Unsupervised acquisition of morphological resources for Ukrainian

Thierry Hamon$^1$ and Natalia Grabar$^2$

$^1$ LIMSI-CNRS, Orsay, Université Paris 13, Sorbonne Paris Cité, France  
$^2$ CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France  
hamon@limsi.fr, natalia.grabar@univ-lille3.fr

Abstract. 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 is necessary for such tools for performing more sophisticated treatments (information retrieval, morpho-syntactic tagging, etc). We propose to acquire morphological resources for Ukrainian language. The method proposed exploits corpora in order to extract words that are related morphologically between them. The method has two versions: without and with processing of prefixes. 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 the corpora, precision varies between 67% and 86%. The results from different corpora are also compared, which shows that there is little redundancy between the corpora. The currently available resource contains 3,315 fully validated pairs of words.

1 Introduction

Morphological resources provide basic and crucial knowledge upon which many Natural Language Processing (NLP) applications are built. During the Part-of-Speech tagging and the lemmatization, morphological lexicon helps to analyze words and to recognize inflected forms of a given word and then to deduce its lemma. Thus, in morphologically rich languages, the recognition of affixes and of inflections allows to disambiguate and to deduce the Part-of-Speech category. In information retrieval and extraction, the needs and the goals go beyond the inflectional morphology. Indeed, such applications often require the identification of relations between derivations and even compounds. Generally, this information is helpful to collect higher number of relevant answers or documents, and then to increase the recall of automatic systems. The processing of unknown words is relevant for many NLP applications because the existing dictionaries or lexical resources are known to be uncompleted. In that respect, if morphological information on words is available, it can be useful to induce the grammatical or syntactical categories, as well as the semantics. In speech recognition, the resources providing groups of words with the same morphological root, are useful for the disambiguation of a spoken sequence and for selection of the most relevant candidate.
Nowadays, such resources are available and widely used in several languages, e.g. CELEX [4] for German, English and Dutch, Démonette [14], lexique.org and Lefff [27] for French, Morph-it [34] for Italian. Those resources provide inflected information associated with words, such as singular and plural forms of noun (\{president, presidents\}), adjectives (\{présidentiel, présidentielle\}) or verbs (\{preside, presided\}). It is less common to find resources that also provide relations between derivational forms (\{president, presidential\}) or between compounds and their basis (\{president, presidoscrapy\}). Moreover, the general-language morphological resources often show partial coverage in specialized domains because such documents involve specific lexicon.

Various methods have been proposed to acquire morphological resources. Among them, various kinds of information can be used in isolation or combined: associations between words in corpora [33, 35]; distributional properties of words in corpora [5]; distribution of letters in words in order to identify frontiers of morphemes [7, 32, 28]; analogy between the word formation in order to deduce or generate new constructed words [24, 12, 15]; frequency of the suffix couple, which insures the reliability of the semantic link between two words [10]; exploitation of dictionaries in order to identify words which are semantically and morphologically related within a given entry [19, 15]; semantically related pairs of terms [12]; samples used by supervised methods in order to induce morphological rules [2, 30, 24].

Several tools are available for morphological analysis in several languages: Flemm and Derif for French [23], Morphisto for German\(^4\), tools for Nguni [3, 25], Indian [1], or Macedonian [17]. Building of morphological resources is an active research topic, including specialized and low-resourced languages. Besides, several methods have been proposed for the automatic acquisition of morphological resources and consequently various types of data can be processed for the acquisition of such resources.

In our work, we propose to tackle the creation of morphological resources for Ukrainian. We propose to take advantage of freely available text corpora which are not annotated syntactically nor semantically. Our method relies on previous works [33, 35] and computes corpus-based associations between words. However, several adaptations are performed to take into account particularities of the Ukrainian language: text encoding, text segmentation and morphological specificities.

In the following, we first propose a description of Ukrainian language and mention some existing works on this language (Section 2). We then present the material in Section 3 and the method in Section 4. Results are described and discussed in Section 5. We conclude and indicate some future works (Section 6).

2 Specificities of the Ukrainian Language

Ukrainian language is part of the Slavic family of languages. It uses Cyrillic alphabet composed of 33 letters and of apostrophe. One particularity of Ukrainian is that the apostrophe plays phonetic role and is not a word separator.

Similarly to other Slavic languages, Ukrainian has a rich inflectional morphology, with seven cases and three genres for common and proper names, numerals, adjectives,

\(^3\) [www.lexique.org](http://www.lexique.org)

\(^4\) [https://code.google.com/p/morphisto](https://code.google.com/p/morphisto)
pronouns and some verbal forms. The derivational morphology plays a key role in the building of the lexical and grammatical structures (e.g. aspect, tense). To illustrate the word creation, we present set of words coined on walk (in (1)).

(1)  

хід (walk), вхід (entrance), вихід (exit), захід (East, sunset, event), прихід (arrival), перехід (crossing, cross walk), відхід (start, departure), відхід (approach), обхід (passing by, walk around)

Thanks to the rich morphology, even if sentences obey to the canonical order subject-verb-object (SVO), the word order is free without introducing stylistic effects. Ambiguities exist at lemma and inflection levels, and it is common to find inflected forms which correspond to different lemmas. These particularities may lead to difficulties with the classical NLP methods, and above all with the POS-tagging. However, they can also facilitate the sentence parsing since inflected information provides useful clues for this task [6]. Concerning the NLP work, we can mention some existing works: since 2010, the Ukrainian language is integrated in the Multex-East POS tagset\(^5\) [9]; UGtag POS tagger has been developed [18]. It exploits dictionary and rules, but does not perform syntactic and morphological disambiguation of words; recognition of named entities [16]; sentiment analysis [26]. As for the corpora, there were endeavor to build the Ukrainian National corpus\(^6\), Polish-Ukrainian [31] and Bulgarian-Ukrainian [29] parallel corpora. However the access to these corpora is restricted and, to our knowledge, there is no free existing corpora or lexicon. Our objective is to contribute to the development of NLP methods and resources dedicated to the Ukrainian language. Such methods and resources are necessary for boosting the development of NLP applications.

### 3 Linguistic Resources

We use three types of resources: (1) corpora (Section 3.1) that allow to acquire morphological resources; (2) set of stop words (Section 3.2) used in order to remove grammatical and invariable words, and (3) set of prefixes (Section 3.3).

#### 3.1 Corpora

The corpora are issued from three sources and represent three different genres:

-- **literary corpus**, composed of texts from Taras Shevchenko's Kobzar (89,289 words);
-- **medical corpus**, composed of medical articles and brochures issued from Medline-Plus [22] (46,230 words). We use those that are translated in Ukrainian;
-- **general corpus**, composed of articles from the Ukrainian Wikipedia pages (1,201,585 articles totaling 246,368,411 words are available in the used version).

Obviously, Wikipedia is the biggest corpus while MedlinePlus is the smallest.

\(^5\) http://nl.ijs.si/ME/V4/

\(^6\) http://ulif.org.ua/
3.2 Stopwords

We use a set of 385 stopword forms issued from an existing resource dedicated to the localization of graphical interfaces\(^7\). However, we observe that this list is incomplete and needs to be augmented through the corpora analysis.

\[(2)\] зі (with), ми (we), на (on), ма (and), мі (you), що (that/what), її (to her), їм (to them)

3.3 Set of Prefixes

The set of 73 prefixes is issued from the existing dictionary [36]. An example is given in (3) together with approximate translations. The prefixes are used for associating words with common bases that may occur after these prefixes, such as in Examples (1)).

\[(3)\] без (without), від (from), екстра (extra), з (perfective meaning), за (behind), до (up to), об (around), най (the most), непер (re), повна (over)

4 Approach for Building Morphological Resources

The corpora are first pre-processed (Section 4.1). We then apply our method to extract morphologically and semantically related pairs of words (Section 4.2), and to process the prefixes (Section 4.3). Besides, the reliable morphological rules from the first set of validated resources are used to prevalidate some of the remaining word pairs (Section 4.4). Finally, the evaluation of the acquired word pairs is performed (Section 4.5).

4.1 Corpus Pre-processing

The corpora are first converted in UTF-8, then a word and sentence tokenization is done. This step takes into account the role of the apostrophe character: inside words, it is not considered as word separator, while at frontiers of words, it shows its traditional quote meaning. Empty words are removed to avoid noise generation during the next step.

4.2 Extraction of Morphologically Related Pairs of Words

The method aims the identification of words which are related morphologically and semantically. We use for this the notion of thematic continuity. Indeed, thematic links exist at the lexical level within a piece of text: words and lexemes from a given semantic field tend to be used together (e.g., hospital, physician, operate). Consequently, words from the same morphological family may also co-occur (e.g., operate, operation). This provides the possibility to automatically find morphologically-related words in corpora.

Similarly to previous works [35], the notion of thematic continuity is approximated with a graphical window of \(W\) words. The morphological proximity between two words is then identified through the \(n\) first characters of each word. To summarize this first

\(^7\) https://github.com/fluxbb/langs/blob/master/Ukrainian/stopwords.txt
method (henceforth, standard method) [11], we collect words which share the same initial string with a length superior to $n$ characters and which co-occur in the same window of $W$ words. This last criteria will be measured with a statistical association measure which measures the frequency of the co-occurrence in comparison to a random association. We use the likelihood ratio [21] i.e. the ratio $\lambda = \frac{L(H_1)}{L(H_2)}$ between the probability to observe the number of co-occurrences of the word $w_1$ and the word $w_2$ according to the hypothesis $H_1$ where words are independent and the probability to observe the number of co-occurrences according to the hypothesis $H_2$ where words are dependent each other (we compute $-2 \log \lambda$). This ratio is computed as follows:

\begin{itemize}
  \item Probability to observe the $H_1$ hypothesis (independence): $L(H_1) = b(c_{12}, c_1, p)b(c_2 - c_1, N - c_1, p)$;
  \item Probability to observe the $H_2$ hypothesis (dependence): $L(H_2) = b(c_{12}, c_1, p_1)b(c_2 - c_1, N - c_1, p_2)$;
  \item Binomial law (probability of a sequence of $k$ success among $n$ draw): $b(k, n, p) = C_n^k p^k (1 - p)^{n-k}$;
  \item Elementary probabilities:
    \begin{itemize}
      \item $p = \frac{c_{12}}{N}$; $p_1 = \frac{c_{12}}{c_1}$; $p_2 = \frac{c_2 - c_{12}}{N - c_1}$;
      \item $c_1$ is the number of occurrences of the word $w_1$,
      \item $c_2$ is the number of the windows where the word $w_2$ occurs,
      \item $c_{12}$ is the number of windows in which $w_1$ and $w_2$ occur,
      \item $N$ is the number of words of the corpus.
    \end{itemize}
\end{itemize}

This association measure is asymmetric and depends on frequency of each word. For instance, there is a higher probability to observe a noun such as canal in the neighbor of its adjective canalized then the opposite. Given the two possible directions, the higher association score is kept. We process and evaluate the proposed pairs independently on their scores: even pairs with low scores may convey correct morphological relations.

The proposed method is applied on the three corpora. We use a window of 10 words on the left and on the right of the pivot word ($W = 21$). The minimal length of the initial string is fixed to 3 characters ($c = 3$) because it allows to keep pairs which may share common roots or bases.

### 4.3 Processing of Prefixes

Prefixes are very frequent in Ukrainian and play an active role in word formation. Prefixes may introduce two problems with the standard method:

\begin{itemize}
  \item they prevent from associating words with the same basis, which are preceded by such prefixes, such as in Examples (1);
  \item they associate words that do not have the same bases (and that have no morphological or semantic relations) but share only the prefix, such as in Examples (4);
\end{itemize}

(4) {заплануйте; запізнюйтесь} ({to plan, being late}), {відповідає; відстань} ({answer, don’t bother}), {переставляйте; перевіряйте} ({move, verify})

In the modified version of the method, we propose to inhibit the prefixes and to focus on the bases of words, even if they occur after these prefixes. In this modified version of the method, the prefixes undergo the following processing:
the known prefixes (Section 3.3) are temporarily removed from the words within a
given word pair, starting from the longest prefix: for instance, за (behind) is applied
before за (perfective meaning);

-- the standard method (Section 4.2) is applied to the remaining strings, with the min-
imal initial string set to 3 characters (c = 3): закритий (closed), відкритому
(Dative of open), відомим (unknown) temporarily become критий, критому and
омим;

-- if a given pair contains the common string with at least 3 characters, the prefixes
are restored and this word pair is kept as candidate. Thus, the pair {закритий,
відкритому} is now a candidate, while the pair {відкритому, відомим} is not
proposed. Notice the latter word pair is proposed by the standard method, but not
by the modified method. On the contrary, the modified method taking into account
the prefixes proposes the former word pair but not the latter one, as expected.

4.4 Rule-based Prevalidation

The proposed method induces very large number of word pairs. Their manual valida-
tion is a very tedious task. We propose to exploit the first set of the validated word
pairs for prevalidating the remaining word pairs. We use for this the morphological
rules issued from these validated pairs. For instance, the word pair {брат, брата}
(brother) provides the morphological rule /а, for the Genitive inflection of nouns. If this
rule is always reliable in the validated dataset, then it can be used safely for prevalidating
other word pairs that show the same rule, such as {імператор, імператора} (imper-
ator), {благовєщенськ, благовєщенська} (Blahoveshesk), {травник, травника} (clover). We use rules that have at least three correct occurrences in the validated dataset.

4.5 Evaluation

The results are evaluated manually by a native speaker. For the evaluation of the results,
the task is to check whether the words from a given pair share the same morphological
basis and are morphologically related.

5 Results

The corpora are pre-processed and the proposed method permits to extract a large set
of word pairs which are expected to be morphologically related. In Table 1, we indicate

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Standard # pairs</th>
<th>Standard Precision</th>
<th>Modified/Prefixes # pairs</th>
<th>Modified/Prefixes Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kobzar</td>
<td>2,546</td>
<td>76.4%</td>
<td>2,550</td>
<td>75.6%</td>
</tr>
<tr>
<td>MedlinePlus</td>
<td>1,961</td>
<td>68.8%</td>
<td>1,757</td>
<td>76.7%</td>
</tr>
<tr>
<td>Wikipedia (evaluated sample)</td>
<td>29,968</td>
<td>65.4%</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Table 1. Number of word pairs and precision
the numbers of the evaluated word pairs. Due to the large number of extractions from Wikipedia (3,108,591) only a sample of 29,968 word pairs is evaluated by now. Precision obtained varies between 65% on Wikipedia, and up to 75% on literary and 77% on medical corpora. Our results are comparable with those obtained in a previous work on the medical French [35]. The second set of results is obtained with the version of the modified method dealing with prefixes. We can see that this modification of the method is suitable for the quality of the results: precision is improved on medical corpus and shows a slight decrease on literary corpus. Several new and correct word pairs are extracted. Same processing will be applied to Wikipedia corpus. The validated set (31,476 pairs) provides 23,326 correct words pairs and will be made available for the research community.

Our results also show that the method can be used in different languages when corpora are available, in order to bootstrap acquisition of morphological resources. Hence, validated word pairs permit to induce 1,176 morphological rules with at least 3 correct occurrences. These rules have been applied to the remaining set with 2,991,890 word pairs and permitted to prevalidate 723,553 word pairs. The first analysis of these pairs indicates that they are correct. Such approach allows to increase the size of the available resource.

In Figure 1, we present the coverage between the resources acquired from the three corpora. We can observe that the intersection is low and that most of the pairs are issued from Wikipedia. 52.6% of word pairs acquired on the medical corpus are also acquired on Wikipedia, while this ratio represents 29.6% of pairs from the literary corpus. Only 8 word pairs are acquired on the three corpora. They correspond to the inflected forms of words (Example (5)). This suggests that several corpora should be processed to reach a good coverage of morphological resources.

(5)  

Some of the acquired pairs contain ambiguous words which can correspond to different lemmas according to the domain. The following examples illustrate this point:
-- {поділися, подіюсь} \(\text{this pair contains forms of the verb to disappear. However, the word поділися, with a different stress accent, corresponds to the verb to share;}

-- the main meaning of the pair {діместь, дім} is to put somewhere. However, the word дім is also an inflected form of child;

-- the main meaning of the pair {гори, горить} is to burn while гори is also an inflected form of mountain.

We consider that such pairs are correct because at least one of their meanings is correct. Among the correct pairs, an important part contains inflections, even if lemmas occur seldom. We also observe derivations (Examples (6)) and compoundings (Examples (7)).

(6) {алергійна, алергія} (\{allergic, allergy\}), {братьсячая, брате} (\{brotherly, brother\}), {вакцинацію, вакцина} (\{vaccination, vaccine\}), {дитину, дитячий} (\{child, childish\})

(7) {ангіопластика, ангиограми} (\{angioplasty, angiogram\}), {бронхіоли, бронхіт} (\{bronchiole, bronchitis\}), {газованих, газоутворення} (\{gaseous, production of gas\})

By comparison with French or English, we identify two specific morphological phenomena in Ukrainian: diminutive forms such as those presented in (8), patronymic forms such as those presented in (9).

(8) {ангелочко, ангел} (angel), {біленькі, білих} (white), {Богданочку, Богдане} (Bohdan), {воленьки, волі} (freedom), {годину, годиночку} (hour)

(9) {Іван, Іванович} (\{Ivan, son of Ivan\}), {Микола, Миколайович} (\{Mykola, son of Mykola\})

Main errors occur when words with the same initial string have no morphological or semantic relations (Examples (10)), and in case of allomorphs which occur within the initial string (the first \(c = 3\) characters) such as in (11).

(10) {криза, крийму} (\{crisis, Crimea\}), {проблем, прокурорської} (\{problem, prosecutor (adj)\})

(11) {ход, хода} (walk), {воля, вільний} (\{freedom, free\})

6 Conclusion and Future Work

We presented an approach for the acquisition of morphological resources for Ukrainian. An unsupervised method is proposed, which does not require annotations or dedicated resources and only relies on the use of corpora. Two versions of method are designed and tested: standard method and modified method with the processing of prefixes. Statistical association metrics between words are used to assess the probability of semantic and morphological relations between words. Our approach has been applied to three Ukrainian corpora: literary, medical and general language. For now, a set of 23,326 word pairs have been judged correct. 723,553 more word pairs are prevalidated with reliable morphological rules. This set will be progressively enriched with more validated data.
and made freely available for the research purposes. The method allows to acquire word pairs with precision which varies between 65% and 77% according to corpora.

One limitation is related to allomorphy which occurs within the initial string. Specific methods or rules will be tested for the processing of such situations. Another problematic point is that manual evaluation is a time-consuming task. To reduce the validation time, we plan to use other association and statistical measures [13, 20], and metrics from the graph theory [8]. Morphological rules induced with the method can also be used for enriching this resource. We plan to use this resource for POS-tagging and information retrieval.

References

36. Клименко, Карпіловська, Карпіловський, Недозим, Словник Афіксальних Морфем Української Мови. Інститут Мовознавства ім. О.О. Потебні Національної Академії Наук України, Київ, Україна (1998)