TASLA 2015

Table des matières

Construction du jeu d'étiquettes pour le parsing du serbe

Aleksandra Miletic, Cécile Fabre, Dejan Stosic
CLLE-ERSS, CNRS UMR 5263, Maison de la Recherche, Université de Toulouse-Jean Jaurès,
5, allées Antonio Machado, Toulouse Cedex 9
aleksandra.miletic@univ-tlse2.fr, cecile.fabre@univ-tlse2.fr, dstosic@univ-tlse2.fr

Résumé. Cet article présente la démarche utilisée pour la construction d'un jeu d'étiquettes syntaxiques destiné à l'élaboration d'un corpus d'entraînement pour le parsing du serbe dans le but de doter le corpus ParCoLab (corpus parallèle serbe-français-anglais) d'une annotation syntaxique. Vu que le serbe ne dispose pas encore de treebank, il est nécessaire d'élaborer manuellement un corpus d'entraînement. Comme la structure et la taille du jeu d'étiquettes peuvent affecter les résultats du parsing, la définition du jeu est une étape cruciale. Dans le choix des étiquettes, nous avons été guidés par deux principes : réconcilier les traditions grammaticales serbe et française pour des raisons techniques et théoriques et maintenir la comparabilité avec les jeux d'étiquettes élaborés pour d'autres langues slaves. Cette démarche aboutit à un jeu de 28 étiquettes qui assurent la cohérence des traitements dans les différents volets du corpus et la possibilité d'exploiter les outils développés pour d'autres langues dans l'élaboration du corpus d'entraînement.

Abstract.

Constructing a syntactic tagset for the parsing of Serbian
This article presents the process of the construction of a syntactic tagset for Serbian. This tagset is intended for the constitution of a training corpus for the parsing of Serbian, in the global aim of linguistic annotation of the ParCoLab corpus, a parallel corpus of Serbian, French and English. Since there are still no treebanks for Serbian, a manually annotated training corpus must be created. As the parsing results can be affected by the structure and size of the tagset, its definition is a crucial stage. In the tag selection process, we were guided by two main goals: to reconcile the Serbian and the French grammar tradition for technical and linguistic reasons and to maintain comparability with existing tagsets for other Slavic languages. This strategy led us to 28 tags that ensure the coherence of annotation between different subcorpora and allow for the exploitation of tools developped for other languages in the manual annotation process.

Mots-clés : Jeu d'étiquettes, parsing, serbe, corpus parallèle
Keywords: Tagset, parsing, Serbian, parallel corpus

1. Introduction

Dans le domaine du TAL, le serbe reste une langue sous-dotée même si de nombreux outils et ressources existent déjà (pour un aperçu, voir Krstev 2008). En matière d'annotation morphosyntaxique, le seul corpus annoté qui soit librement disponible est celui développé dans le cadre du projet MULTEXT-east (Krstev et al., 2004), et le premier étiqueteur consacré au traitement de cette langue est BTagger, distribué en 2012 (Gesmundo, Samardžić, 2012). Aujourd'hui, il n'existe pas de treebank pour le serbe, et les premières expériences de parsing de cette langue sont très récentes (Jakovljević et al., 2014). Ce paysage a été enrichi récemment par ParCoLab, un corpus de textes littéraires en français, serbe et anglais (Stosic, 2015). Ce corpus parallèle a un intérêt important aussi bien pour la linguistique que pour le TAL car il est susceptible de favoriser des études contrastives théoriques sur ces trois langues, et de devenir une ressource importante pour le développement de différents outils TAL, notamment dans le domaine de la traduction automatique et de la traduction assistée par ordinateur. Or, pour que ce potentiel soit réalisé, il est nécessaire de doter ParCoLab d'annotations linguistiques de différents niveaux. Comme on dispose des outils nécessaires à l'annotation morphosyntaxique de ce corpus (Miletic, 2013), on se propose de développer les ressources pour le parsing et d'enrichir ParCoLab avec une annotation syntaxique.

Etant donné la tradition bien établie du parsing pour le français et l'anglais, reflétée dans le nombre et la diversité des ressources et outils disponibles pour le traitement de ces deux langues (Abeillé et al., 2003, Candito et al., 2010, Marcus et al., 1993, Petrov et al., 2006), l'annotation de ces deux volets du corpus n'est pas considérée comme problématique. Il n'en est pas de même pour le serbe : comme il n'existe pas encore de treebank pour cette langue, l'annotation du volet serbe de ParCoLab avec un parser statistique exige d'abord le

développement des ressources nécessaires à ce type de méthodes, notamment d'un corpus d'entraînement annoté manuellement. Pour pouvoir entamer l'annotation manuelle, il faut d'abord définir le jeu d'étiquettes à utiliser, autrement dit, déterminer quelles relations syntaxiques seront identifiées et codées dans le corpus. Vu que le nombre d'étiquettes, ainsi que leur définition, peuvent affecter l'exactitude du parsing, cette étape n'est pas anodine et exige une réflexion linguistique approfondie. Dans cet article, nous présentons le jeu d'étiquettes que nous avons établi à cette fin.

Plusieurs principes ont guidé le travail d'élaboration du jeu d'étiquettes. Tout d'abord, l'un des usages prévus de ParCoLab est de servir de support à des recherches linguistiques pour la communauté scientifique serbe aussi bien que française. Il fallait donc réconcilier deux traditions grammaticales différentes. La nécessité d'avoir un jeu d'étiquettes comparable vient aussi des contraintes techniques puisque nous allons annoter le volet français de ParCoLab avec le parser Talismane (Urieli, 2013) et que nous envisageons de tester le même outil sur le sous-corpus serbe. Comme le jeu d'étiquettes de Talismane est basé sur celui du *French Treebank* en dépendances (Candito et al., 2009) (dorénavant FTBDep), il diffère de manière importante des jeux élaborés pour les langues slaves (Hajič et al., 1988, Merkler et al., 2013). Les compromis qui sont à faire doivent donc assurer la cohérence des traitements linguistiques dans les deux volets du corpus et permettre une comparaison plus directe des résultats du parsing. Enfin, nous avons jugé intéressant de maintenir une comparabilité avec les jeux d'étiquettes déjà existants pour d'autres langues proches du serbe, notamment le croate. Ceci nous laisse la possibilité, suggérée par les travaux de (Agić et al., 2013), d'exploiter les outils développés pour cette langue dans le traitement du serbe (voir section 3).

Nous tenons à souligner qu'il s'agit ici d'une version préliminaire du jeu d'étiquettes, basée sur une réflexion théorique. Il est par conséquent fort probable que la couverture des phénomènes syntaxiques du serbe ne soit pas parfaite. Pour contrer ce problème, avant d'entamer l'élaboration du corpus d'entraînement, notre jeu d'étiquettes sera mis à l'épreuve d'un échantillon conséquent de texte authentique, ce qui nous permettra d'identifier et combler les éventuelles lacunes.

Dans la suite de cet article, nous présentons d'abord quelques spécificités du serbe et les travaux existants en parsing de cette langue (section 2). Nous définissons ensuite notre problématique et donnons une description brève du corpus ParCoLab (section 3). Dans la section 4, nous présentons le jeu d'étiquettes dans sa totalité, en justifiant quelques choix qui nous ont semblé importants. Nous clôturons enfin cet article en donnant une conclusion et des pistes pour les travaux à venir (section 5).

2. Parsing du serbe

L'annotation syntaxique automatique (ou *parsing*) repose aujourd'hui généralement sur l'utilisation des parsers (logiciels d'analyse syntaxique) statistiques, qui effectuent l'apprentissage des règles d'annotation à partir d'un corpus d'entraînement. Il s'agit d'un échantillon de texte annoté manuellement qui permet au parser de déterminer la probabilité de différentes analyses syntaxiques d'une phrase. Ainsi, une fois lancé sur un texte inconnu, le parser est capable de sélectionner l'analyse la plus probable et peut être utilisé à annoter la totalité d'un corpus de manière automatique (Kübler et al., 2009).

Pour pouvoir entamer l'élaboration d'un corpus d'entraînement, il est d'abord nécessaire de déterminer quelles fonctions syntaxiques seront annotées. Les appellations attribuées à ces fonctions constituent ce que l'on nomme *jeu d'étiquettes*. La structure du jeu d'étiquettes dépend avant tout du fonctionnement syntaxique de la langue en question, mais elle est conditionnée aussi par les choix théoriques retenus et par les spécificités de l'apprentissage automatique. La taille du jeu peut varier en fonction de l'usage envisagé du corpus : les applications TAL favorisent souvent des jeux restreints, alors que l'exploitation d'un corpus dans le domaine linguistique exige une granularité plus fine. Vu que la taille du jeu et la définition des étiquettes peuvent affecter les résultats du parsing, la définition d'un jeu d'étiquettes constitue une étape préalable cruciale.

2.1. Spécificités du serbe

Tout comme les autres langues slaves, le serbe dispose d'une morphologie flexionnelle riche : à titre d'illustration, les noms varient selon le genre (masculin, féminin ou neutre), le nombre (singulier, pluriel et paucal) et le cas (nominatif, génitif, datif, accusatif, vocatif, instrumental ou locatif). En plus de ces trois catégories, les adjectifs sont également affectés par celles du degré de comparaison (positif, comparatif ou superlatif) et de l'aspect (défini ou indéfini). Tous les pronoms se déclinent, et certains distinguent le genre, le nombre et la personne (première, deuxième ou troisième). L'indication de certaines fonctions syntaxiques étant portée par le marquage casuel, la structure de la phrase est très flexible. Même si l'ordre des constituants canonique est SVO, les ordres SOV, VOS, VSO, OVS et OSV sont non seulement grammaticaux, mais fréquents

(cf. par exemple Stanojčić, Popović, 2011, p.367, Ivić 2005). Il est également possible d'avoir des constituants discontinus, comme dans l'exemple 1.

[Lep-u	{ste	kuć-u]	kupi-li}.
beau-ACC.SG.F	être[PRS.2PL]	maison-ACC.SG.F	acheter-PTCP.PL.M

Vous avez acheté une belle maison / C'est une belle maison que vous avez achetée.

Exemple 1 : Constituant discontinu en serbe

Ici, l'un des constituants est délimité par des crochets (*Lepu kuću*), et l'autre par des accolades (*ste kupili*). On voit donc qu'il est possible d'insérer le verbe auxiliaire entre l'adjectif épithète et le nom auquel cet adjectif est rattaché.

Il est généralement considéré que le parsing de ce type de langues doit être basé sur l'analyse en dépendances, et non pas sur l'analyse en constituants, cette dernière ne disposant pas de mécanismes pour gérer la discontinuité des constituants (Buchholz, Marsi, 2006, Nivre et al., 2007). Or, à la différence d'autres langues slaves comme le tchèque et le russe, qui ont une tradition importante en syntaxe de dépendances (cf. Sgall et al., 1986, Мельчук, 1995), la description syntaxique du serbe repose traditionnellement sur l'analyse en constituants (Stanojčić, Popović, 2011, Ivić, 2005). Ceci signifie qu'il n'existe pas encore de formalisme pour l'annotation du serbe en syntaxe de dépendances[1], ce qui peut être l'une des causes du manque de travaux sur le parsing du serbe.

En effet, le premier travail sur le parsing de cette langue a été publié très récemment (Jakovljević et al., 2014). Ces expériences initiales ont été effectuées sur un treebank en cours de développement, basé sur le corpus *AlfaNum* (Sečujski, 2009). Pour la constitution du treebank, Jakovljević et ses collaborateurs reprennent le jeu d'étiquettes de *Prague Dependency Treebank* (dorénavant PDT) (Hajič, 1998). Ce jeu contient 28 étiquettes, dont 15 annotent les fonctions syntaxiques principales comme sujet, objet, prédicatif nominal etc., alors que les étiquettes restantes sont consacrées aux éléments considérés comme auxiliaires : verbes auxiliaires, mots emphatiques, différents types de ponctuation. L'annotation des fonctions syntaxiques est de faible granularité : à titre d'exemple, il existe une seule étiquette Obj (objet), alors que le tchèque distingue l'objet direct et l'objet indirect. (Jakovljević et al., 2014) utilisent ce jeu d'étiquettes en le modifiant de manière minimale pour optimiser le traitement des particules interrogatives, de l'auxiliaire *hteti* 'vouloir' et de certains cardinaux.

L'avantage principal de l'utilisation du jeu d'étiquettes de PDT pour l'annotation du serbe réside dans le fait que le guide d'annotation de PDT est disponible sur internet. En effet, afin d'assurer la cohérence des annotations manuelles, l'élaboration d'un corpus d'entraînement nécessite un ensemble de règles de traitement détaillées, appelé le guide d'annotation. L'accès à un guide existant permet d'entamer l'élaboration du corpus d'entraînement sans devoir consacrer un temps important à la définition des règles d'annotation. La même approche a été utilisée dans l'élaboration des premiers treebanks du croate (Hrvatska ovisnosna banka stabala ou HOBS, cf. (Tadić, 2007)) et du slovène (Slovene Dependency Treebank, cf. (Džeroski et al., 2006)). Vu la proximité typologique du tchèque avec le croate et le slovène, il a été possible d'adapter le jeu élaboré pour PDT à l'annotation syntaxique de ces deux langues. Cependant, cette stratégie a été remise en question dans des travaux plus récents : suite aux remarques des annotateurs humains selon lesquelles le jeu de PDT n'était pas intuitif dans son application au croate (Berović et al., 2012), un nouveau jeu de 15 étiquettes a été établi. Ce jeu reste fondé sur les principes de base de PDT, avec la réduction de taille obtenue par la simplification du traitement de certains éléments. La pertinence de ces choix a été justifiée par les résultats sur l'accord inter-annotateurs présentés dans (Agić, Merkler, 2013), qui montrent que l'utilisation du jeu d'étiquettes réduit apporte une hausse de 7,22 points pour le score LAS[2], et de 2,13 points pour le score UAS[3]. Le nouveau jeu a ensuite été utilisé dans l'élaboration d'un corpus de messages électroniques du croate (Merkler et al., 2013) et de *SETimes.hr*, un treebank du croate basé sur des textes journalistiques (Agić, Merkler, 2013). On peut remarquer la même tendance dans le parsing du slovène : lors de l'élaboration du JOS Corpus (Jezikoslovno označevanje slovenščine 'Annotation linguistique du slovène', (Erjavec et al., 2010)), un deuxième treebank du slovène, il a été noté que les annotateurs humains avaient des difficultés à maintenir la cohérence des annotations avec le jeu d'étiquettes basé sur le PDT. Par conséquent, le JOS Corpus a été élaboré en utilisant un jeu minimaliste de 10 étiquettes (*id.*).

[1] Un des relecteurs nous a signalé les travaux de P. Mrazović, que malheureusement nous n'avons pas réussi à nous procurer avant de terminer cet article.

[2] *Labeled attachment score* : pourcentage des tokens pour lesquels le parser a bien identifié le gouverneur et le type de la relation.

[3] *Unlabeled attachment score* : pourcentage des tokens pour lesquels le parser a bien identifié le gouverneur, sans tenir compte du type de relation attribuée.

Etant donné ces expériences, nous avons décidé de ne pas fonder notre jeu d'étiquettes sur celui du PDT, mais d'en concevoir un nouveau, tout en cherchant à respecter les différentes contraintes posées par la nature plurilingue de notre corpus et son usage envisagé.

3. Méthodologie et données

3.1. Méthodologie de construction du jeu d'étiquettes

Comme il a été mentionné ci-dessus, nous avons sélectionné le parser Talismane (Urieli, 2013) pour effectuer l'annotation du volet français du corpus. Ce parser, paramétré et testé sur le français, atteint une exactitude de 86,9 - 88,0% pour le score LAS et de 89,5 - 90,4% pour le score UAS (*id.*, p. 154), en fonction de la configuraiton utilisée. Ces résultats sont comparables à ceux obtenus par d'autres parsers disponibles pour le français, à savoir Berkeley (Petrov et al., 2006), MSTParser (McDonald et al., 2006), et MaltParser (Nivre et al., 2006), dont les performances sur le français sont présentées dans (Candito et al., 2010). Talismane a l'avantage de proposer également la possibilité de créer une approche hybride en intégrant des règles grammaticales afin d'améliorer la qualité de la sortie.

Nous envisageons également de paramétrer cet outil sur le corpus serbe pour tester sa capacité à s'adapter à une langue à ordre de constituants flexible. Si les résultats sont satisfaisants, il sera utilisé pour le parsing de la totalité du volet serbe de ParCoLab. Pour préserver l'intérêt scientifique de cette démarche et pouvoir comparer les résultats de l'outil sur ces deux langues, il est indispensable de maintenir un degré de comparabilité entre les deux jeux d'étiquettes. Cette démarche n'est pas simple, d'abord à cause des différences structurelles entre le français et le serbe, et ensuite à cause des différences entre les deux traditions grammaticales. Pour pouvoir faire les rapprochements nécessaires, nous avons été obligés d'adapter certaines analyses syntaxiques admises dans la tradition grammaticale serbe. Les plus importants de ces choix sont présentés et discutés dans la Section 4.

Un autre ensemble de compromis a été nécessaire pour respecter notre décision de garder notre jeu d'étiquettes proche du jeu croate utilisé dans SETimes.hr (Merkler et al., 2013), et ceci par souci d'accélérer l'élaboration du corpus d'entraînement. En effet, le croate dispose déjà des modèles de parsing : dans (Agić, Merkler, 2013), trois logiciels sont entraînés et testés sur le croate, et le modèle le plus performant, à savoir celui de MSTParser (McDonald et al., 2006), est mis à la disposition de la communauté scientifique. Etant donné la proximité syntaxique et morphologique du serbe et du croate, il est envisageable d'utiliser ce modèle pour effectuer une première annotation automatique du corpus d'entraînement, qui sera ensuite manuellement corrigée, à la fois pour éliminer les erreurs de parsing et pour rendre les annotations conformes à notre jeu d'étiquettes. Les résultats présentées dans (Agić et al., 2013) prouvent la pertinence de cette approche : dans ces expériences, MSTParser, entraîné exclusivement sur des données du croate, atteint le même niveau d'exactitude sur les échantillons du croate et du serbe. Afin de pouvoir tester cette possibilité, le jeu d'étiquettes ciblé doit garder un minimum de comparabilité avec le jeu intégré dans le modèle pour le croate de MSTParser. Sans cela, la correction de l'annotation de sortie exigerait trop d'ajustements et présenterait peu d'avantage par rapport à l'annotation manuelle pure.

En prenant en compte ces deux contraintes (comparabilité avec le jeu d'étiquettes de Talismane et celui de SETimes.hr), nous partons des fonctions syntaxiques traditionnellement utilisées en serbe (Stanojčić, Popović, 2011, Ivić, 2005) et établissons un jeu de 28 étiquettes (cf. § 4).

3.2. ParCoLab, corpus parallèle serbe-français-anglais

ParCoLab est un corpus parallèle trilingue. Il contient des textes originaux en serbe, français et anglais, ainsi que des traductions professionnelles de ces textes dans les deux autres langues du corpus. Aujourd'hui, il contient exclusivement des textes littéraires, mais une diversification en termes de genres est prévue pour un avenir proche, avec l'apport de textes provenant du web. Le corpus comporte à présent 1 650 501 tokens, dont 639 555 en serbe, 658 373 en français et 352 573 en anglais. Leur distribution par type de texte (original ou traduction) est donnée dans le Tableau 1.

Un échantillon de 150 000 tokens issu de la partie du corpus contenant des textes originaux en serbe a été transformé en corpus d'entraînement pour l'annotation morpho-syntaxique. Le jeu d'étiquettes morphosyntaxiques utilisé compte 47 étiquettes (Miletic, 2013), ce qui présente un compromis entre les deux jeux utilisés pour l'étiquetage du serbe dans la littérature : le premier est celui proposé dans le cadre du projet MULTEXT-East, qui encode toutes les informations morphosyntaxiques et compte plus de 900 tags (Krstev et al., 2004), et le deuxième est un jeu minimaliste de 15 tags, n'encodant que la partie du discours (Utvić, 2011)

Des 150 000 tokens que le sous-corpus comporte, 100 000 ont été annotés manuellement, et 50 000 ont fait l'objet d'un étiquetage automatique suivi d'une correction manuelle. Afin de profiter de la haute qualité des annotations morphosyntaxiques dans les expériences du parsing, c'est le même sous-corpus qui nous servira de base dans l'élaboration d'un corpus d'entraînement pour le parsing.[4]

	Volet français	**Volet serbe**	**Volet anglais**
	originaux	**originaux**	**originaux**
	113 973	459 708	229 044
	serbe → français	**français → serbe**	**français → anglais**
	388 326	104 310	123 529
	anglais → français	**anglais → serbe**	**serbe → anglais**
	156 074	75 537	-
TOTAL :	658 373	639 555	352 573

TABLEAU 1 : Nombre de tokens provenant des textes originaux et des traductions dans ParCoLab

Cependant, pour limiter le nombre d'étiquettes, le jeu de (Miletic 2013) encode seulement la partie du discours principale et sa sous-catégorie, en ajoutant le degré de comparaison pour les adjectifs et les adverbes. Comme une partie des fonctions syntaxiques en serbe sont indiquées par le marquage casuel, nous envisageons d'élaborer un module d'identification de cas qui nous permettra d'intégrer dans l'annotation morpho-syntaxique cette information cruciale pour le parsing.

4. Un nouveau jeu d'étiquettes pour le parsing du serbe

4.1. Présentation des étiquettes

Le jeu que nous proposons compte 28 étiquettes. Leur présentation accompagnée d'une brève définition de leur application est donnée dans le Tableau 2.

Etiquette		Définition	Exemple
1.	Pred	prédicat ; dans les temps composés, accordé au participe du verbe principal	*Filip **jede*** 'Filip **mange**' ; *Filip je **jeo*** 'Filip a **mangé**'
2.	AuxV	verbe auxiliaire dans les temps composés	*Filip **je** jeo* 'Filip **a** mangé'
3.	AuxVNeg	forme synthétique de verbe auxiliaire nié	*Filip **nije** jeo* 'Filip **n'a pas** mangé'
4.	Suj	sujet au nominatif	***Filip** jede* '**Filip** mange'
5.	SujLog	sujet logique exprimé au génitif, datif ou accusatif	cf. Exemple 4
6.	ObjDir	objet direct, exprimé à l'accusatif ou génitif	*Filip jede **jabuku*** 'Filip mange **une pomme**' ; *Filip pije **mleka*** 'Filip boit **du lait**'
7.	ObjIndir	objet indirect, exprimé au datif	*Filip daje jabuku **Milici*** 'Filip donne une pomme **à Milica**'
8.	ObjPrep	objet prépositionnel, réalisé sous forme d'un groupe prépositionnel régi par le verbe	*Filip misli **na porodicu*** 'Filip pense **à sa famille**'
9.	ComplInf	complément d'un verbe modal ou aspectuel sous forme d'un infinitif	cf. Exemple 11
10.	DepVAdv	dépendant adverbial d'un verbe	*Filip **vredno** radi* 'Filip travaille **assidument**'
11.	DepVCas	dépendant d'un verbe sous forme d'un GN fléchi	*Milica ide **kući*** 'Milica va **à la maison'**
12.	DepVPrep	dépendant d'un verbe sous forme d'un GP	*Milica radi **sa Filipom*** 'Milica travaille **avec Filip'**
13.	AttrSuj	complément d'un verbe attributif qui s'accorde avec le sujet	cf. Exemple 2
14.	AttrObj	complément d'un verbe attributif qui	cf. Exemple 3

[4] Davantage d'informations sur le corpus en question et l'étiquetage morphosyntaxique de ParCoLab peuvent être trouvées dans (Balvet et al., 2014) et (Miletic, 2013).

		s'accorde avec l'objet	
15.	Ep	épithète d'un nom sous forme d'un adjectif qui s'accorde avec le nom	*Alan kupuje **lepu** kuću* 'Alain achète une **belle** maison'
16.	DepNCas	dépendant d'un nom sous forme d'un GN	*kuća **moga strica*** 'maison de mon oncle'
17.	DepNPrep	dépendant d'un nom sous forme d'un GP	*kolač **sa višnjama*** 'gâteau **aux cerises**'
18.	DepAdjAdv	dépendant adverbial d'un adjectif	***vrlo** lepa kuća* '**très** belle maison'
19.	Ap	apposition	*Ivo Andrić, **pisac i nobelovac*** 'Ivo Andric, **écrivain et prix nobel**'
20.	EpDet	épithète occupant une place non canonique, en tête de phrase et/ou détaché du nom par des virgules	***Umoran**, Filip se vratio kući* '**Fatigué**, Filip est rentré'
21.	Sub	tout subordonnant sauf les relatifs	*Javiću se **kad** stignem* 'Je t'appellerai **quand** je serai arrivé'
22.	PredRel	prédicat de proposition relative	*Video sam čoveka koji se **doselio*** 'J'ai vu l'homme qui **vient d'emménager'**[5]
23..	Coord	conjonction de coordination	*Filip **i** Alan se smeju* 'Filip **et** Alain rient'
24.	DepCoord	éléments coordonnés sauf le premier (voir ci-dessous)	*Filip i **Alan** se smeju* 'Filip et **Alain** rient'
25.	CPrep	complément d'une préposition	*kolač sa **višnjama*** 'gâteau aux **cerises**'
26.	Elp	ellipse	*Stigao je **umoran*** '[~~il~~] est arrivé **fatigué**'[6]
27.	Neg	particule de négation	*Filip **ne** jede jabuku* 'Filip **ne** mange **pas** la pomme'
28.	Punc	ponctuation	*Filip se vratio**!*** 'Filip est rentré **!**'

TABLEAU 2 : Jeu d'étiquettes proposé

Ce jeu contient 13 étiquettes de plus que celui utilisé pour le croate dans SETimes.hr (Agić et al., 2013). Cependant, pour la majorité des étiquettes, il s'agit simplement d'une augmentation de granularité. Le Tableau 3 résume les différences principales.

SETimes.hr	**ParCoLab**
Sb (sujet)	Suj
	SujLog
Obj (objet)	ObjDir
	ObjIndir
	ObjPrep
Atr (modifieur nominal)	Ep
	DepNCas
	DepNPrep

TABLEAU 3: Correspondances d'étiquettes entre SETimes.hr et ParCoLab

Les distinctions faites dans notre jeu d'étiquettes sont basées sur des critères formels, notamment le cas et la catégorie du constituant (cf. Tableau 2). Par conséquent, la conversion de l'étiquette globale utilisée par SETimes.hr vers les étiquettes plus spécifiques de notre jeu peut se faire de manière automatique. Des différences plus fondamentales concernent le traitement de la fonction désignée dans SETimes.hr comme l'attribut verbal, des groupes prépositionnels et de la coordination, ce qui sera expliqué plus en détail dans la sous-section 4.2.

Quant à la comparabilité de notre jeu d'étiquettes avec celui de FTBDep, repris par Talismane, nous constatons d'abord que leurs tailles respectives sont proches. Le jeu de FTBDep compte 21 étiquettes de base, utilisées pour l'annotation automatique, et 8 étiquettes plus spécifiques, réservées à l'annotation manuelle (Candito et al., 2009). Nous prendrons ici en considération seulement les 21 étiquettes destinées au parsing.

[5] Le traitement des relatives est repris de FTBDep. Le prédicat de la relative est rattaché au prédicat de la proposition indépendante. Ceci permet d'annoter le relatif avec la fonction qu'il exerce au sein de la proposition relative. Sinon il serait annoté simplement en tant que subordonnant et son rôle syntaxique dans la relative serait perdu.

[6] Nous reprenons le traitement de l'ellipse du PDT. Si un constituant est absent de la phrase, les nœuds qui dépendraient de lui sont annotés avec l'étiquette Elp et rattachés à l'endroit où le nœud manquant se trouverait dans l'arbre. Dans l'exemple donné, c'est l'épithète détaché *umoran* 'fatigué' qui porterait l'étiquette Elp, vu que le sujet duquel il dépend n'est pas présent dans la phrase.

A la différence de FTBDep, qui dispose d'une étiquette pour le sujet, nous faisons la distinction entre le sujet et le sujet logique. Comme indiqué ci-dessus, cette opposition est basée sur des critères morphosyntaxiques et une correspondance directe peut être établie entre les étiquettes de notre corpus et celle de FTBDep. Les étiquettes pour l'attribut du sujet (*ats*) et l'attribut de l'objet (*ato*) de FTBDep ont des équivalents directs dans les étiquettes *AttrSuj* et *AttrObj*. Le traitement de la fonction objet est proche dans les deux jeux : l'étiquette *obj* de FTBDep (objet direct) correspond à celle de *ObjDir* dans notre jeu, *a_obj* et *de_obj* (objet indirect introduit par *à* ou *de*, respectivement) sont équivalentes de *ObjIndir*, et *p_obj* (objet introduit par une préposition autre que *à* et *de*) correspond à *ObjPrep*. La différence la plus importante entre ces deux jeux d'étiquettes concerne un aspect de l'application de l'étiquette *obj* dans FTBDep, et qui n'est pas repris dans ParCoLab. Ce point est discuté dans la sous-section suivante.

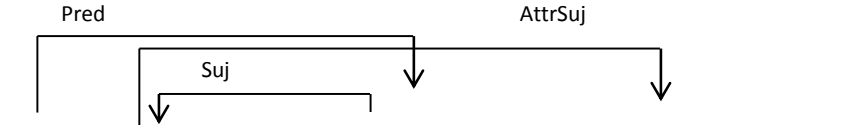
Un autre point commun que notre jeu d'étiquettes partage avec SETimes.hr et FTBDep est la décision de ne pas encoder la différence entre les compléments et les modifieurs. Comme cette distinction ne repose pas clairement sur des critères de surface accessibles à un parser, et qu'elle peut être problématique même pour les annotateurs humains, nous l'avons omise de notre jeu d'étiquettes. En revanche, nous considérons ces différentes fonctions comme des dépendants et précisons dans les étiquettes la catégorie de leur gouverneur et celle du dépendant lui-même. Ainsi, l'étiquette *DepNCas* désigne le dépendant d'un nom qui a la forme d'un GN fléchi, alors que *DepVPrep* indique le dépendant d'un verbe sous forme d'un GP. La structure des étiquettes permettra par la suite de faire des regroupements des fonctions s'il se prouve que la granularité actuelle est trop fine.

4.2. Justification des choix

Cette partie de l'article est consacrée à la justification de plusieurs choix faits dans la construction du jeu d'étiquettes. Certains d'entre eux sont techniques, conditionnés par les caractéristiques intrinsèques des algorithmes de parsing, alors que d'autres concernent plutôt l'analyse linguistique de différentes fonctions. Dans chacun des points discutés, nous essaierons d'expliciter notre position par rapport à la tradition grammaticale serbe, ainsi que par rapport aux jeux d'étiquettes de SETimes.hr et de FTBDep.

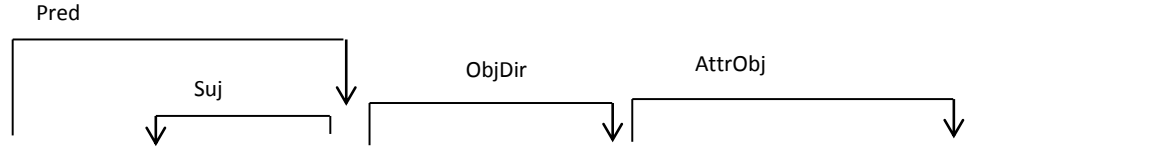
Attribut du sujet et attribut de l'objet direct

Ces deux fonctions ne sont pas reconnues dans les travaux sur la syntaxe du serbe. Elles sont réparties entre les fonctions des prédicatifs nominal, complémentaire et optionnel, qui correspondent respectivement aux compléments du verbe *biti* 'être', ceux des autres verbes essentiellement attributifs (cf. (Riegel, 1981)) comme *proglasiti (se)* '(se) proclamer', *smatrati (se)* '(se) considérer', *prozvati (se)* '(se) nommer', et ceux des verbes occasionnellement attributifs (*id.*). Cependant, il n'y a pas de critères formels pour distinguer ces constituants : les trois favorisent la position post-verbale (quoiqu'ils puissent être placés devant le verbe), admettent des dépendants sous forme de groupe nominal ou adjectival et si le dépendant se réalise sous forme d'un groupe adjectival, il prend les marques du genre et du nombre soit du sujet soit de l'objet direct. Par conséquent, nous remplaçons la distinction traditionnelle citée ci-dessus par celle du jeu d'étiquettes de FTBDep et introduisons les étiquettes pour l'attribut de sujet (*AttrSuj*) et l'attribut d'objet (*AttrObj*) (cf. exemples 2 et 3).



'Marko devient stressé.'

Exemple 2: Etiquette AttrSuj

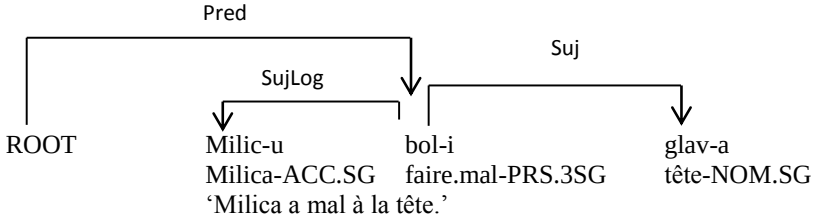


'Ils considèrent Marko intelligent.'

Exemple 3: Etiquette AttrObj

Sujet logique

En serbe, le sujet est typiquement exprimé au nominatif et désigne l'agent du processus ou l'expérienceur de l'état décrit par le verbe (*Milica čita knjigu* 'Milica-NOM lit livre-ACC'). Cependant, un groupe de verbes exprimant un état physique ou mental exigent que leur expérienceur soit au datif ou à l'accusatif (cf. exemple 4). Ce constituant est désigné dans la littérature comme sujet logique (Ivic, 2005, Stanojčić, Popović, 2011).



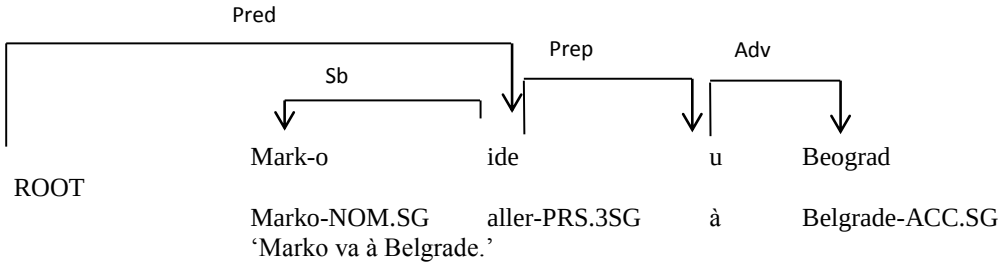
Exemple 4: Etiquette SujLog

Comme il existe un critère formel de distinction par rapport au sujet au nominatif, nous avons décidé d'introduire cette distinction dans notre jeu d'étiquettes : le sujet typique au nominatif sera annoté comme *Suj*, alors que pour le sujet logique on utilisera l'étiquette *SujLog*.

On peut remarquer que formellement ce constituant peut coïncider avec celui d'*ObjDir*, les deux étant des GN au génitif ou à l'accusatif. Ils montrent cependant des comportements différents quant à la linéarisation : le sujet logique est typiquement antéposé au verbe, avec une préférence pour la position initiale dans la phrase, alors que la position canonique de l'objet direct est à droite du verbe. Il est vrai que les deux sont mobiles et peuvent prendre la position typique de l'autre si la focalisation de la phrase l'exige. Il reste donc à voir si un parser sera capable d'opérer cette distinction.

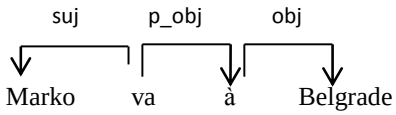
Groupes prépositionnels

Les jeux d'étiquettes de SETimes.hr et FTBDep adoptent deux approches différentes pour le traitement des groupes prépositionnels. Dans SETimes.hr, toutes les prépositions sont liées à leur gouverneur par la relation *Prep*, alors que c'est le complément de la préposition qui porte l'étiquette de la fonction exercée par le groupe prépositionnel dans la phrase (cf. Exemple 5).



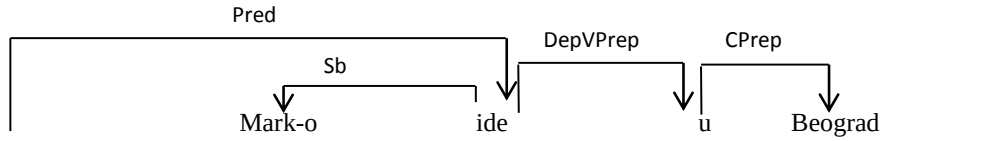
Exemple 5: Traitement des groupes prépositionnels dans SETimes.hr

En revanche, dans FTBDep, on annote la préposition avec la fonction du groupe prépositionnel, alors que le complément de la préposition est annoté en tant que *obj* (cf. exemple 6).



Exemple 6: Traitement des groupes prépositionnels dans FTBDep

Nous reprenons l'approche du FTBDep avec une différence : au lieu d'étendre le domaine d'application de l'étiquette de l'objet direct aux compléments de préposition, nous définissions une étiquette spécialisée, *CPrep*. Ceci résulte dans le traitement présenté dans l'exemple 7.

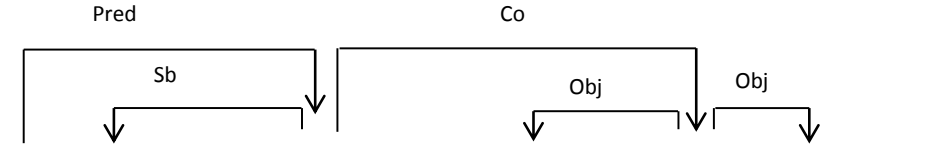


'Marko va à Belgrade.'

Exemple 7: Traitement des groupes prépositionnels dans ParCoLab

Coordination

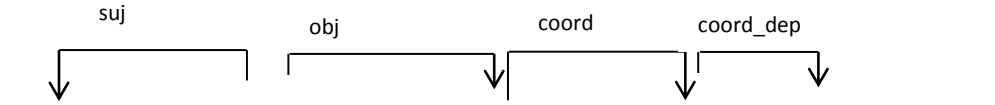
Les formalismes d'annotation de SETimes.hr et de FTBDep diffèrent également dans leur traitement des constructions de coordination. Le premier reprend l'approche proposée par PDT, qui consiste à utiliser la fonction *Co* pour relier la conjonction de coordination au gouverneur des constituants coordonnés, et d'ensuite relier les coordonnés à la conjonction par l'étiquette de la fonction qu'ils exercent dans la phrase (cf. exemple 8).



'Alain mange des poires et des pommes.'

Exemple 8: Traitement de coordination dans SETimes.hr

Bien qu'en accord avec l'intuition linguistique, ce traitement a des défauts du point de vue technique : ici, le parser est obligé de déterminer si la forme *kruške* fait partie d'une coordination avant d'identifier l'objet du verbe *jede*. Autrement dit, le parser doit reconnaître la coordination avant de pouvoir décider quel est le statut des coordonnés (Urieli, 2013). FTBDep propose une alternative que, tout en étant moins conforme à l'intuition linguistique, permet un traitement plus simple pour le parser (cf. exemple 9).



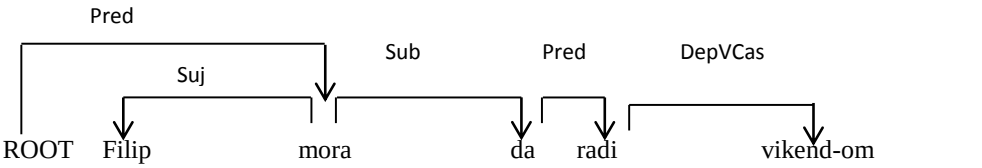
'Alain mange des poires et des pommes.'

Exemple 9: Traitement de la coordination dans FTBDep

Avec ce traitement, le parser identifie d'abord la fonction du premier coordonné pour passer ensuite à l'analyse de la construction coordonnée. Même si la fonction des coordonnés autres que le premier n'est pas explicitement notée, elle peut être récupérée de l'étiquette du premier coordonné. Pour notre jeu d'étiquettes, nous adoptons cette approche et utilisons par conséquent l'étiquette *Coord* pour lier la conjonction de coordination au premier coordonné, et l'étiquette *DepCoord* pour lier tous les coordonnés sauf le premier à la conjonction.

Traitement du prédicat complexe

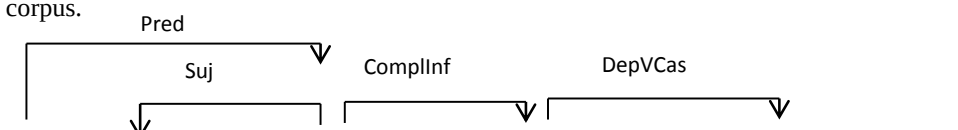
La tradition grammaticale serbe considère comme « prédicats complexes » les constructions avec les verbes modaux et aspectuels (Stanojčić, Popović, 2011, p.269). Ces verbes sont le plus couramment complémentés par la construction *da + Vprezent* 'que +V présent', comme dans l'exemple 10.



'Filip doit travailler le weekend.'

Exemple 10: Prédicat complexe sous forme *da + Vprezent* 'que + Vprésent'

Cependant, une complémentation en infinitif est également possible. Pour étiqueter cette construction alternative, deux traitements différents peuvent être envisagés. Il est possible de faire un rapprochement avec la construction équivalente *da + Vprezent* et d'annoter l'infinitif complément du verbe principal comme une subordonnée. Néanmoins, ceci veut dire que l'étiquette *Sub*, qui est dédiée à l'annotation des subordonnants, deviendrait également applicable aux verbes. De même, ses dépendants changeraient de manière importante : dans son emploi canonique, cette étiquette a un descendant *Pred* qui correspond au prédicat de la subordonnée, auquel sont ensuite rattachés les dépendants typiques du prédicat. Si on appliquait cette étiquette à l'infinitif, il n'y aurait plus de descendants *Pred*, et ce sont les dépendants du prédicat qui seraient attachés directement à l'étiquette *Sub*. Pour éviter la multiplication des contextes possibles pour cette étiquette, nous choisissons de considérer qu'il s'agit d'un complément verbal spécifique et introduisons une nouvelle étiquette, *ComplInf*. Cette approche nous permet également de maintenir la distinction linguistique entre ces deux constructions dans le corpus.



Exemple 11 : Utilisation de l'étiquette ComplInf

La discussion menée dans cette section montre que le jeu d'étiquettes que nous proposons respecte certaines distinctions traditionnellement admises dans la syntaxe du serbe (par exemple, pour le sujet et l'objet). Néanmoins, quelques adaptations ont également été nécessaires, notamment pour les fonctions regroupées sous le nom du prédicatif dans la grammaire serbe. Ces compromis sont justifiés par la contrainte de comparabilité entre notre jeu d'étiquettes, celui de FTBDep et celui de SETimes.hr. Grâce à un degré de comparabilité élevé, nous espérons à la fois maintenir une cohérence d'annotation entre les volets serbe et français de ParCoLab, tester les performances de Talismane sur le serbe, et exploiter les ressources existantes pour le croate dans l'élaboration du corpus d'entraînement.

5. Conclusion

Ce travail présente la première version d'un jeu d'étiquettes syntaxiques pour l'élaboration d'un corpus d'entraînement pour le parsing du serbe. Il s'inscrit dans le projet de doter le corpus parallèle serbe-français-anglais ParCoLab d'une couche d'annotation syntaxique.

En identifiant les fonctions syntaxiques qui doivent être représentées par le jeu, nous avons pris en compte les fonctions traditionnellement utilisées dans la syntaxe du serbe, mais nous avons également été guidés par le besoin de maintenir la comparabilité avec les jeux d'étiquettes d'un treebank du croate SETimes.hr et de French Treebank en dépendances. Cette approche a été choisie avec deux objectifs principaux : d'abord, nous souhaitons utiliser les modèles de parsing développés pour le croate sur notre corpus d'entraînement et accélérer ainsi son élaboration ; deuxièmement, une fois le corpus d'entraînement prêt, nous prévoyons de tester le parser Talismane (Urieli, 2013) et, si ses performances sont satisfaisantes, l'utiliser pour annoter la totalité du sous-corpus serbe de ParCoLab. Cette démarche aboutit à un jeu de 28 étiquettes. Cette taille dépasse celle des jeux de SETimes,hr et de FTBDep (15 et 21 étiquettes respectivement), mais la structure de notre jeu permet d'établir des correspondances nécessaires et a l'avantage de rendre possible la représentation des sous-types des fonctions syntaxiques principales (sujet, objet, dépendants nominaux, etc.).

Comme nous l'avons déjà indiqué, il s'agit d'une proposition initiale, fondée sur une réflexion théorique. Pour vérifier la pertinence de nos choix et combler d'éventuelles lacunes, nous testerons ce jeu d'étiquettes sur un échantillon du corpus. Une fois la version finale du jeu arrêtée, elle sera utilisée pour l'annotation manuelle du corpus d'entraînement.

Références

ABEILLE, A., CLEMENT, L., TOUSSENEL, F. (2003). Building a treebank for French. Dans: A. Abeillé, éd. *Treebanks.* Dordrecht : Kluwer, 165-187.

AGIC, Ž., MERKLER, D. (2013). Three Syntactic Formalisms for Data-Driven Dependency Parsing of Croatian. *LNCS* 8082, 560-567.

AGIC, Ž., MERKLER, D., BEROVIC, D. (2013). Parsing Croatian and Serbian by Using Croatian Dependency Treebanks. Actes de *SPMRL* à *EMNLP*, 22-33.

BALVET, A., STOSIC, D., MILETIC, A. (2014). TALC-sef, A Manually-Revised POS-Tagged Litterary Corpus in Serbian, English, and French. Actes de *LREC 2014*, 4105-4110.

BEROVIC, D., AGIC, Ž.,TADIC, M. (2012). Croatian Dependency Treebank: Recent Development and Initial Experiments. Actes de *LREC 2012*, 1902-1906.

BUCHHOLZ, S., MARSI, E. (2006). CoNLL-X shared task on multilingual dependency parsing. Actes de *Tenth Conference on Computational Natural Language Learning*, 149-164.

CANDITO, M., CRABBE, B., FALCO, M. (2009). Dépendances syntaxiques de surface pour le français. *Rapport technique, Université Paris 7.*

CANDITO, M., NIVRE, J., DENIS, P., HENESTROZA ANGUIANO, E. (2010). Benchmarking of Statistical Dependency Parsers for French. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 108-116.

DZEROSKI, S., ERJAVEC, T., LEDINEK N., PAJAS, P., ŽABOKRTSKY, Z., ŽELE, A. (2006). Towards a Slovene Dependency Treebank. Actes de *LREC 2006*, 1388-1391.

ERJAVEC, T. (2004). *MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora.* Actes de *LREC 2004*, 1535-1538.

ERJAVEC, T., FISER , D., KREK, S., LEDINEK, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. Actes de *LREC2010*, 1806-1809.

GESMUNDO, A., SAMARDZIC, T. (2012). Lemmatising Serbian as a category tagging with bidirectional sequence classification. Actes de *LREC 2012*, 2103-2106.

HAJIC, J. (1998). Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. *Issues of Valency and Meaning.* Prague : Karolinum, 106-132.

HAJIC, J., HAJICOVA, E. (1997). Syntactic tagging in the Prague Treebank. *Proceedings of the Second European Seminar "Language Applications for a Multilingual Europe"*, 55-68.

HAJIC, J., PANEVOVA, J., BURANOVA, E., URESOVA, Z., BEMOVA, A. (1999). Annotations at analytical level: Instructions for annotators. *Rapport technique, UK MFF UFAL, Prague.*

IDE, N., VERONIS, J. (1994). MULTEXT (Multilingual text tools and corpora). *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, 588-592.

IVIC, M. éd., 2005. *Sintaksa savremenog srpskog jezika.* Beograd: Institut za srpski jezik SANU.

JAKOVLJEVIC, B., KOVACEVIC, A., SECUJSKI, M., MARKOVIC, M. (2014). A Dependency Treebank for Serbian: Initial Experiments. *Speech and Computer Lecture Notes in Computer Science* 8773, 42-49.

KRSTEV, C. (2008), *Processing of Serbian. Automata, Texts and Electonic Dictionaries*, Belgrade, Faculty of Philology, University of Belgrade.

KRSTEV, C., VITAS, D., ERJAVEC, T. (2004). MULTEXT-East resources for Serbian. Actes de *7. mednarodne multikonference Informacijska druzba IS 2004 Jezikovne tehnologije*, 108-114.

KUBLER, S., MCDONALD, R., NIVRE, J. (2009). Dependency parsing. *Synthesis lectures on Human Language Technologies*, 1(1), 1-127.

LEDINEK, N., ŽELE, A. (2005). Building of the Slovene dependency treebank corpus according to the Prague dependency treebank corpus. *Proceedings of Grammar and Corpus*.

MARCUS, M. P., SANTORINI, B., MARCINKIEWICZ, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313-330.

MCDONALD, R., LERMAN, K., PEREIRA, F. (2006). Multilingual Dependency Parsing with a Two-Stage Discriminative Parser. *Proceedings of the Tenth Conference on Computational Natural Laguage Learning*, 216-220.

MEL'CUK, I. (1988). *Dependency Syntax: Theory and Practice.*State University of New York Press.

MEL'CUK, I. (2011). Dependency in language. *Proceedings of DepLing 2011*, 1-16.

MERKLER, D., AGIC, Ž., AGIC, A. (2013). Babel Treebank of Public Messages in Croatian. *Procedia-Social and Behavioral Sciences* 95, 490-497.

MILETIC, A. (2013). Annotation semi-automatique en parties du discours d'un corpus littéraire serbe. *Mémoire de Master, Université Charles de Gaulle Lille 3*.

NIVRE, J., HALL, J., KUBLER, S., MCDONALD, R., NILSSON, J., RIEDEL, S., YURET, D. (2007). The CoNLL 2007 shared task on dependency parsing. *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*, 915-932.

NIVRE, J., HALL, J., NILSSON, J. (2006). MaltParser A Data-Driven Parser-Generator for Dependency Parsing. *Proceedings of LREC 2006* 6, 2216-2219.

PETROV, S., BARRETT, L., THIBAUX, R., KLEIN, D. (2006). Learning accurate, compact, and interpretable tree annotation. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 433-440.

RIEGEL, M. (1981). Verbes essentiellement ou occasionnellement attributifs. *L'information grammaticale* 10, 23-27.

RIEGEL, M., PELLAT, J.-C., RIOUL, R. (1999) *Grammaire méthodique du français*, 5e éd. mise à jour. Paris : Presses Universitaires de France.

SECUJSKI , M. (2009). Automatic part-of-speech tagging of texts in the Serbian language. *Thèse de doctorat, Faculté des Sciences Techniques de Novi Sad*.

SGALL, P., HAJICOVA, E., PANEVOVA, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspect.* Dordrecht : Kluwer.

STANOJCIC, Ž., POPOVIC, L. (2011). *Gramatika srpskog jezika*. 14 éd. Beograd: Zavod za udžbenike.

STOSIC, D. (2015). ParCoLab (beta), A Parallel Corpus of French, Serbian and English. *Toulouse, France: CLLE-ERSS, CNRS & Université de Toulouse 2.* (http://parcolab.univ-tlse2.fr)

TADIC, M. (2000). Building the Croatian-English parallel corpus. *Proceedings of LREC 2000*, 523-530.

TADIC, M. (2007). Building the Croatian Dependency Treebank: the initial stages. *Suvremena Lingvistika* 33(63), 85-92.

URIELI, A. (2013). Analyse syntaxique robuste du français : concilier méthodes statistiques et connaissances linguistiques dans l'outil Talismane. *Thèse de doctorat, Université Toulouse II le Mirail.*

UTVIC, M. (2011). Annotating the Corpus of contemporary Serbian. *Proceedings of INFOtheca '12*, 36-47.

VITAS, D., KRSTEV, C., OBRADOVIC I., POPOVIC, LJ., PAVLOVIC-LAZETIC, G. (2012). The Serbian Language in the Digital Age. *META-NET White Paper Series*. Springer. http://www.meta-net.eu/whitepapers

МЕЛЬЧУК, И. А. (1995). *Русский язык в модели «Смысл ↔ Текст».* Wiener Slawistischer Almanach/ Škola «Jazyki russkoj kul´tury»: Vienne/Moscou.

Acquisition non supervisée de ressources morphologiques en ukrainien

Natalia Grabar[1] Thierry Hamon[2]
(1) CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France
natalia.grabar@univ-lille3.fr
(2) LIMSI-CNRS, BP133, Orsay; Université Paris 13, Sorbonne Paris Cité, France
hamon@limsi.fr

Résumé. La disponibilité de ressources morphologiques est un besoin important et récurrent car elles permettent le développement des outils et applications de TAL dans une langue. De telles ressources fournissent, en effet, les informations de base dont ces outils ont besoin pour effectuer des traitements plus évolués (recherche d'information, étiquetage morpho-syntaxiques, etc). Nous proposons d'effectuer l'acquisition de ressources morphologiques pour la langue ukrainienne, qui est une langue peu dotée actuellement. La méthode proposée exploite des corpus afin d'en extraire les mots qui sont liés morphologiquement entre eux. La force d'association entre ces mots indique la probabilité du lien morphologique et sémantique entre eux. Nous utilisons trois corpus (littéraire, médical et encyclopédique) et évaluons les résultats obtenus. Selon les corpus, la précision varie entre 67 % et 86 %. Les résultats sont aussi comparés entre les corpus, ce qui montre que la redondance est assez faible. La ressource actuellement disponible contient 3 315 paires de mots validées.

Abstract.
Unsupervized acquisition of morphological resources for Ukrainian.
Availability of morphological resources is an important and recurrent need because they allow the development of NLP tools and applications for a given language. Indeed, such resources provide basic information which are necessary for such tools for performing more sophisticated treatments (information retrieval, morpho-syntactic tagging, etc). We propose to acquire morphological resources for Ukrainian language, that is under-resourced at the time being. The method proposed exploits corpora in order to extract words that are related morphologically between them. The association strength between these words indicates their probability to have a morphological and semantic relation between them. We use three corpora (literary, medical and general-language) and evaluate the results obtained. According to corpora, precision varies between 67% and 86%. The results from different corpora are also compared, which shows that there is little redundancy between the copora. The currently available resource contains 3,315 validated pairs of words.

Mots-clés : Ukrainien, langues peu dotées, corpus, morphologie, acquisition de ressources, méthodes non supervisées.

Keywords: Ukrainian, low-resourced languages, corpora, morphology, acquisition of resources, unsupervised methods.

1 Introduction

Les ressources morphologiques constituent une connaissance de base pour plusieurs applications TAL. Souvent, il s'agit de la première brique qui est construite et utilisée dans la chaîne de traitement. Voila quelques exemples de telles applications :

– En *étiquetage morpho-syntaxique et lemmatisation*, il est nécessaire d'avoir un lexique approprié pour bien analyser les mots. Il est ainsi important de pouvoir reconnaître les formes fléchies d'un mot donné et de

déduire leur lemme. Pour les langues à morphologie riche en particulier, la possibilité de repérer les suffixes et flexions des mots permet de déduire et de désambiguïser leur étiquette morpho-syntaxique ;
– En *recherche et extraction d'information*, les besoins et objectifs dépassent la morphologie flexionnelle. En effet, il est souvent nécessaire d'aller au-delà des formes fléchies et de détecter également les liens entre les formes dérivées ou même composées. Typiquement, cela permet d'augmenter le rappel des systèmes automatiques et de collecter plus de réponses ou de documents pertinents ;
– Le *traitement de mots inconnus* concerne une multitude d'applications TAL. La raison principale est que les dictionnaires et ressources existants sont souvent incomplets. Tandis que, si des informations morphologiques sur les mots sont disponibles, celles-ci peuvent être très utile pour les traitements automatiques, notamment pour induire leur catégorie syntaxique ou leur sémantique ;
– En *reconnaissance de la parole*, les ressources par familles de mots peuvent être très utiles afin de désambiguïser une séquence ou bien de trouver le candidat le plus convenable pour un contexte.

Pour plusieurs langues, de telles ressources sont maintenant disponibles et largement utilisées, comme par exemple CELEX (Burnage, 1990) pour l'allemand, l'anglais et le néerlandais, Démonette (Hathout & Namer, 2014), lexique.org [1] et Lefff (Sagot *et al.*, 2006) pour le français, Morph-it (Zanchetta & Baroni, 2005) pour l'italien, etc. De telles ressources comportent au moins les informations flexionnelles sur le lexique d'une langue, comme les formes des noms {*président*; *présidents*}, adjectifs {*présidentiel*; *présidentielle*} ou verbes {*présider*; *présidons*}. Il est beaucoup plus rare de disposer de ressources qui permettent de relier aussi les formes dérivationnelles {*président*; *présidentiel*} ou compositionnelles {*président*; *présidoscopie*}. Notons que dans les domaines de spécialité la question de ressources morphologiques occupe également une place importante (McCray *et al.*, 1994; Grabar & Zweigenbaum, 1999; Zweigenbaum *et al.*, 2003), car les langues de spécialité comportent un lexique spécifique souvent absent des dictionnaires standards de la langue générale.

En plus de ressources morphologiques, plusieurs méthodes ont été proposées pour l'acquisition de ressources morphologiques. Parmi les approches existantes, nous pouvons par exemple mentionner les suivantes (une méthode donnée peut combiner plusieurs principes et approches) :
– exploitation des associations entre les mots dans les corpus (Xu & Croft, 1998; Zweigenbaum *et al.*, 2003) ;
– exploitation des propriétés distributionnelles des mots dans les corpus (Claveau & Kijak, 2014) ;
– exploitation des distributions de lettres dans les mots pour détecter les frontières des morphèmes et bases (Déjean, 1998; Urrea, 2000; Schone & Jurafsky, 2001) ;
– exploitation des analogies dans la formation des mots pour déduire ou générer de nouvelles formes et élargir ansi le dictionnaire (Pirrelli & Yvon, 1999; Grabar & Zweigenbaum, 1999; Hathout, 2001) ;
– exploitation de la fréquence du couple des suffixes de deux mots donnés, qui assure alors la fiabilité du lien sémantique entre ces mots (Gaussier, 1999) ;
– exploitation de dictionnaires existants et de la structure des informations dans les articles dictionnariques pour détecter les mots liés sémantiquement et morphologiquement (Pentheroudakis & Vanderwende, 1993; Hathout, 2001; Krovetz, 1993) ;
– exploitation de paires de termes en relations sémantiques (Grabar & Zweigenbaum, 1999) ;
– exploitation d'une base d'exemples et de méthodes supervisées pour déduire des règles morphologiques (van den Bosch *et al.*, 1996; Theron & Cloete, 1997; Pirrelli & Yvon, 1999).

Des outils pour l'analyse morphologique sont également disponibles pour plusieurs langues : le français (Namer, 2009), l'allemand [2], les langues Nguni (Bosch *et al.*, 2008; Pretorius & Bosch, 2009), les langues indiennes (Abeera *et al.*, 2012), le macédonien (Kostov, 2013), etc.

Nous pouvons voir qu'il s'agit d'un axe de recherche assez actif et que les langues de spécialité (comme la médecine), mais aussi les langues peu dotées (le macédonien, les langues Nguni et indiennes dans les travaux cités plus haut) peuvent disposer de ressources et d'outils pour le traitement des mots au niveau morphologique. Nous avons aussi vu qu'il existe plusieurs méthodes pour l'acquisition de ressources morphologiques et que, de ce fait, différents types de données peuvent être traités afin d'acquérir les ressources morphologiques.

Dans notre travail, nous proposons d'aborder la question de construction de ressources morphologiques pour l'ukrainien, qui est une langue slave et actuellement peu dotée. Nous allons exploiter les corpus de textes. Il

1. www.lexique.org
2. https://code.google.com/p/morphisto

s'agit de ressources librement disponibles et ne disposant pas d'annotations syntaxiques ou sémantiques. La méthode utilisée s'appuie sur les travaux antérieurs (Xu & Croft, 1998; Zweigenbaum *et al.*, 2003) et exploite les associations entre les mots. Plusieurs adaptations sont effectuées pour traiter la langue ukrainienne : encodage des corpus, segmentation des textes, quelques spécificités morphologiques de cette langue.

Nous proposons d'abord une description de la langue ukrainienne et indiquons quelques travaux existants (section 2). Nous présentons ensuite le matériel utilisé (section 3), et les étapes de la méthode (section 4). Nous décrivons et discutons les résultats obtenus (sections 5), et concluons avec des orientations pour les travaux futurs (section 6).

2 Spécificités de l'ukrainien

L'ukrainien fait partie de la famille des langues slaves et utilise un alphabet cyrillique composé de 33 lettres et l'apostrophe. Une des particularités de l'ukrainien est que l'apostrophe joue un rôle phonétique et non pas de séparation de mots. Par exemple, dans le mot об'єкт (objet), l'apostrophe permet de ne pas palataliser la consonne "б" devant la voyelle molle "є".

Comme c'est le cas de toutes les langues slaves, l'ukrainien est une langue morphologiquement riche. Par exemple, les informations flexionnelles sont utilisées pour décrire jusqu'à sept cas et trois genres pour les noms communs et propres, adjectifs, pronoms et certaines formes verbales. La morphologie dérivationnelle est également très présente dans la formation des constructions grammaticales (par exemple, aspect, temps) et lexicales. En (1) et (2), nous présentons quelques mots de deux séries, *marcher* et *fermer/ouvrir*, respectivement. Quant à la morphologie compositionnelle, elle est largement utilisée dans la langue ukrainienne, ce qui semble être le cas d'autres langues slaves également (Loginova-Clouet, 2014).

(1) хід (*marche*), вхід (*entrée*), вихід (*sortie*), захід (*est, coucher (de soleil), événement*), прихід (*arrivée*), перехід (*traversée, passage piéton*), відхід (*départ*), підхід (*approche*), дохід (*approche encore plus proche du but, les revenus*), прохід (*passage à l'intérieur d'un obstacle comme le bois, une haie*), обхід (*passage à coté, contournement*)

(2) критий (*couvert*), закритий (*fermé*), відкритий (*ouvert*), напівзакритий (*demi-fermé*), напіввідкритий (*demi-ouvert*), прикритий (*un peu fermé, recouvert*), перекритий (*séparé, bloqué*)

Ayant une morphologie riche, cela permet à l'ukrainien d'avoir un ordre des mots assez libre sans introduire pour autant d'effets stylistiques particuliers, même si l'ordre canonique des phrases reste sujet-verbe-objet (SVO).

Notons aussi qu'il peut exister des ambiguïtés au niveau des lemmes, mais surtout au niveau des formes fléchies. Entre la multitude de genres, de cas et de différentes formes verbales combinés avec l'accent tonique présent à l'oral (mais pas à l'écrit), il est assez commun de trouver des formes fléchies qui peuvent correspondre à plusieurs lemmes de différentes parties de discours.

Ces particularités, communes à la plupart des langues slaves, peuvent entraîner des difficultés pour les méthodes classiques de TAL, et en premier lieu à l'étiquetage morpho-syntaxique. Elles peuvent également cependant faciliter l'analyse syntaxique car les informations flexionnelles fournissent des indices très utiles à cette tâche (Collins *et al.*, 1999).

Du point de vue du Traitement Automatique des Langues, il s'agit d'une langue peu dotée, et peu de travaux peuvent être mentionnés à cet égard :
– depuis 2010, l'ukrainien est intégré dans le jeu d'étiquettes morpho-syntaxiques Multex-East [3] (Erjavec, 2012) ;
– il existe un étiqueteur morpho-syntaxique UGtag (Kotsyba *et al.*, 2009) fonctionnant à base de règles et de dictionnaires, mais qui n'effectue pas la désambiguïsation syntaxique et morphologique des mots ;

3. http://nl.ijs.si/ME/V4/

– une méthode de reconnaissance des entités nommées a été proposée (Katrenko & Adriaans, 2007) ;
– il existe également un travail sur la détection et l'analyse de sentiments (Romanyshyn, 2013).
Actuellement, il ne semble pas exister des corpus ou des lexiques librement disponibles pour l'ukrainien. Notre objectif est de contribuer à la description de cette langue et à l'évolution de ressources nécessaires pour les travaux de recherche en TAL et dans d'autres disciplines.

3 Données linguistiques

Nous utilisons deux types de données : (1) les corpus (section 3.1) nous permettent d'effectuer l'acquisition de ressources morphologiques ; (2) un ensemble de mots vides (section 3.2) pour ne pas prendre en compte les mots grammaticaux.

3.1 Corpus

Les corpus proviennent de trois sources, représentant trois genres différents : un corpus littéraire, un corpus de spécialité (textes médicaux) et un corpus encyclopédique provenant de Wikipédia :
– les oeuvres inclus dans Kobzar de Taras Shevtchenko, qui est un des fondateurs de la langue littéraire ukrainienne ;
– les articles et brochures médicales provenant de MedlinePlus (Miller *et al.*, 2000), dont une partie est traduite en ukrainien (à côté d'autres langues) ;
– les articles de Wikipédia en ukrainien. Actuellement, Wikipédia en ukrainien fournit 1 201 585 articles.
Dans le tableau 1 nous indiquons les tailles de ces corpus. Wikipédia est, bien sûr, le plus grand des corpus, alors que MedlinePlus est le plus petit.

Corpus	Taille (nombre d'occurrence des mots.)
Kobzar	89 289
MedlinePlus	46 230
Wikipédia	246 368 411

Table 1 – Taille des corpus

3.2 Mots vides

Une liste de mots vides comporte 385 formes, issue d'une ressource existante destinée à l'internationalisation d'interfaces graphiques [4]. En (3), nous présentons quelques mots vides de cette liste. Cependant, nous avons pu observé qu'il sera nécessaire d'augmenter cette ressource par une analyse en corpus.

(3) зі (*avec*), ми (*nous*), на (*sur*), та (*et*), ти (*tu*), ще (*encore*), що (*que*), їй (*à elle*), їм (*à eux*)

4 Approche pour la constitution de ressources morphologiques

La méthode générale se décompose en quatre étapes : la préparation de corpus (section 4.1), l'extraction de paires de mots liés sémantiquement et morphologiquement (section 4.2), et leur évaluation (section 4.3).

4. https://github.com/fluxbb/langs/blob/master/Ukrainian/stopwords.txt

4.1 Préparation de corpus

Trois étapes sont effectuées :
– les corpus sont convertis en UTF-8 ;
– la segmentation en mots est effectuée. Elle doit prendre en compte la valeur spécifique de l'apostrophe lorsqu'il apparaît à l'intérieur des mots et auquel cas il ne peut pas être un caractère permettant la segmentation. En revanche, lorsqu'il apparaît aux extrémités des mots il a sa valeur habituelle de guillemets ;
– les mots vides sont supprimés afin d'alléger les traitements ultérieurs.

4.2 Extraction de paires de mots liés morphologiquement

L'objectif de la méthode est de détecter les mots liés sémantiquement et morphologiquement en corpus. Nous exploitons pour ceci la notion de continuité thématique du discours. Il existe en effet des liens thématiques lexicaux au sein des textes. D'une part, les locuteurs ont tendance à employer les mots d'un champs sémantique donné (par exemple, *hôpital*, *médecin*, *opérer*). D'autre part, il peuvent également employer des mots d'une même famille morphologique (*opérer*, *opération*). Dans cette situation, il est possible de trouver des mots d'une même famille morphologique dans les séquences de corpus.

Comme dans le travail précédent (Zweigenbaum *et al.*, 2003), la notion de continuité thématique est approximée à l'aide d'une fenêtre glissante de M mots. La proximité morphologique entre deux mots est indiquée par les n premiers caractères du mot. En résumé, nous recensons les mots qui partagent la même chaîne de caractères initiale de longueur supérieure ou égale à c et qui se trouvent souvent dans une même fenêtre de M mots. Ce dernier critère sera mis en œuvre par une mesure statistique d'association qui évalue dans quelle mesure cette cooccurrence est plus fréquente que ce que donnerait le hasard. Nous exploitons le rapport de vraisemblance ("likelihood ratio") (Manning & Schütze, 1999) : rapport $\lambda = \frac{L(H_1)}{L(H_2)}$ entre la probabilité d'observer le nombre de cooccurrences du mot m_2 avec le mot m_1 dans l'hypothèse H_1 où les mots sont indépendants et la probabilité d'observer leur nombre de cooccurrences dans l'hypothèse H_2 où les mots sont dépendants (on calcule $-2\log\lambda$).

Les données pour calculer ce rapport sont les suivantes. Probabilité de l'observation selon H_1 (indépendance) : $L(H_1) = b(c_{12}, c_1, p)b(c_2 - c_1, N - c_1, p)$; probabilité de l'observation selon H_2 (dépendance) : $L(H_2) = b(c_{12}, c_1, p_1)b(c_2 - c_1, N - c_1, p_2)$; loi binômiale (probabilité d'une séquence de k succès parmi n tirages) : $b(k, n, p) = C_k^n p^k (1-p)^{n-k}$; probabilités élémentaires : $p = \frac{c_{12}}{N}$; $p_1 = \frac{c_{12}}{c_1}$; $p_2 = \frac{c_2 - c_{12}}{N - c_1}$; c_1 est le nombre d'occurrences du mot m_1, c_2 est le nombre de fenêtres où apparaît le mot m_2, c_{12} est le nombre de fenêtres où cooccurrent les mot m_1 et m_2, N est la taille du corpus.

Cette mesure d'association est asymétrique car elle dépend différemment de la fréquence propre de chaque mot. Par exemple, on a plus de chances d'observer un nom comme *canal* dans le voisinage de son adjectif *canalaire* que l'inverse. Le score d'association le plus fort des deux directions est conservé. Ce critère d'association est utilisé pour classer les paires de mots : lorsque cette mesure est plus élevée la probabilité que les mots de la paires soient liés morphologiquement est plus élevée. Cependant, nous traitons et évaluons toutes les paires proposées car même avec une mesure faible il est possible de trouver des mots liés morphologiquement.

Cette approche est appliquée sur les trois corpus. La fenêtre exploitée est 10 mots à gauche et à droite par rapport au mot pivot ($M = 21$). Nous utilisons la longueur de la chaîne initiale commune de 3 caractères ($c = 3$) car cela permet de garder les paires dont les mots partagent probablement des bases communes. Nous nous attendons à ce que ce paramétrage permette d'obtenir une bonne précision.

4.3 Évaluation

Comme il n'existe pas de ressources de référence, l'évaluation est effectuée manuellement par un locuteur de la langue. Les paires ont été présentées de manière ordonnée en fonction du rapport de vraissemblance. Cette évaluation donne une idée de la précision des résultats. La question est posée lors de l'évaluation est la suivante : *Est-ce que cette paire de mots serait utile en recherche d'information pour l'extension de*

la requête ? En d'autres mots, est-ce qu'en recherche d'information, une paire de mots données permettra d'augmenter le rappel sans trop détériorer la précision. Il s'agit donc d'un cadre d'évaluation assez ciblé, mais pour lequel il est important de veiller à la précision des ressources.

5 Résultats

Les corpus sont traités et permettent d'extraire un nombre assez important de paires de mots supposés être reliés morphologiquement. Dans le tableau 2, première colonne, nous indiquons le nombre de paires de mots validées. Notons qu'à partir du corpus Wikipédia, nous avons extrait 3 108 591 paires de mots mais nous n'avons évalué qu'un échantillon de 6 950 paires pour le moment. Dans la dernière colonne du tableau, nous indiquons la précision observée. La précision élevée obtenue sur le corpus Wikipedia peut s'expliquer par le fait que, pour l'instant, nous n'avons validé que les paires qui montrent une force d'association la plus élevée. Il est possible que la précision globale de ces paires de mots diminuera avec l'augmentation de l'ensemble évalué. Pour les deux autres corpus, nous observons une précision moins bonne : 67 % pour le corpus médical et 76 % pour le corpus littéraire. Aussi, la précision moins importante obtenue sur le corpus médical peut être due à la taille de ce corpus, notre méthode étant sujette au volume de données traitées. Cependant, ces résultats sont assez comparables avec ceux obtenus dans un travail précédent sur le français médical (Zweigenbaum *et al.*, 2003). Dans ce travail, nous avons pris la longueur de la chaîne initiale de quatre caractères et avons obtenu une précision moyenne de 75,6 % au 5000è rang (fenêtre de 150 mots). Dans le travail actuel, nous avons diminué la chaîne initiale à trois caractères car les bases dans la langue ukrainienne sont souvent plus courtes qu'en français médical. Pour cette même raison, le risque d'extraire de fausses propositions augmente. Par contre, comme nous effectuons la recherche de mots liés morphologiquement dans une fenêtre de 21 mots, cela réduit ce risque. Si nous nous positionnons au rang de 5 000 paires de mots (comme dans le travail précédent), la précision observée reste de 86 %. En moyenne entre les trois corpus exploités dans notre travail, nous obtenons donc une précision comparable à celle observée dans le travail précédent, réalisé sur le français.

Corpus	Nombre de paires	Précision
Kobzar	2 603	76
MedlinePlus	1 961	67
Wikipédia (échantillon validé)	6 950	86

Table 2 – Nombre de paires de mots et précision

L'ensemble validé fournit actuellement 3 315 paires de mots jugées comme correctes. Cette ressource sera mise à disposition de la communauté scientifique.

Le travail présenté dans cet article montre aussi que cette méthode peut être transposée sur différentes langues, pour lesquelles des corpus sont disponibles, afin d'amorcer l'acquisition de ressources morphologiques. Les ressources acquises de cette manière peuvent ensuite servir pour déduire les règles les plus fréquentes, de même que des règles moins fréquentes, pour une langue et de compléter ainsi ces premières ressources (Grabar & Zweigenbaum, 1999) grâce à l'exploitation de l'analogie qui existe dans la formation de mots.

À la figure 1, nous présentons le recouvrement entre les ressources acquises à partir des trois corpus. Il s'agit de l'ensemble de paires extraites : pour le corpus Wikipédia seulement une partie de ces paires est évaluée actuellement. Nous pouvons voir qu'il existe peu de recouvrement entre les corpus. Wikipédia fournit une énorme part de paires de mots (évaluées et non évoluées). C'est aussi Wikipédia qui a des sous-ensembles communs avec les deux autres corpus. Seulement huit paires de mots sont partagées par les 3 corpus. Elles sont présentées en (4). Il s'agit de formes fléchies de mots dont la traduction est indiquée entre parenthèses. Notons qu'il n'existe pas de paires de mots communes seulement entre le corpus médical et littéraire. Ces observations indiquent qu'il est nécessaire de traiter plusieurs corpus provenant de différents genres et domaines pour avoir une couverture acceptable pour une nouvelle langue. La situation est bien sûr plus difficile pour une langue à morphologie riche comme l'ukrainien.

(4) {руки; руку} (*main*), {серця; серцем} (*coeur*), {кров; крові} (*sang*), {ліжка; ліжку} (*lit*), {новими; нові} (*nouveau*), {одна; одну} (*seule*), {стало; стали} (*devenir*), {кров; кров'ю} (*sang*)

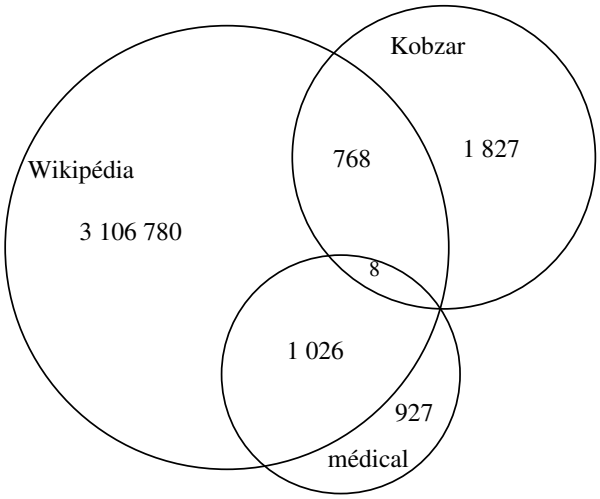


Figure 1 – Recouvrement de ressources acquises sur différents corpus.

Parmi les paires de mots extraites, certaines comportaient des mots ambigus qui, selon les contextes, peuvent correspondre à des lemmes différents. En voici quelques exemples :
– {поділися; поділось} : cette paire contient les formes du verbe *disparaître*, cependant le mot поділися, avec un accent tonique différent, correspond au lemme du verbe *partager* ;
– dans la paire {дітись; діти}, le sens principal est *se mettre quelque part* mais le mot діти correspond aussi à une forme flexionnelle de *enfant* ;
– dans la paire {гори; горить}, le sens principal est *brûler*, tandis que гори correspond aussi à une forme flexionnelle de *montagne*.
Pour de telles situations, nous avons considéré qu'il s'agit d'extractions correctes car elles peuvent apporter des résultats supplémentaires et corrects dans un contexte de recherche d'information. Pour ce qui est de l'ambiguïté contextuelle, elle devrait être traitée avec des méthodes spécifiques.

Parmi les paires de mots correctes, nous avons un grand nombre de formes flexionnelles, même si les lemmes y apparaissent rarement. Nous trouvons aussi des paires avec les dérivations (exemples en (5)) et compositions (exemples en (6)).

(5) {алергійна; алергія} ({*allergique*; *allergie*}), {братерськая; брате} ({*fraternelle*; *frère*}), {вакцинацію; вакцина} ({*vaccination*; *vaccin*}), {дитину; дитячий} ({*enfant*; *enfantin*})

(6) {ангіопластика; ангіограми} ({*angioplasie*; *angiogramme*}), {бронхіоли; бронхіт} ({*bronchiole*; *bronchite*}), {газованих; газоутворення} ({*gazéeux*; *production de gaz*})

En comparaison avec la langue française, dans les ressources extraites sur les corpus en ukrainien, nous avons deux nouveaux cas de figures : les formes diminutives comme dans les exemples en (7), où ангеляточко (*petit ange*) est formé sur ангел (*ange*) ; et les patronymes, comme les exemples en (8).

(7) {ангеляточко; ангел} (*ange*), {біленькі; білих} (*blanc*), {Богданочку; Богдане} (*Bohdan, nom propre*), {воленьки; волі} (*liberté*), {годину; годиночку} (*heure*)

(8) {Іван; Іванович} ({*Jean*; *fils de Jean*}), {Микола; Миколайович} ({*Nicolas*; *fils de Nicolas*})

Quant aux erreurs, elles sont assez typiques de ce type de méthode. Nous avons essentiellement détecté deux

types d'erreurs que nous avons également observées sur les données en français et en anglais (Zweigenbaum *et al.*, 2003; Grabar & Zweigenbaum, 1999) :
– les mots qui ont les mêmes chaînes initiales sans pourtant avoir un lien sémantique ou morphologique entre eux, comme dans les exemples en (9) ;
– les mots qui comportent les mêmes préfixes sans que le reste des mots soient lié sémantiquement ou morphologiquement, comme dans les exemples en (10).

Les préfixes apportent du bruit, comme dans les exemples en (10), mais aussi du silence car ils empêchent de faire lien entre les mots liés morphologiquement. Par exemple, les mots des séries présentées en (1) et (2) ne peuvent pas être mis en relation avec la méthode actuelle. Il s'agit d'un aspect de la méthode qui doit être amélioré. Dans les travaux futurs, nous prévoyons d'utiliser les préfixes communs de la langue ukrainienne, comme par exemple ceux fournis par un dictionnaire existant (Клименко *et al.*, 1998). Nous espérons ainsi dépasser cette limite.

(9) {криза; криму} ({*crise*; *Crimée*}), {проблем; прокурорської} ({*problème*; *procureur (adj)*})

(10) {заплануйте; запізнюйтесь} ({*planifier*; *être en retard*}), {відповідає; відстань} ({*répondre*; *laisser tranquille*}), {переставляйте; перевірте} ({*déplacer*; *vérifier*})

Une autre limite de la méthode concerne son incapacité à traiter les allomorphies qui apparaissent dans les trois premiers caractères (la contrainte du paramétrage utilisé dans le travail présenté ici), comme dans les exemples en (11).

(11) {хід; хода} (*marche*), {воля; вільний} ({*liberté*; *libre*})

6 Conclusion et travaux futurs

Nous avons proposé un travail sur l'acquisition de ressources morphologiques pour la langue ukrainienne. Nous exploitons pour ceci une méthode non supervisée qui ne requiert pas d'annotations ni de ressources spécifiques. Celle-ci est seulement basée sur l'utilisation de corpus bruts. Ces deux aspects correspondent à son originalité et ses avantages. Les mesures d'association statistique entre les mots permettent d'apprécier la probabilité du lien sémantique et morphologique qui existe entre ces mots. La méthode est appliquée à trois corpus qui représentent les genres différents de la langue : littéraire, médical et la langue générale. L'ensemble de paires de mots jugées comme correctes est actuellement de 3 315. Cet ensemble sera progressivement complété avec les données extraites à partir de Wikipédia et qu'il reste à traiter. La ressource validée sera mise à disposition de la communauté scientifique.

La méthode permet d'acquérir plusieurs paires de mots, avec une précision variant entre 67 % et 86 % selon le corpus. Ces résultats sont comparables avec les expériences menées sur la langue médicale en français.

L'expérience présentée ici montre que cette méthode peut être appliquée aux corpus en différentes langues afin d'acquérir les ressources morphologiques. Nous serons ainsi intéressés de tester cette méthodes sur d'autres langues et corpus.

Nous avons noté deux limites de la méthode : la préfixation, que nous proposons de traiter grâce à l'utilisation d'un ensemble de préfixes connus de la langue (Клименко *et al.*, 1998) et qui sont en nombre fini, et l'allomorphie qui apparaît au début des mots. Ce dernier point sera plus difficile à résoudre. Un autre point délicat concerne l'évaluation des paires de mots extraites. Il s'agit en effet d'une tâche très longue et lourde. Nous prévoyons d'exploiter d'autres indicateurs en plus des mesures d'association calculées sur le corpus lors de l'extraction des données. D'autres mesures statistiques (Hamon *et al.*, 2012; Loukachevitch & Nokel, 2013) de même que l'exploitation de la théorie des graphes (Diestel, 2005) vont nous permettre d'alléger cette étape de la méthode.

Les ressources acquises avec cette méthode peuvent être utilisées pour l'induction de règles morphologiques et servir ensuite à enrichir cette ressource. Nous prévoyons aussi d'utiliser cette ressource et les ressources

dérivées pour l'étiquetage morpho-syntaxique et la recherche d'information en ukrainien.

Références

Abeera V., Aparna S., Rekha R., Kumar M., Dhanalakshmi V., Soman K. & Rajendran S. (2012). Morphological analyzer for Malayalam using machine learning. *Data Engineering and Management, LNCS*, **6411**, 252–254.

Bosch S., Pretorius L. & Fleisch A. (2008). Experimental bootstrapping of morphological analysers for Nguni languages. *Nordic Journal of African Studies*, **17**(2), 66–88.

Burnage G. (1990). *CELEX - A Guide for Users*. University of Nijmegen : Centre for Lexical Information.

Claveau V. & Kijak E. (2014). Generating and using probabilistic morphological resources for the biomedical domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 3348–3354.

Collins M., Hajic J., Ramshaw L. & Tillmann C. (1999). A statistical parser for czech. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, p. 505–512, College Park, Maryland, USA : Association for Computational Linguistics.

Déjean H. (1998). Morphemes as necessary concept for structures discovery from untagged corpora. In *Workshop on Paradigms and Grounding in Natural Language Learning*, p. 295–299, Adelaide.

Diestel R. (2005). *Graph Theory*. New-York : Springer-Verlag Heidelberg.

Erjavec T. (2012). Multext-east : Morphosyntactic resources for central and eastern european languages. *Language Resources and Evaluation*, **46**(1), 131–142.

Gaussier E. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In A. Kehler & A. Stolcke, Eds., *ACL workshop on Unsupervised Methods in Natural Language Learning*, College Park, Md.

Grabar N. & Zweigenbaum P. (1999). Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. In *Traitement Automatique de Langues Naturelles (TALN)*, p. 175–184.

Hamon T., Engström C., Manser M., Badji Z., Grabar N. & Silvestrov S. (2012). Combining compositionality and pagerank for the identification of semantic relations between biomedical words. In *BIONLP NAACL*, p. 109–117.

Hathout N. (2001). Analogies morpho-syntaxiques. In *Traitement Automatique des Langues Naturelles (TALN)*, Tours.

Hathout N. & Namer F. (2014). La base lexicale Démonette : entre sémantique constructionnelle et morphologie dérivationnelle. In *TALN*, p. 208–219.

Katrenko S. & Adriaans P. (2007). Named entity recognition for ukrainian : A resource-light approach. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, p. 88–93, Prague, Czech Republic : Association for Computational Linguistics.

Kostov J. (2013). *Le verbe macédonien : pour un traitement informatique de nature linguistique et applications didactiques (réalisation d'un conjugueur)*. Thèse de doctorat, INaLCO, Paris, France.

Kotsyba N., Mykulyak A. & Shevchenko I. V. (2009). Ugtag : morphological analyzer and tagger for the ukrainian language. In *Proceedings of the international conference Practical Applications in Language and Computers (PALC 2009)*.

Krovetz R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, p. 191–202.

Loginova-Clouet E. (2014). *Traitement automatique des termes composés : segmentation, traduction et variation*. Thèse de doctorat, Université de Nantes, Nantes, France.

Loukachevitch N. & Nokel M. (2013). An experimental study of term extraction for real information-retrieval thesauri. In *TIA*, p. 1–8.

Manning C. D. & Schütze H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA : MIT Press.

McCray A. T., Srinivasan S. & Browne A. C. (1994). Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual SCAMC*, p. 235–239.

Miller N., Lacroix E. & Backus J. (2000). MEDLINEplus : building and maintaining the national library of medicine's consumer health web service. *Bull Med Libr Assoc*, **88**(1), 11–7.

Namer F. (2009). *Morphologie, Lexique et TAL : l'analyseur DériF. TIC et Sciences cognitives*. London : Hermes Sciences Publishing.

Pentheroudakis J. & Vanderwende L. (1993). Automatically identifying morphological relations in machine-readable dictionaries. In *Ninth annual confenrence of the UW Center for the New OED an d Text Research*, p. 114–131.

Pirrelli V. & Yvon F. (1999). The hidden dimension : a paradigmatic view of data-driven NLP. *JETAI*, **11**, 391–408.

Pretorius L. & Bosch S. (2009). Exploiting cross-linguistic similarities in Zulu and Xhosa computational morphology. In *AFLAT*, p. 96–103.

Romanyshyn M. (2013). Rule-based sentiment analysis of ukrainian reviews. *International Journal of Artificial Intelligence & Applications (IJAIA)*, **4**(4), 103–111.

Sagot B., Clément L., Villemonte de la Clergerie E. & Boullier P. (2006). The Lefff 2 syntactic lexicon for french : architecture, acquisition, use. In *Proceedings of LREC*.

Schone P. & Jurafsky D. (2001). Knowledge-free induction of inflectional morphologies. In *Proceedings of NAACL'01*, p. 1–9.

Theron P. & Cloete I. (1997). Automatic acquisition of two-level morphological rules. In *ANLP*, p. 103–110.

Urrea A. M. (2000). Automatic discovery of affixes by means of a corpus : a catalog of Spanish affixes. *Journal of quantitative linguistics*, **7**(2), 97–114.

van den Bosch A., Daelemans W. & Weijters T. (1996). Morphological analysis as classification : an inductive-learning approach. In *International Conference on Computational Linguistics (COLING)*.

Xu J. & Croft B. W. (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems*, **16**(1), 61–81.

Zanchetta E. & Baroni M. (2005). Morph-it ! a free corpus-based morphological resource for the italian language. *Corpus Linguistics 2005*, **1**(1).

Zweigenbaum P., Hadouche F. & Grabar N. (2003). Apprentissage de relations morphologiques en corpus. In *Traitement Automatique des Langues Naturelles (TALN)*, p. 285–294.

Клименко Н. Ф., Карпіловська Є. А., Карпіловський В. С. & Недозим Т. І. (1998). *Словник Афіксальних Морфем Української Мови*. Київ, Україна : Інститут Мовознавства ім. О.О. Потебні Національної Академії Наук України.

Modèle de document pour TALN 2015
Représentation des expressions composées en macédonien en tant qu'entrées lexicales en Unitex

Aneta Rafajlovska[1], Katerina Zdravkova[2]
Université Sts Cyrille et Méthode, Skopje
Faculté de science informatique et de génie informatique
[1]r.aneta@yahoo.com, [2]katerina.zdravkova@finki.ukim.mk

Résumé. Le logiciel de traitement de corpus Unitex 3.0 a été utilisé pour obtenir la flexion automatique de mots simples et des mots composés en langue macédonienne. En utilisant les graphes de flexion des mots simples, nous avons réussi à représenter les expressions composées du corpus en tant qu'entrées lexicales dans un dictionnaire DELAC en Unitex. En outre, nous avons créé des transducteurs à états-finis qui permettent de fléchir les expressions composées et nous avons obtenu automatiquement toutes leurs formes fléchies que nous avons stockées dans un dictionnaire DELACF (DELA de formes Composées Fléchies).

Abstract.

Representation of Multiword Expressions in Macedonian as Lexical Entries in Unitex
The corpus processing system – Unitex 3.0 was used to obtain the automatic inflection of the simple word forms and the multiword expressions in Macedonian. Based on the inflection graphs of the simple word forms we managed to represent the multiword expressions retrieved from the corpus as lexical entries in a DELAC dictionary in Unitex. We also created inflection finite-state transducers for the multiword expressions and as a result we managed to obtain automatically all the inflected forms of the multiword expressions in the form of a DELACF dictionary of compound inflected forms.

Mots-clés : expressions composées, mots composés, mots simples, flexion automatique, transducteurs à états-finis de flexion, Unitex, Multiflex

Keywords: multiword expressions, compound words, simple word forms, automatic inflection, inflection finite-state transducers, Unitex, Multiflex

1 Expressions composées

Les expressions composées ou les mots composés représentent un problème linguistique assez important, surtout à cause de la difficulté de les définir et de les représenter. Par conséquent, les mots composés sont énumérés parmi les problèmes majeurs en traitement automatique des langues (TAL), générant des ambigüités conséquentes (Sag, Baldwin, Bond, Copestake, Flickinger, 2002). Les études récentes dans le domaine du TAL ont incité les linguistes à abordes de nouvelles théories linguistiques et à développer différentes approches par rapport à la syntaxe et la lexicologie dès les années 1960 (Léon, 2004).

Il existe de nombreuses définitions linguistiques et pragmatiques pour les expressions composées. Toutefois, il est généralement admis que les expressions composées contiennent au moins deux ou plusieurs mots qui représentent un seul lexème. Dans ce sens un lexème signifie une seule unité lexicale. De ce fait, les expressions composées contiennent deux ou plusieurs mots, mais ils représentent un seul ensemble qui peut différer du sens premier de ces mots pris séparément. Elles posent également un problème de représentation, car l'unité représntative dans un lexique linéaire est le mot, ce qui les exclut dans les dictionnaires (Gross, 1986).

D'un point de vue de la nature, les expressions composées en macédonien peuvent être des adverbes composés, des noms composés ou des verbes composés, comme dans les exemples suivants : « под услов да » (*pod uslov da*, ADV) - « sous condition que/de », « во врска со » (*vo vrska so*, ADV) – « à propos de / concernant », « во недостиг на » (*vo nedostig na*, ADV) – « à défaut de / en absence de », « високо друштво » (*visoko društvo*, NOM) – « haute société », « здрав разум » (*zdrav razum*, NOM) – « bon sens / lucidité », « роден крај » (roden kraj, NOM) – « pays natal », « доби премија » (*dobi premija*, V) – « gagner le gros lot », « има предвид » (*ima predvid*, V) « prendre en

considération / tenir compte de », « ги зема работите во свои раце » (*gi zema rabotite vo svoi race*, V) – « prendre le contrôle de » etc.

2 Méthodologie

Notre objectif premier est de représenter les expressions composées en tant qu'entrées lexicales dans un dictionnaire morphologique pour pouvoir obtenir automatiquement leur flexion. Notre travail consiste en quatre étapes essentielles :

- Extraction et annotation des expressions composées macédoniennes du corpus
- Création du dictionnaire des mots simples
- Flexion automatique des mots simples
- Création du dictionnaire des mots composés
- Flexion automatique des mots composés

La création de cette ressource Unitex a pris un peu plus de trois mois. Le dictionnaire des mots simples (Figure 1) contient 280 entrées lexicales, ce qui permet d'obtenir automatiquement le dictionnaire des formes fléchies avec, environ 3 762 entrées (Figure 4). Le dictionnaire des mots composés contient 184 mots composés (Figure 5). Il permet d'obtenir automatiquement le dictionnaire des formes fléchies qui contient 1 454 entrées (Figure 8).

3 Particularités morphologiques du macédonien

Les catégories grammaticales des noms et des adjectifs en macédonien sont le genre, le nombre et la définitude qui s'exprime par un suffixe appelé « article défini ». Les étiquettes de ces catégories sont les suivantes :

1. Genre
- Masculin m
- Féminin f
- Neutre n

2. Nombre
- Singulier s
- Pluriel simple p
- Pluriel compté i
- Pluriel collectif z

3. Article définie
- Indéterminé U
- Proximal (objets qui se trouvent à côté de celui qui parle) P
- Distal (objets qui se trouvent loin de celui qui parle) D

4 Dictionnaire et flexion automatique des mots simples

Pour pouvoir aborder le problème linguistique en question et pour obtenir la flexion automatique des mots simples de même que des mots composés, nous avons utilisé Unitex 3.0. Les dictionnaires électroniques distribués en Unitex utilisent la structure DELA (Dictionnaires Electroniques du LADL) (Paumier, 2006). Nous avons créé un dictionnaire macédonien des formes lexicale des mots simples au format DELA en Unitex (Figure 1). Le dictionnaire contient les formes lexicales des mots suivies par une virgule et le nom du graphe de flexion qui sera utilisé pour obtenir toutes les formes fléchies automatiquement.

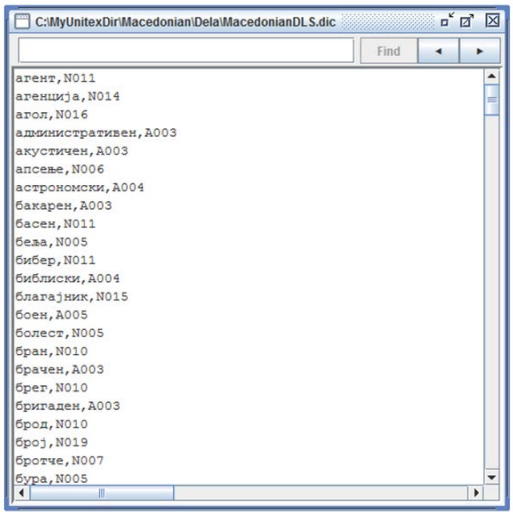


Figure 1 : Extrait du dictionnaire des mots simples dans le format DELA

4.1 Graphes de flexion des мots simples

Les graphes en Unitex compilés sous le format *.fst2* ont été utilisés en vue d'obtenir toutes les formes fléchies des lemmes des mots simples. Pour couvrir toutes les formes lexicales du dictionnaire des mots simples, nous avons créé 19 graphes de flexion (transducteurs à états-finis) pour les noms et 4 graphes pour les adjectifs.

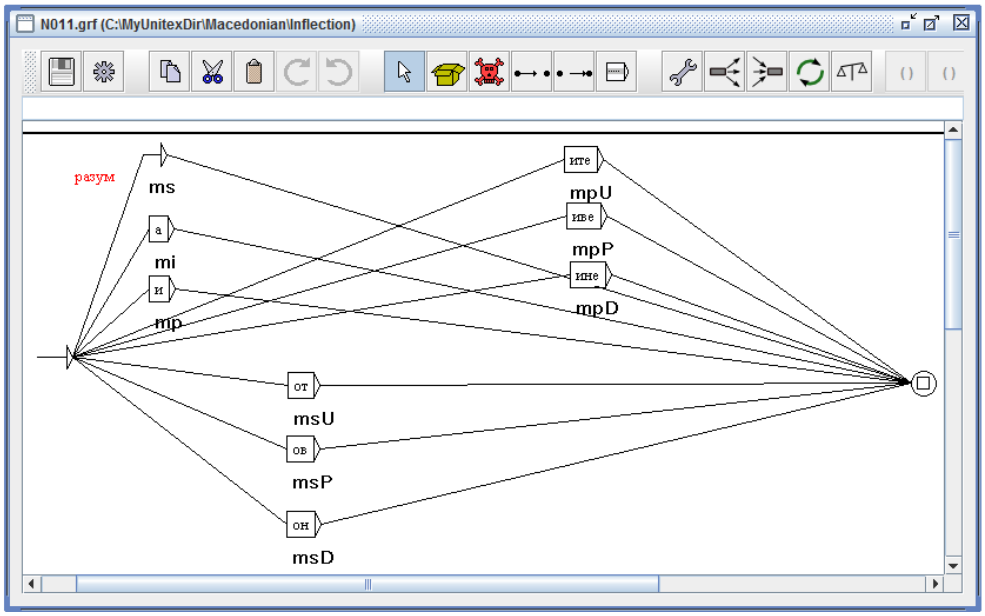


Figure 2 : Un graphe de flexion pour les noms masculins

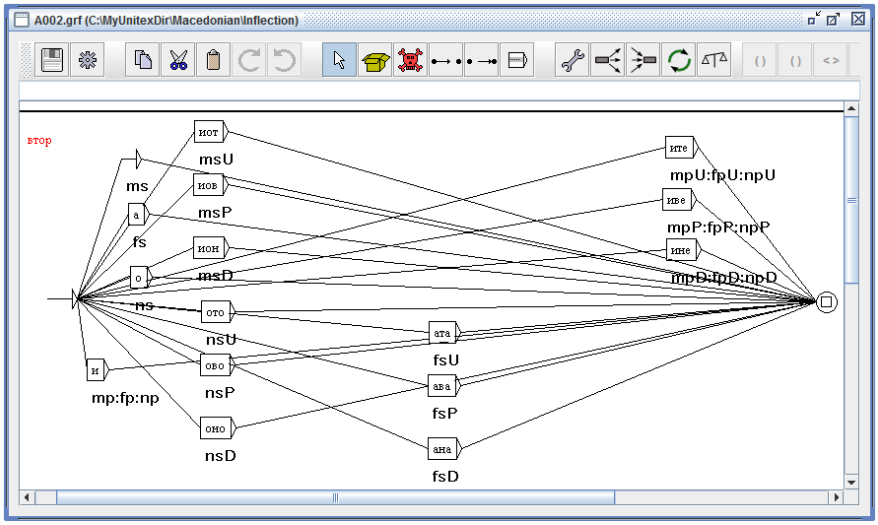


Figure 3 : Un graphe de flexion pour les adjectifs

Dans les Figures 2 et 3, nous pouvons observer les différents suffixes ajoutés à la forme lexicale du mot avec les codes flexionnels pour la catégorie grammaticale. Pour le nom 'разум' (razum) 'sens' et tous les autres noms qui utilisent le graphe N011 pour la flexion automatique, les suffixes –a et –i seront ajoutés à la forme lexicale pour former les deux formes du pluriel, ce qui donne dans le cas de 'razum' - 'razuma' et 'razumi'. Pour l'article défini indéterminé le suffixe –ot sera ajouté ce qui donne 'razumot', le suffixe –ov pour les objets proches 'razumov' ce qui peut être traduit par 'celui-ci', et le suffixe –on 'razumon' pour les objets lointains ce qui peut être traduit par 'celui-là'. Au pluriel les articles définis seront exprimés par les suffixes -ite 'razumite' (indéterminé), -ive 'razumive' (ceux-ci) et –ine 'razumine' (ceux-là). En revanche, les adjectifs varient aussi selon le genre (Figure 3), la forme lexicale est au masculin singulier 'vtor'. Le suffixe –a est ajouté pour créer la forme féminine 'vtora', et le suffixe –o est ajouté pour créer la forme neutre 'vtoro'. La forme plurielle est la même pour tous les trois genres formée par le suffixe –i 'vtori'. Tous les trois genres prennent les trois formes de l'article défini, de même que la forme plurielle, pour le masculin 'vtoriot', 'vtoriov' et 'vtorion' ; pour le féminin 'vtorata', 'vtorava' et 'vtorana' ; pour le neutre 'vtoroto', 'vtorovo' et 'vtorono' et pour le pluriel 'vtorite', 'vtorive' et 'vtorine'.

4.2 Dictionnaire macédonien de formes fléchies des mots simples

Après avoir construit le Dictionnaire macédonien de lemmes de mots simples (Figure 1) nous avons appliqué les graphes de flexion et nous avons obtenu automatiquement le Dictionnaire des formes fléchies (Figure 4).

Le Dictionnaire des formes fléchies (Figure 4) comprend la forme fléchie suivie d'une virgule et de la forme lexicale du mot, puis d' un point et du code flexionnel décrivant catégorie grammaticale.

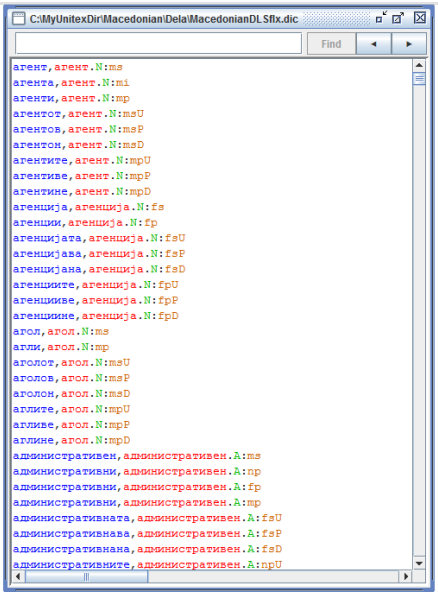


Figure 4 : Extrait du dictionnaire de formes fléchies des mots simples

5 Dictionnaire et flexion automatique des mots composés

L'objectif principal étant la flexion automatique des mots composés, nous avons pris la liste des mots composés du corpus (la traduction macédonienne du roman *Tour du monde en quatre-vingts jours* de Jules Verne) et nous avons créé un dictionnaire DELAC (DELA de formes composées) sous Unitex (Figure 5). Unitex utilise le formalisme Multiflex (Savary, 2008), qui représente une approche graphique pour représenter la flexion des expressions composées. Les graphes de flexion des expressions composées réutilisent les graphes de flexion de ces composants - les mots simples.

Le dictionnaire contient la forme fléchie du premier constituant de l'expression composée, suivie de la forme lexicale entre parenthèses, puis du graphe de flexion du mot, du deuxième constituant et de sa forme lexicale, puis du graphe de flexion du deuxième mot, d'une virgule et du nom du graphe de flexion de l'expression composée.

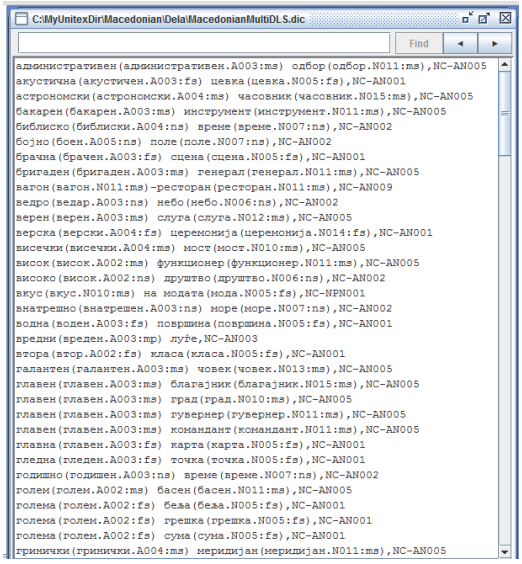


Figure 5 : Extrait du dictionnaire des mots composés

5.1 Graphes de flexion de mots composés

Le graphe de flexion de mots composés (Figure 6) est utilisé pour la flexion de mots composés du type Adjectif-Nom (AN). Les composants du mot composé sont définis par le caractère $1 qui est le premier composant, l'adjectif 'млади' (mladi) 'jeune'; le caractère $2 représente l'espace, et le caractère $3 représente le nom 'години' (godini) 'âge'. Le nom, soit le troisième élément est un féminin pluriel, donc l'adjectif ou le premier élément ne peut être qu'au pluriel Nb=p et au féminin, ce qui est déterminée par Gen=f. En outre, l'adjectif peut prendre le suffixe des trois articles définis qui est déterminé par Def=$d.

Tout d'abord, l'adjectif 'млади' (mladi) 'jeune' est inclus dans le dictionnaire des mots simples et un graphe de flexion lui est associé. Le même principe s'applique pour le nom 'години' (godini) 'âge'. L'entrée dans le dictionnaire des mots simples pour l'adjectif est le suivant : млад,A002 ; et pour le nom : година,N005. L'entrée dans le dictionnaire des mots contenus dans l'expression composée est la suivante : млади(млад.A002:fp) години,NC-AN006. Nous pouvons observer sur le graphe de la Figure 6 que le troisième élément (le nom) ne change pas et qu'il ne faut pas substituer l'entrée le graphe qui lui est associé. En revanche, l'adjectif peut prendre différentes formes et donc il est obligatoire de citer le graphe de flexion de ce mot. Enfin, après l'application du graphe de flexion de l'expression composée Figure 6, le logiciel produit toutes les formes fléchies de l'expression composée, sans produire les formes qui ne sont pas permises, comme, par exemple, la forme du nom au singulier, ni les formes avec l'article défini pour le nom. Dans le dictionnaire des formes fléchies des expressions composées apparaissent seulement les formes au pluriel :

младите години, млади години.N:pf
младиве години, млади години.N:pf
младине години, млади години.N:pf
млади години, млади години.N:pf

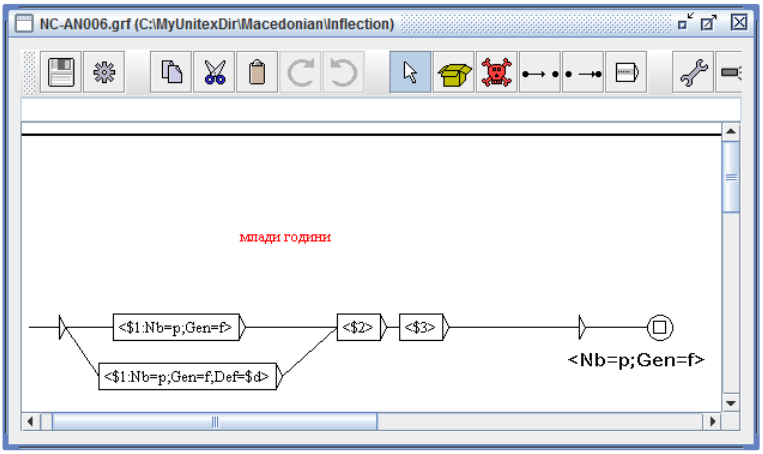


Figure 6 : Un graphe de flexion de mots composés du type AN

Le graphe représenté à la Figure 7 est utilisé pour la flexion des mots composés du type Nom Préposition Nom (NPN), dans le cas où les noms peuvent être fléchis tous les deux. Le caractère $1 représente le premier composant, qui est en fait le nom 'куќичка' (kukjichka) 'coquille', le caractère $2 représente l'espace, le caractère $3 représente la préposition 'на' (na) 'de', le caractère $4 représente l'espace, et le caractère $5 représente le nom 'полжав' (polzhav) 'colimaçon'. Les deux noms peuvent être en n'importe quel nombre Nb=$n et ils peuvent prendre tous les suffixes de l'article défini Def=$d.

Les noms 'куќичка' (kukjichka) 'coquille' et 'полжав' (polzhav) 'colimaçon' sont inclus dans le dictionnaire des mots simples et un graphe de flexion leur est associé. L'entrée dans le dictionnaire des mots simples pour les deux noms est la suivante : куќичка,N005; et полжав,N011. L'entrée dans le dictionnaire des mots composés de l'expression composée est le suivant : куќичка(куќичка.N005:fs) на полжав(полжав.N011:ms),NC-NPN003. Nous pouvons observer que les deux graphes de flexion des noms sont cités, ainsi que le graphe de flexion de l'expression composée Figure 7. Enfin, après l'application du graphe de flexion de l'expression composée le logiciel produit toutes les formes fléchies de l'expression composée. Dans le dictionnaire des formes fléchies des expressions composées apparaissent les formes suivantes:

куќичката на полжав, куќичка на полжав.N:s

куќичката на полжавот, куќичка на полжав.N:s
куќичкава на полжав, куќичка на полжав.N:s
куќичкава на полжавов, куќичка на полжав.N:s
куќичкана на полжав, куќичка на полжав.N:s
куќичкана на полжавон, куќичка на полжав.N:s
куќичките на полжави, куќичка на полжав.N:p
куќичките на полжавите, куќичка на полжав.N:p
куќичките на полжавиве, куќичка на полжав.N:p
куќичкине на полжави, куќичка на полжав.N:p
куќичкине на полжавине, куќичка на полжав.N:p
куќичка на полжав, куќичка на полжав.N:s
куќичка на полжавот, куќичка на полжав.N:s
куќичка на полжавов, куќичка на полжав.N:s
куќичка на полжавон, куќичка на полжав.N:s
куќички на полжави, куќичка на полжав.N:p
куќички на полжавите, куќичка на полжав.N:p
куќички на полжавиве, куќичка на полжав.N:p
куќички на полжавине, куќичка на полжав.N:p

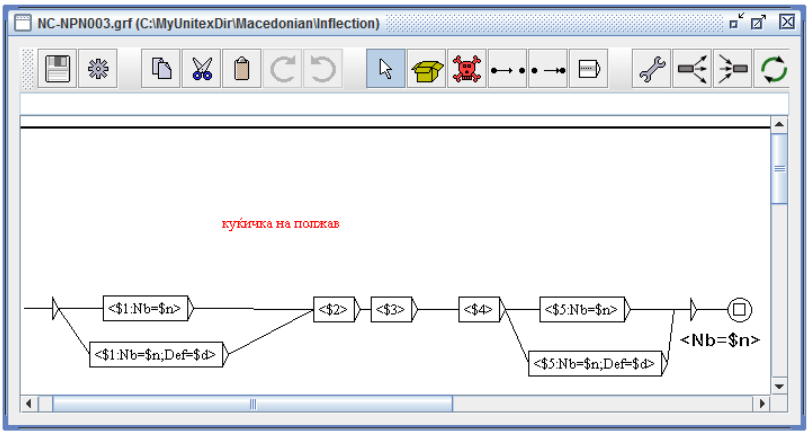


Figure 7 : Un graphe de flexion de mots composés du type NPN

5.2 Dictionnaire macédonien de formes fléchies des mots composés

La Figure 8 montre une partie du dictionnaire DELACF (DELA de formes Composées Fléchies) obtenu automatiquement par l'application des graphes au dictionnaire des mots composés Figure 5. À la gauche, toutes les formes fléchies des expressions composées sont citées, suivies de la forme lexicale, et des codes flexionnels.

Toutes les expressions composées nominales représentées sur Figure 8 ont été extraites du corpus. Leur forme lexicale est représentée sur le côté droit, avec la règle d'annotation utilisée pour obtenir la forme fléchie. Les formes fléchies obtenues en utilisant cette règle sont représentées sur le côté gauche du dictionnaire.

Il est à noter que l'expression composée nominale 'акустична цевка' (akustichna cevka) 'tuyau acoustique', apparaît dans sa forme plurielle 'акустични цевки' (akustichni cevki), 'tuyaux acoustiques', dans le roman de Verne. De même, 'бакарни инструменти' (bakarni instrumenti) 'instruments de cuivre', 'библиски времиња' (bibliski vreminja) 'temps bibliques', etc. y sont au pluriel, mais nous avons utilisé la forme au singulier en tant que forme lexicale citée dans le dictionnaire des mots composés.

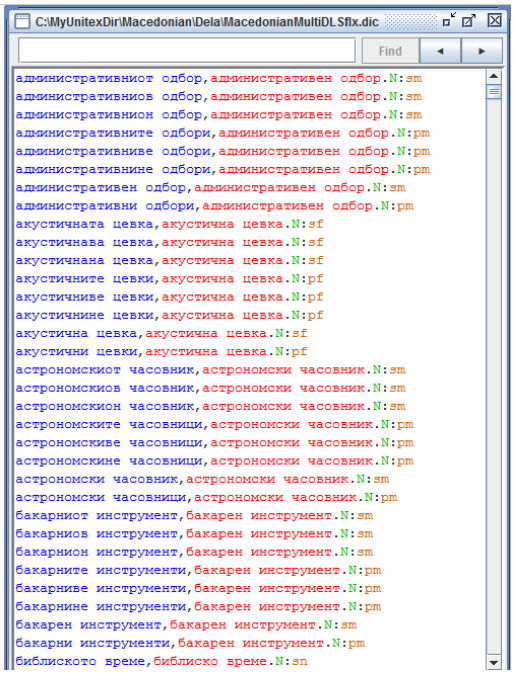


Figure 8 : Extrait du dictionnaire de formes fléchies des mots composés

6 Conclusion

Nous avons utilisé Unitex pour créer un dictionnaire et des graphes de flexion de mots simples, afin de pouvoir utiliser ces ressources pour pouvoir représenter les mots composés et d'obtenir leur flexion automatiquement. Par la suite, nous avons créé le dictionnaire des mots composés du macédonien et grâce aux graphes de flexion nous avons compilé le dictionnaire de toutes les formes fléchies des mots composés. L'examen linguistique montre que les graphes couvrent toutes les formes fléchies des mots simples et composés. Il est assez facile d'étendre cette ressource Unitex à d'autres corpus. Les actions nécessaires consistent à ajouter une annotation manuelle de la forme lexicale des mots simples et de lui associer un graphe de flexion et, si nécessaire, de modifier le graphe ou d'en créer un nouveau. Ensuite, ajouter la forme lexicale du mot composé et de lui associer le graphe correspondant, ou si nécessaire, de le modifier ou d'en créer un nouveau. Ainsi toutes les formes fléchies serraient-elles obtenues automatiquement.

Références

GROSS M.(1986). Lexicon-Grammar. The representation of compound words. Actes de *Eleventh International Conference on Computational Linguistics*, 1-6.

LEON J.(2004). Lexies, synapsies, synthèmes: le renouveau des études lexicales en France au début des années 1960. *History of Linguistics in Texts and Concepts*, 405-418.

PAUMIER, S. (2006). Unitex 3.0 User Manual. *Université Paris-Est.*

SAG I. A., BALDWIN T., BOND F. COPESTAKE A., FLICKINGER D.(2002). Multiword Expressions: A Pain in the Neck for NLP. *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing.*

SAVARY, A. (2008) "Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches", *Linguistic Issues in Language Technology,* 1(2):1–53.