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*)

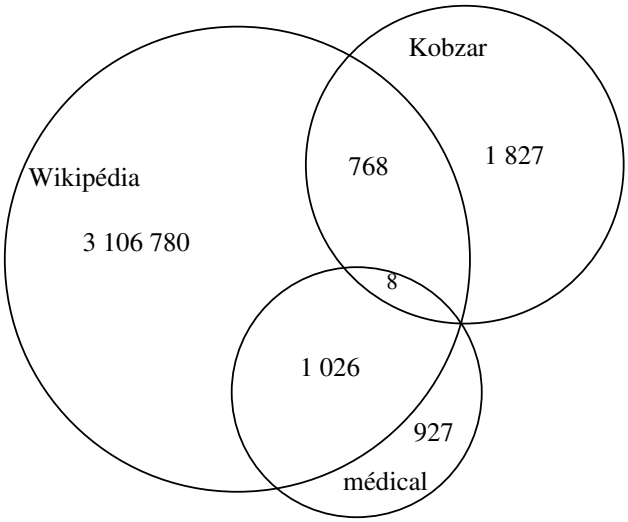

Figure 1 – Recouvrement de ressources acquises sur différents corpus.

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) {запланyйте; 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.