Principles of organising a common morphological tagset for PolUKR (Polish-Ukrainian Parallel Corpus)*

Natalia Kotsyba1, Olga Shypivska2, and Magdalena Turska3

1 Institute of Slavic Studies, Polish Academy of Sciences
2 ULIF, National Academy of Sciences of Ukraine
3 Warsaw University

Abstract
In this article we discuss some theoretical issues concerning the organization of a common morphological tagset for Polish and Ukrainian which should be the basis for a consistent search through PolUKR, the parallel corpus for the two languages. Keywords: morphosyntactic tagset, grammatical categories, contrastive grammar, PolUKR, parallel corpora, Slavic languages

1 Introduction. Existing approaches and theoretical background issues
PolUKR is intended first of all for human users, ranging from professional linguists to translators and language learners, although it could also be used for computer-aided language processing. The corpus layout and search facilities are expected to help a convenient and objective linguistic research, this is why we aim to create a simple intuitive common set of grammatical categories to search through the corpus. The present version of the corpus is the extension of the pilot project that appeared in 2005 as one of the first parallel corpora for Slavic languages, cf. Kotsyba 2006, Turska 2007, Turska 2006. Consistent morphosyntactic tagging is one of the main tasks of the present version of the project.

Monolingual corpora have been developed for practically all Slavic languages, and in each case the issue of morphosyntactic tagging was resolved in its own way, most often according to the dominant grammar description. As for parallel corpora including two or more Slavic languages (the few existing examples concern mostly Balkan languages: Multext-East 2003, Paskaleva 2007, Vita 2007), we have encountered two ways to approach morphological tagging of bi-texts. One of them presupposes the use of independent annotation schemes from corresponding monolingual corpora1, and is an efficient solution for quick results. However, it is

*http://corpus.domeczek.pl. The current, extended version is based at the Institute of Slavic Studies, Polish Academy of Sciences, and is partially supported by MNiSW (Ministry of Science and Higher Education) grant N N 104 0403 33.
1Regensburg Parallel Corpus: http://www-cgi.uni-regensburg.de/Fakultaeten/Slavistik/Corpus/parallel/parallel.html.
hardly convenient even for experienced linguists, who will have to learn annotation peculiarities for each represented language, not to mention non-professional users.

Another option is to create a common tagset. The first common tagset embracing Slavic languages\(^2\) was developed for 11 structurally different languages in the framework of the project MULTEXT 2000 and its extension MULTEXT-East 2004 (Multext-East, 2003). All existing morphosyntactic categories and their possible values were basically combined into a common matrix according to the descriptions found in the academic grammars of the languages. As the authors themselves say, much work was done to find as many matches across the languages' grammatical categories as possible. Still, the resulting tagset was later (Przepiórkowski, 2003), (Vita, 2007) criticized for its inconsistency and the lack of clear organizing principles\(^3\).

2 Tagset for PolUKR. Input material

The morphological tagset for Ukrainian was developed in the 1960’s by a research group at the Institute of Linguistics of the Ukrainian National Academy of Sciences. It was used with only slight changes by the ULIF group and is described in Shirokov 2005. The tagset consists of two-letter symbols that correspond to prototypical word forms with a common set of morphological values. This solution can be considered as sooner technical than user-oriented. One of the features of the tagset codes is that they include both small and capital, Latin and Cyrillic, Unicode symbols. Some of Latin and Cyrillic letters are visually identical. Thus, it is difficult for a human to both identify and use them. The tags are not mnemonic either.

An example of a tagged text in Ukrainian (chains of two-letter codes mean that there are several grammatical interpretations for a given form):

"Озвалась<гИ> сторожа<КИЙРЙВ>  
I<СПССN0Ь0> пролунав<гЕ> відбiй<ЙИЙВ> .<e>  
Хоч<Ь0СПГС> покарати<гФ> можуть<ГЮ>  
Я<МQ> , друже<ЙК> , мчу<ГГ> мерщiй<Н0> .<e>  
Сховай<гЬ> в<ПВПППР> душi<КРКДКПКАКУКШЙПЙАЙУЙШ>  
прощальний<АИАВ> трем<ГЖ> .<e>  
Як<Ь0СПН0ЙИ> вiрилось<ГЛ> , що<Ь0СПМСМХ> ми<МА>  
pідем<гЖ>".

The existing tagset for Polish was developed for the IPI PAN corpus (Korpus, 2004). We accept the IPI PAN’s tag layout\(^4\) as the more intuitive and mnemonic in comparison with the MULTTEXT pattern. As for the distribution of categories, some of IPI PAN purely technical solutions, like creating separate categories for words with atypical syntactic behaviour ("winien" "bedzie"), seem to be impractical for human-oriented search. The categories have been regrouped to represent

\(^2\)Neither of our two languages was considered there.

\(^3\)For example, the treatment of participles needs more clarification there. The Czech form that corresponds both functionally and etymologically to active participles in other Slavic languages is called transgressive, while for Bulgarian the equivalent form is identified as a gerund, which unnecessarily increases the quantity of tags and dissociates comparable categories.

\(^4\)Like, for example, [verb:perf:sg:past].
a more traditional distribution of parts of speech. The tagset has been extended
to cover Ukrainian-specific morphological classes.

In general, we try to compare language systems rather than their grammatical
descriptions, to match identical categories with differing names. We have decided
to organize the common Polish–Ukrainian tagset bearing in mind also other Slavic
languages that might be later engaged in parallel corpora projects. Part of the
task has already been done in the framework of the earlier mentioned MULTTEXT-
East. Apart from this linguistic typological approach, we were guided by reasons
of practicality in some of the solutions. One example is eschewing the category
of predicative, which is considered a separate part of speech in the Russian and
(partly) the Polish grammatical tradition, although equivalent lexemes might be
treated as adverbs or nouns in descriptions of other Slavic languages.  

3 Some problems and proposed solutions

The preliminary comparison of the available tagsets developed for Ukrainian and
Polish shows that the approaches used to describe the morphological systems differ considerably and mapping the tagsets demands a careful analysis of theoretical
grammatical foundations. The mere adding up of the categories gives a discourag-
gingly small common set—only 6 categories coincide formally. As many as 21 of
the categories from this formal union of tagsets (counting 50 entries) are unique
for the Ukrainian morphological system and 23 are unique for the Polish one.

Methodological (ULIF’s approach is more empirical, while IPI PAN’s is more
generative) and at times pragmatic differences lead to a different conceptualization
of categories. For instance, the ULIF scheme treats comparative and superlative
adjectives as adjectival lexemes, while according to the IPI PAN tagset adjectives
and adverbs can be characterized by degree. Both tagsets include the category of
predicative, although the scope of its grammatical meaning differs considerably.
A separate solution had to be found for each of these issues.

We have decided to leave out purely syntactic information other than the stan-
dard differentiation between prepositions, conjunctions and particles. For the sake
of simplicity some subcategorizations were ignored, e.g., the case that a preposi-
tion requires and types of conjunctions. The same was done for purely semantic
information issues, e.g., Ukrainian pseudoreflexive verbs ending with „-ся” have
been added to the general group of verbs. Thus, the verbal paradigm was reduced
to the half of its size.

The Ukrainian tagger singles out proper names, which are a semantic category.
Having in mind the possibility of future semantic tagging we have decided to keep
this information and allow it to be searchable. Therefore, a special tag "propnouns" as opposed to the general "gnouns" was introduced, and the noun paradigm was
doubled for the time being.

---

5 More arguments for this solution are given in 3.4 below.
6 Here and further we refer to Korpus 2004, Przepiórkowski 2003, Wolinski 2003 for Polish and
7 Information about the required case for prepositions is redundant as it belongs to the dic-
tionary level.
3.1 Participles

Adverbial participles are characterized by temporal sequence dependency. This dependency is often called "tense" in traditional grammars, since for simplification the main verb is treated as a reference point, as the speech time in a simple sentence. Hence, participles expressing states that are simultaneous with or prior to the main action are called present and past, respectively. This simplification is possible due to the "indifference" of the main verb towards the actual tense. A similar scheme could have been used for Ukrainian, were it not that it has a participle that can only be used with past main verbs. This participle is characterised not only by temporal sequence dependence but also by tense, cf. the forms:

1. verb, adverbial participle, perfective aspect, past tense, active voice, tag: "VW", example: прочитавши (having read);
2. verb, adverbial participle, imperfective aspect, past tense, active voice, tag: "UW", example: читавши (reading in the past);
3. verb, adverbial participle, imperfective aspect, *present tense, active voice, tag: "UQ", example: читаючи (reading).

This makes it necessary to keep the proper category of tense for this form, whereas sequence dependency is expressed by aspect. The tense-independent form роблячи, corresponding to Polish робиць is interchangeable with робивши as far as the mentioned grammatical restrictions are concerned. Further discrimination of their meanings exceeds the scope of the present paper.

The table below shows existing participles for Polish and Ukrainian. Forms with asterisks are ungrammatical, presented here only for exemplification purposes.

<table>
<thead>
<tr>
<th>active</th>
<th>active</th>
<th>active</th>
<th>passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>adverbial</td>
<td>adverbial</td>
<td>adjective</td>
<td>adjective</td>
</tr>
<tr>
<td>past</td>
<td>tense-irrelevant</td>
<td>tense-irrelevant</td>
<td>tense-irrelevant</td>
</tr>
<tr>
<td>робивши</td>
<td>роблячи</td>
<td>*роблячий</td>
<td>роблений</td>
</tr>
<tr>
<td>*робиwszy</td>
<td>робiс</td>
<td>робiсь</td>
<td>робiонy</td>
</tr>
<tr>
<td>(while) doing (in the past)</td>
<td>(while) doing</td>
<td>(the one who is) doing</td>
<td>(being) done</td>
</tr>
</tbody>
</table>

To differentiate between situations with absolute and relative tense restrictions we will use the terms: simultaneous participle (pcon), anterior participle (pant) and simultaneous past participle (ppast), where the last one is specific to Ukrainian.

3.2 Pronouns

Pronominal groups were added to their corresponding word classes: pronouns proper are grouped with nouns, proadjectives (differentiated only for Ukrainian)
are grouped with adjectives. Following the IPI PAN version, we divided personal pronouns into 1–2 and 3 person. Person can be further specified during the search through a lemma (as we considered assigning a tag for one word a redundancy). The Polish tagset treats the reflexive pronoun *siebie* as a separate flexeme, while Ukrainian considers it a 3p pronoun. For the sake of simplicity this word is searchable within the 3p pronoun group. Proadverbs are not differentiated for either language.

### 3.3 Adjectives

Polish post-prepositional adjectives (adjp) that correspond to the adjectival formant of adverbial expressions like *polsku* in *po polsku* were moved to the adverbs section, as similar expressions in their full form are referred to as adverbs in the Ukrainian tagset. Such formants do not function independently as adjectives.

Another productive Polish adjectival formant, ad-adjectival adjective (adj+a), *polsko-* is included in the adjectives proper group. In the Ukrainian grammatical dictionary such formants are not grammaticalized but presented as part of the lexicon. The possibility of searching according to this tag remains only at the level of the query language.

The degree of comparison for Polish is defined at the tag level, while for Ukrainian the comparative and superlative degrees of both adjectives and adverbs are independent lexemes. An algorithm has been developed and implemented to differentiate between degrees of Ukrainian adjectives.

Cardinal numerals are treated as adjectives for both languages.

### 3.4 Predicatives

Predicatives\(^9\) are probably the most problematic category from a theoretical point of view. They can be found in the IPI PAN Corpus, the Ukrainian explanatory dictionary SUM and the Russian National Corpus, and the intersection of translation equivalents in all of them is remarkably low.

Predicatives are most numerous in the Russian corpus. A relatively new explanatory dictionary by Tatiana Efremova (Efremova, 2000) contains about 1200 predicatives. Many adverbs that can function as predicates in a sentence are presented as both adverbs and predicatives. Comparative degrees of adverbs are treated as predicatives as well, which leads to an increase of homonyms that do not differ from the point of view of lexical semantics. Many translation equivalents of such lexemes will be considered adverbs for Ukrainian: *на плавь*, *напидпітку* (afloat, drunk) (UGD, 2004), but still the set of predicative words there amounts to 176. We have identified 26 predicative forms in the 15 mln segment of the IPI PAN corpus. Due to the purely syntactic nature of predicatives one may question

\(^9\)The category of predicative was introduced in 1928 by Lev Shcherba and included words like нельзя, можно, надо, пара, весель etc. (this group basically corresponds to the core of Polish predicatives in the IPI PAN Corpus) as “it was difficult to assign them to any part of speech”. Shcherba put those words in the same group with such as холодно, светло, весело; готов, должен, рад, болен, быть навеселе, находиться, замужем (it is cold, ready, glad, ill, drunk, married) and called them the category of state. Sources: http://spravka.gramota.ru, http://forum.gramota.ru. Cf. also Morfologia 1998.
the necessity of differentiating them as a separate part of speech at all. They are defined by functional, not morphological characteristics, which might be compared to calling sentence predicates, or subjects, parts of speech.

Given the variability in understanding of the category of predicatives, we had to regroup their sets. The existence of homonymic forms belonging to other parts of speech helped in assigning these to existing categories.

The following subgroups were removed from the Ukrainian predicatives (the category that appears after an arrow is where those uses will be found):

1. predicatives of semelfactive origin: зирк, крут-верть → interjections (as many of them are already categorized as interjections in the dictionary);
2. state words of adverbial origin: зимно, безвітряно, безсніжно (it is cold, not windy, no snow) → adverbs;
3. diminutive verbs like спінькати, спатоньки, спатки, ходитоньки, їстючки, їстки, їстоньки, їсточки → infinitives.

The most frequent predicative in the IPI PAN corpus to, when used in the subject position in sentences like To ksią˙gka (This is a book), was combined with its "subst" uses. The words like mo˙zna, trzeba, wolno, wiadomo, trza, niepodobna, podobna, doś, dosy´c, being in general marked by modal semantics, as well as their Ukrainian equivalents, are referred to as modal adverbs. The remaining handfuls of deverbatives: słucha´c, wida´c, sta´c, ču´c, zna´c (one hears, one can see, it is felt, it is known), and denominatives: szkoda, potrzeba, żal, usłys, strach, pora, czas, brak, śmiech, are treated as adverbs.

4 Search options

Conversion tables have been prepared to automatically convert Ukrainian tags into the common chain format taking into account all the above mentioned nuances.

As we have seen, mapping of languages, even in the case of structurally close ones, inevitably leads to further granulation of categories, which can be useful for advanced language research but might be distracting for an average user. In order to meet different requirements we have decided to enable three levels of search:

- by exact form;
- by lemma with additional morphological options presented in a special table;
- using Poliqarp-like tag formulas (for advanced users who will have to learn the query language and the tag lexicon).

Table 2. Morphological restrictions for lemma-based search. Language- or tagset-specific categories are marked respectively (PL) or (UA):

10 A set of Ukrainian predicatives was extracted from the Ukrainian grammatical dictionary (UGD, 2004), courtesy of ULIF.
11 Diminutive verbs are almost exclusively used in speaking with or about children. Cf. also Czech diminutive verbs ending in -inkat, like spinkat (sleep), blinkat (vomit).
12 http://korpus.pl/?page=poliqarp.
Since we try to consider the linguistic background of potential users, morphological categories and their values in the second search option roughly correspond to traditional sets of grammatical categories and are presented in Table 2, which
is also the basis for the graphic interface of this search option. A possibility to select several subcategories of different categories seems to be a way out in the situation of unclear POS boundaries.

Certain selectional restrictions at the ready-options search level connected with the objective grammatical characteristics (like impossibility of selecting simultaneously infinitive and the genitive case) are planned but will not be specified here.

5 Suggestions for future work and conclusions

We have tried to use as much of the grammatical information given by the existing available taggers as possible. There is no end to further atomization, but the picture we receive is already quite informative. The following tasks may still be undertaken in the future:

- Dividing Ukrainian adjectival into adjectives proper and adjectival passive participles (those that still can be related to actively used verbs);
- Singling out more diminutive verbs (rule based) and treating them as normal verbs (considering the fact that they possess an infinitival form);
- Further classification of particles/qubliks;
- Semantic tagging.

It is clear that the task of creating a common tagset, especially for structurally differing languages (leaving alone the multiplication of the problem with each added language), belongs to the realm of the typological linguistics and the present impossibility of constructing a uniform and conflictless tagset is rather the mirror of the current situation in the theory of grammar and cannot be resolved as one of "by-tasks" ad hoc. The current, "encyclopaedic" approach of linguists towards the issue of word classes can be illustrated by the following citation: "The solution to the generality problem that is usually adopted (often implicitly...) is that one defines word classes on a language-particular basis, and then the word class that includes most words for things and persons is called ‘noun,’ the word class that includes most words for actions and processes is called ‘verb,’ and the word class that includes most words for properties is called ‘adjective.’ However, the subclass problem has not been solved or even addressed satisfactorily, and the use of word-class notions in a general or cross-linguistic sense remains problematic" (Haspelman, 2001). The problem is also closely connected with the idea of a universal language and a universal grammar. The Universal Networking Language project is an example of such a language currently being developed by a large international group of linguists and its success may bring solutions to the problem of tagsets in general in the future. We believe that parallel corpora can also serve as a good material database for further theoretical linguistic developments, including not only grammatical classification itself, but also more semantically and ontologically oriented issues.

References


Volodymyr Shirokov et al. (2005), В. А. Широков, О. В. Бутаков, Т. О. Грабейнхіна, О. М. Костишнін, М. Ю. Крітин, Т. П. Любченко, О. Г. Рабулець, О. О. Сидоренкo, Н. М. Сидорчук, І. В. Шевченко, О. О. Шипинська, К. М. Якіменко. *Корпусна лінгвістика*. Київ: Довіра.


Magdalena Turska and Natalia Kotsyba (2007), Polish–Ukrainian Parallel Corpus and
its Possible Applications, in *Proceedings of the International Conference on Practical Applications in Language and Computers*, 7-9 April, Łódź, Peter Lang GmbH.


