois plus que dans BOBO, mais huit fois moins que dans FLAT).

Ces différents chiffres pour chacune des quatre descriptions, indiqués en ordonnant les descriptions le long du spectre, sont représentés à la figure 1. Ils montrent bien que l'utilisation de la notion de zone flexionnelle, qui généralise la notion d'espace thématique (qui dans notre modèle correspond à celle de schème radical), permet de rendre compte de la morphologie verbale du français d'une façon plus économique en termes de longueur de description et donc de complexité par rapport aux descriptions traditionnelles mais aussi par rapport à une description récente et originale (Bonami *et al.*, 2008). On peut noter que nous n'aurions pas obtenu ce résultat si nous n'avions pas pris en compte la longueur de description du lexique, mais seulement celle de la grammaire morphologique. Cela dit, puisque la répartition de l'information morphologique entre lexique et règles varie d'une description à une autre, cela n'aurait pas grand sens d'évaluer la longueur de description de la grammaire morphologique seule.

6. Conclusion

Dans cet article, nous avons présenté un modèle formel de la morphologie flexionnelle, qui repose sur des notions nouvelles comme celle de zone flexionnelle. De nom-

breux phénomènes flexionnels non canoniques peuvent être définis formellement dans ce modèle, et notamment la supplétion, la déponence, l'hétéroclise, la défectivité et la surabondance. Nous avons présenté ces phénomènes à partir de données issues de diverses langues (français, latin, italien, kurde sorani, persan, croate et slovaque), et nous avons montré comment ces données pouvaient être représentées dans notre modèle. Nous avons plus particulièrement insisté sur la flexion verbale du français, en indiquant quatre différentes façons de la décrire dans notre modèle et en mesurant leur complexité, au moyen d'une métrique issue de la théorie de l'information et adaptée à notre modèle et notamment au caractère structuré des descriptions. Les résultats quantitatifs obtenus sur nos quatre descriptions montrent que des façons traditionnelles de décrire la morphologie verbale du français au moyen de nombreuses classes flexionnelles, mais également un modèle récent et radicalement différent (Bonami *et al.*, 2008), sont tous d'une complexité plus élevée qu'une description que nous avons développée en utilisant des notions que nous avons introduites dans notre modèle, et en particulier les notions de zone flexionnelle, de schème radical et de schème flexionnel, de façon à équilibrer au mieux la répartition des informations morphologiques entre le lexique et la grammaire morphologiques.

7. Bibliographie

Aronoff M., *Morphology by Itself*, MIT Press, 1994.

Baerman M., « Deponency in Serbo-Croatian », , Typological Database on Deponency, Surrey Morphology Group, CMC, University of Surrey, 2006. Online Database : `http://www.smg.surrey.ac.uk/deponency/Examples/Serbo-Croatian.htm`.

Baerman M., « Morphological Typology of Deponency », *in* M. Baerman, G. G. Corbett, D. Brown, A. Hippisley (eds), *Deponency and Morphological Mismatches*, vol. 145, The British Academy, Oxford University Press, p. 1-19, 2007.

Baerman M., Corbett G. G., Brown D., *Defective Paradigms*, Oxford University Press, Oxford, Royaume-Uni, 2010. Proceedings of the British Academy 145.

Baerman M., Corbett G. G., Brown D., Hippisley A. (eds), *Deponency and Morphological Mismatches*, Oxford University Press, 2007.

Bane M., « Quantifying and Measuring Morphological Complexity », *in* C. B. Chang, H. J. Haynie (eds), *Proceedings of the 26th West Coast Conference on Formal Linguistics*, Sommerville, États-Unis, 2008.

Baroni M., Matiasek J., Trost H., « Unsupervised discovery of morphologically related words based on orthographic and semantic similarity », *Proceedings of the ACL Workshop on Morphological and Phonological Learning*, p. 48-57, 2002.

Beesley K. R., Karttunen L., *Finite State Morphology*, CSLI, 2003.

Bernhard D., « Apprentissage non supervisé de familles morphologiques par classification ascendante hiérarchique », *Actes de la 14e conférence sur le traitement automatique des langues naturelles – TALN 2007*, June, 5-8, Toulouse, France, p. 367-376, 2007.

Bernhard D., « MorphoNet : Exploring the Use of Community Structure for Unsupervised Morpheme Analysis », *Multilingual Information Access Evaluation Vol. I, 10th Workshop of the*

Cross-Language Evaluation Forum, CLEF 2009, Revised Selected Papers, Springer, Corfu, Grèce, 2010.

Bonami O., Boyé G., « Suppletion and dependency in inflectional morphology », *in* F. V. Eynde, L. Hellan, D. Beerman (eds), *The Proceedings of the HPSG '01 Conference*, CSLI Publications, Stanford, États-Unis, 2002.

Bonami O., Boyé G., « Supplétion et classes flexionnelles dans la conjugaison du français », *Langages*, vol. 152, p. 102-126, 2003.

Bonami O., Boyé G., « Deriving inflectional irregularity », *Proceedings of the 13th International Connference on HPSG*, CSLI Publications, Stanford, États-Unis, p. 39-59, 2006.

Bonami O., Boyé G., « La morphologie flexionnelle est-elle une fonction ? », *in* I. Choi-Jonin, M. Duval, O. Soutet (eds), *Typologie et comparatisme. Hommages offerts à Alain Lemaréchal*, Peeters, Louvain, Belgique, p. 21-35, 2010.

Bonami O., Boyé G., Giraudo H., Voga M., « Quels verbes sont réguliers en français ? », *Actes du premier Congrès Mondial de Linguistique Française*, p. 1511-1523, 2008.

Boyé G., « Régularité et classes flexionnelles dans la conjugaison du français », *in* M. Roché, G. Boyé, N. Hathout, S. Lignon, M. Plénat (eds), *Des unités morphologiques au lexique*, Hermes Science, 02, 2011.

Boyé G., Problèmes de morpho-phonologie verbale en français, espagnol et italien, PhD thesis, Université Paris 7, 2000.

Boyé G., « Suppletion », *in* K. Brown (ed.), *Encyclopedia of Language and Linguistics (2nd ed.)*, vol. 12, Elsevier, Oxford, Royaume-Uni, p. 297-299, 2006.

Brown D., Chumakina M., Corbett G. G., Popova G., Spencer A., « Defining 'periphrasis' : key notions », *Morphology*, 2012. À paraître.

Cartoni B., « Lexical Morphology in Machine Translation : A Feasibility Study », *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, Athens, Grèce, p. 130-138, March, 2009.

Chomsky N., Halle M., *The sound pattern of English*, Harper and Row, 1968.

Corbett G. G., « Agreement : the range of the phenomenon and the principles of the Surrey database of agreement », *Transactions of the philological society*, vol. 101, p. 155–202, 2003.

Corbett G. G., « Canonical Typology, suppletion and possible words », *Language*, vol. 83, p. 8-42, 2007a.

Corbett G. G., « Deponency, Syncretism, and What Lies Between », *in* M. Baerman, G. G. Corbett, D. Brown, A. Hippisley (eds), *Deponency and Morphological Mismatches*, vol. 145, The British Academy, Oxford University Press, p. 21-43, 2007b.

Corbett G. G., Fraser N., « Network Morphology : a DATR account of Russian nominal inflection », *Journal of Linguistics*, vol. 29, p. 113-142, 1993.

Creutz M., Lagus K., « Unsupervised Discovery of Morphemes », *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, p. 21-30, 2002.

Deléger L., Namer F., Zweigenbaum P., « Morphosemantic parsing of medical compound words : Transferring a French analyzer to English », *International Journal of Medical Informatics*, vol. 78, p. 48-55, 2009.

Demberg V., « A Language-Independent Unsupervised Model for Morphological Segmentation », *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, p. 920-927, June, 2007.

Ernout A., Thomas F., *Syntaxe Latine*, 2 edn, Kliensieck, Paris, 1953.

Fradin B., *Nouvelles approches en morphologie*, Presses Universitaires de France, Paris, France, 2003.

Fradin B., Kerleroux F., « Troubles with Lexemes », *in* G. Booij, A. R. Janet de Cesaris, Sergio Scalise (eds), *Selected papers from the Third Mediterranean Morphology Meeting*, Topics in Morphology, IULA-Universitat Pompeu Fabra, Barcelona, Espagne, p. 177-196, September 20-22 2001, 2003.

Gaussier E., « Unsupervised learning of derivational morphology from inflectional lexicons », *Proceedings of the Workshop on Unsupervised Methods in Natural Language Processing*, University of Maryland, 1999.

Goldsmith J., « Unsupervised Learning of the Morphology of a Natural Language », *Computational Linguistics*, vol. 27, n° 2, p. 153-198, 2001.

Hahn U., Honeck M., Shulz S., « Subword-Based Text Retrieval », *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*, January 06 - 09, Big Island, Hawaii, 2003.

Halle M., Marantz A., « Distributed morphology and the pieces of inflection », *in* K. Hale, S. J. Keyser (eds), *The view from building 20*, MIT Press, Cambridge, États-Unis, p. 111-176, 1993.

Harris Z. S., « From Phoneme to Morpheme », *Language*, vol. 31, n° 2, p. 190-222, 1955.

Hathout N., « From WordNet to CELEX : acquiring morphological links from dictionaries of synonyms », *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Espagne, p. 1478-1484, 2002.

Hippisley A., « Declarative Deponency : A Network Morphology Account of Morphological Mismatches », *in* M. Baerman, G. G. Corbett, D. Brown, A. Hippisley (eds), *Deponency and Morphological Mismatches*, vol. 145, The British Academy, Oxford University Press, p. 145-173, 2007.

Hockett C. F., « Two models of linguistc descriptions », *Words*, vol. 10, p. 210-234, 1954.

Jacquemin C., « Guessing morphology from terms and corpora », *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 156 - 165, 1997.

Juola P., « Assessing Linguistic Complexity », *in* M. Miestamo, K. Sinnemäki, F. Karlsson (eds), *Language Complexity : Typology, Contact, Change*, John Benjamins Press, Amsterdam, Pays-Bas, 2008.

Keshava S., « A simpler, intuitive approach to morpheme induction », *In PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*, p. 31-35, 2006.

Kilani-Schoch M., Dressler W. U., *Morphologie naturelle et flexion du verbe français*, Gunter Narr Verlag, Tübingen, Allemagne, 2005.

Kiparsky P., « Blocking and periphrasis in inflectional paradigms », *Yearbook of Morphology 2004*, Springer, Dordrecht, Pays-Bas, p. 113-135, 2005.

Koskenniemi K., « A general computational model for word-form recognition and production », *Proceedings of the 22nd annual meeting of the Association for Computational Linguistics*, p. 178-181, 1984.

Lepage Y., « Solving analogies on words : an algorithm », *Proceedings of the 17th international conference on Computational Linguistics*, p. 728-734, 1998.

Lieber R., *Deconstructing Morphology : Word Formation in Syntactic Theory*, University of Chicago Press, Chicago, États-Unis, 1992.

Lovis C., Michel P.-A., Baud R., Scherrer J.-R., « Word Segmentation Processing : A Way to Exponentially Extend Medical Dictionaries », *in* R. A. Greenes, H. E. Peterson, D. J. Protti (eds), *Proceedings of the 8th World Congress on Medical Informatics*, p. 28-32, 1995.

Matthews P. H., *Morphology*, Cambridge University Press, Cambridge, Royaume-Uni, 1974.

McCarus E. N., A Kurdish Grammar : descriptive analysis of the Kurdish of Sulaimaniya, Iraq, PhD thesis, American Council of Learned Societies, New-York, États-Unis, 1958.

McWhorter J., « The world's simplest grammars are creole grammars », *Linguistic Typology*, vol. 5, p. 125-166, 2001.

Moreau F., Claveau V., Sébillot P., « Automatic morphological query expansion using analogy-based machine learning », *Proceedings of the European Conference on Information Retrieval, ECIR 07*, Rome, Italie, April, 2007.

Namer F., « Morphologie, Lexique et TAL : l'analyseur DériF », *TIC et Sciences cognitives*, Hermes Sciences Publishing, Londres, Royaume-Uni, 2009.

Pinker S., *Words and Rules*, Basic Books, New-York, NY, États-Unis, 1999.

Pirelli V., Battista M., « The Paradigmatic Dimension of Stem Allomorphy in Italian Verb Inflection », *Italian Journal of Linguistics*, vol. , p. 307-380, 2000.

Porter M. F., « An algorithm for suffix stripping », *Program*, vol. 14, n° 3, p. 130-137, 1980.

Pratt A. W., Pacak M. G., « Automated processing of medical English », *Proceedings of the 1969 conference on Computational linguistics*, p. 1-23, 1969.

Rissanen J., « Universal coding, information, prediction, and estimation », *IEEE Transactions on Information Theory*, vol. 30, n° 4, p. 629-636, 1984.

Robins R. H., « In defense of WP », *Transactions of the Philological Society 1959*, vol. 58, p. 116-144, 1959.

Sagot B., « The Lefff, a freely available, accurate and large-coverage lexicon for French », *Proceedings of the 7th Language Resource and Evaluation Conference*, La Valette, Malte, 2010.

Snyder B., Barzilay R., « Unsupervised Multilingual Learning for Morphological Segmentation », *Proceedings of ACL-08*, Columbus, États-Unis, p. 737-745, June, 2008.

Spiegler S., Golenia B., Flach P., « PROMODES : A Probabilistic Generative Model for Word Decomposition », *Working Notes for the CLEF 2009 Workshop*, Corfu, Grèce, 2009.

Stroppa N., Yvon F., « Du quatrième de proportion comme principe inductif : une proposition et son application à l'apprentissage de la morphologie », *Traitement Automatique des Langues*, vol. 47, p. 33-59, 2006.

Stump G. T., *Inflectional Morphology. A Theory of Paradigm Structure*, Cambridge University Press, Cambridge, Royaume-Uni, 2001.

Stump G. T., « Heteroclisis and Paradigm Linkage », *Language*, vol. 82, p. 279-322, 2006.

Tepper M., Xia F., « Inducing Morphemes Using Light Knowledge », *Journal of ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 9, n° 3, p. 1-38, 2010.

Thackston W., « Sorani Kurdish : A Reference Grammar with Selected Readings », 2006, Publié en ligne.

Thornton A. M., « Towards a Typology of overabundance », December, 2010, Presented at the Décembrettes 7, Toulouse, France.

van den Bosch A., Daelemans W., « Memory-based morphological analysis », *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, p. 285-292, 1999.

Walther G., « A Derivational Account for Sorani Kurdish Passives. », 2011a, Communication au 4th International Conference on Iranian Linguistics (ICIL4). 17-19 juin 2011. Uppsala, Suède.

Walther G., « Latin Passive Morphology Revisited », 2011b, Communication au colloque de la la Linguistic Association of Great-Britain (LAGB 2011). 7-10 septembre 2011. Manchester, Royaume-Uni.

Walther G., « Measuring Morphological Complexity », *Linguistica*, 2011c. Internal and External Boundaries of Morphology. À paraître.

Xanthos A., *Apprentissage automatique de la morphologie — Le cas des structures racine-schème*, vol. 48 of *Sciences pour la Communication*, Peter Lang, 2008.

Zauner A., *Praktická príručka slovenského pravopisu*, Vydavateľstvo Osveta, Martin, Slovaquie, 1973.

Zweigenbaum P., Grabar N., « Liens morphologiques et structuration de terminologie », *Actes de IC 2000 : Ingénierie des Connaissances*, p. 325-334, 2000.