УДК 811.162.1=162.2’374.822:004.65

Natalia Kotsyba,
Faculty of „Artes Liberales”, Warsaw University, 69, Nowy Świat str., Warsaw, 00-046, Poland

OVERVIEW OF THE UKRAINIAN LANGUAGE RESOURCES WITHIN THE MULTILINGUAL EUROPEAN MULTEXT-EAST PROJECT, V.4

The article presents an overview of computational resources for the Ukrainian language within a multilingual European MULTEXT-East project (MTE, http://nl.ijs.si/ME/V4) freely available for researchers since May 2010, including a formal representation of morphosyntactic specifications consisting of 1239 unique grammatical tags in the XML, TEI-5 compatible, format and a morphosyntactic lexicon covering over 200000 wordforms with lemmas and morphosyntactic codes.

Key words – computational language resources, NLP, TEI, Text Encoding Initiative, standards, Ukrainian language, morphosyntactic specifications, morphosyntactic lexicon.

У статті представлено огляд комп’ютерних ресурсів для української мови, створених в рамках багатомовного європейського проекту MULTEXT-East (MTE, http://nl.ijs.si/ME/V4), доступних безкоштовно для дослідницьких цілей від травня 2010 року. Ресурси охоплюють формальну репрезентацію морфологічно-синтаксичних специфікацій 1239 унікальних граматичних тагів у форматі XML, згідному з вимогами TEI-5, та морфологічно-синтаксичний лексикон на понад 200000 словоформ разом з лемами та тагами.

Ключові слова – комп’ютерні мовні ресурси, обробка природньої мови, TEI (Ініціатива Кодування Текстів), стандарти, українська мова, морфологічно-синтаксичні специфікації, граматичний таг, лема, морфологічно-синтаксичний лексикон.

Introduction. Aim of the article

Due to historical reasons, developing of computational resources for the Ukrainian language was discouraged in the times of their rapid growth for widely used world languages like English or Russian, which is the reason why at present there is still no solid computational linguistic base for Ukrainian in terms of both materials and original theoretical works, cf. [14:4]. One of the consequences of this situation is a continuing strong orientation at the modern Russian corpus linguistics, which, notwithstanding the strong post-Soviet scientific heritage, itself is largely influenced by the developments of English linguistic resources. Hence, there is a considerable gap between the modern Ukrainian corpus and computational linguistics and the most recent work in this field done in the Western world. This is the reason why worldwide initiatives involving Ukrainian are beneficial for following good practices in the field, and knowledge about them should be disseminated among present and potential researchers. Thus, the purpose of the article is to present to a wide audience the existing linguistic resources for Ukrainian developed within a recent international project in a possibly accessible way, shedding also the light on some linguistic nuances and the preceding work in theoretical linguistics that led to taking certain decisions connected with the linguistic organization of the specifications, and encourage their further use by researchers.

Existing morphosyntactic encodings for Ukrainian

The task of computational morphosyntactic description of Ukrainian has been approached by several researchers/research groups both from the theoretical and practical perspectives. A system of 383 synthetic morphosyntactic codes is used in the National Corpus of Ukrainian and is well documented [15]. It is based on the Ukrainian Grammatical Dictionary developed by Igor V. Shevchenko in the 90-ties of the

XX-th century. Unfortunately, the morphosyntactic information is still not available for search in the corpus due to unresolved disambiguation in the texts. Another corpus of Ukrainian [mova.info] enables online search with morphosyntactic restrictions but they are not documented at all, at least at the time this paper is being written[1]. Neither of the two above mentioned corpus creating initiatives makes any tagging resources available to a wider public for individual tagging purposes, hence, potential users have to deal with a situation of informational vacuum in this regard. The existing schemes are characterized by exclusive use for a single project.

A purely theoretical approach is presented in [10], [11]. The latter gives inter alia a detailed account of the tagset architecture used in the project to be presented in this paper [11:197—230], comparing it to the earlier proposals of the same author [10].

General structure of the MTE resources

The international MULTEXT-East project [8] is dedicated to a uniform, harmonized presentation of language resources, enabling their further use in various information systems and easy data interchange. It was launched as MULTEXT project for six Western European languages in 1995 and further extended to some Central and Eastern European languages, under the name of MULTEXT-East (further MTE for short). Since 1998 it has been an on-going project. Each new version aimed at extending either the number of languages or types of resources within the existing language parts, correcting detected errors, eliminating inconsistencies, or updating the format of the data in order to be up-to-date with the latest technological advancements. The Ukrainian language was included in version 4, May 2010.

The specifications for all 16 languages in version 4 are licensed under the Creative Commons licence Attribution-ShareAlike 3.0, which means that they are freely available from the project's website for download. The lexicon is available free of charge for use for academic purposes and is available upon registration.

Due to the XML family technologies and the uniform XML encoding used in MTE, different types of resources, such as the specifications, lexica, and the morphosyntactically annotated corpus[2] are well integrated, "making it possible to easily move between different representations of the same data" [3:1].

Common part of the morphosyntactic specifications

MTE morphosyntactic specifications are "a TEI P5 document that provides the definition of the attributes and values used by the various languages for word-level syntactic annotation, i.e., they provide a formal grammar for the morphosyntactic properties of the languages covered" [3:2]. Apart from the formal parts the specifications contain comments, bibliography, various metainformation. The specifications consist of the front matter, the common part, describing features that are common for all the languages, and the language particular part. The common part includes definition of categories (parts-of-speech) and their possible features, comprising attribute and their values. "The morphosyntactic specifications also define the mapping between the feature-structures and morphosyntactic descriptions (MSDs), which are compact strings used in the morphosyntactic lexica and for corpus annotation" [3:2]. MSDs are similar to what is known as POS (part of speech) tags in that they are grammatical codes carrying morphosyntactic information at the wordform level with the exception that they have analytic, not synthetic representation, and far more detailed to be associated with the POSes only, especially in the case of such morphologically rich languages as Ukrainian.

There are 12 main morphosyntactic categories which correspond to traditional linguistic parts of speech[3]. Each category has its own attributes and their values, information about which of the 16 languages uses each particular attribute-value pair is also present. While most categories are the same for the

[1] NB: There is a considerable amount of grammatical errors in KUM due to automatic disambiguation. A relatively detailed analysis can be found in [4], forthcoming, the draft is available at http://domeczek.pl/~natko/papers/guide_ukrcorpora.pdf

[2] Relevant language versions of George Orwell's novel „1984" make parts of the multilingual parallel corpus of MTE. By the time version 4 appeared no Ukrainian translation of the full text existed which is the reason why it was not included into the release of 2010.

[3] One of the categories, residual (R), is used for technical reasons, such as for unknown wordforms.

languages (with such exceptions as e.g. absence of articles in the Slavic parts), their attribute-values combinations differ significantly, both due to objective reasons, as well as using different linguistic traditions for language descriptions. The category code is the first element of the resulting tag, and each attribute takes a fixed position after it. Every such position is encoded by a one-character code.

Language specific part of the morphosyntactic specifications

Language particular parts use an appropriate selection of attributes for the given language, so that the tags are less cumbersome and include only essential information. For example, the general MTE Noun category has 14 attributes, but Ukrainian uses only 5 of them. The Ukrainian part of the specifications also includes localization of all the linguistic terminology across the specifications, leaving the codes English-based for simplicity. A fragment of the specifications in XML format:

```
<div select="uk" type="section" xml:id="msd.V-uk">
<head xml:lang="en">Ukrainian Verb</head>
<table n="msd.cat" xml:id="msd.cat.V-uk" select="uk">
<head xml:lang="en">Specification for Verb</head>
<row role="type">
<cell xml:lang="en" role="position">0</cell>
<cell role="name" xml:lang="en">CATEGORY</cell>
<cell role="name" xml:lang="uk">частина_мови </cell>
<cell role="value" xml:lang="en">Verb</cell>
<cell role="value" xml:lang="uk">Дієслово</cell>
<cell role="code" xml:lang="en">V</cell>
</row>
<row role="attribute">
<cell xml:lang="en" role="position">1</cell>
<cell role="name" xml:lang="en">Type</cell>
<cell role="name" xml:lang="uk">тип</cell>
<cell role="values">
<table>
<row role="value">
<cell role="name" xml:lang="en">main</cell>
<cell role="name" xml:lang="uk">основне</cell>
<cell role="code" xml:lang="en">m</cell>
</row>
<row role="value">
<cell role="name" xml:lang="en">auxiliary</cell>
<cell role="name" xml:lang="uk">допоміжне</cell>
<cell role="code" xml:lang="en">a</cell>
</row>
</table>
</cell>
</row>
…
```

The core of Ukrainian specifications is based on the Ukrainian Grammatical Dictionary (UGD) developed by Igor V. Shevchenko (http://lcorp.ulif.org.ua/dictua), and a morphological analyzer (UGTag) which uses an extended version of the UGD. The additional features in comparison with the UGD embrace among others: the degree attribute for adjectives and adverbs, full paradigms of adjectival participles, pronouns as a separate part of speech with detailed semantic categorization. The degree of the data reorganization and extension can be demonstrated by the quantity of the resulting tags used in ULIF corpus – 383 unique tags [15: 420—434] and 1239 tags in MTE version (considering that some of ULIF related tags like passivity of verbs was disposed of). The difference is otherwise largely due to a detailed description of pronouns in the MTE version. Morphosyntactic data were also rearranged to better reflect MTE categorisations.

Below are listed and explained some of the specific features of grammar presentation of Ukrainian in MTE.

Ukrainian *pluralia tantum* nouns are not encoded directly but can be identified by the absence of a value of Gender ("-"). The Gender value "common" is assigned to nouns that can combine with adjectives in either feminine or masculine, e.g. *сирота* or either neutral or masculine gender, e.g. *Самоа*.

Gerunds are not differentiated, but could be treated as a special class of nouns, *nota bene*: they possess aspect.

No voice category is used for Ukrainian verbs as all verbal forms are active (adjectival/attributive participles are treated as adjectives).

Relative adjectives (Ukr. “відносні прикметники”) are labelled "o(rdinal)" for the sake of consistency with the Slovene tagset, where this term translates Slovene *vrstni* (*pridevniki*).

The feature "Animate" in adjectives is used to differentiate between two accusative masculine forms.

Adjectival participles are grouped with adjectives and are characterized by voice, quasi-tense and aspect. Although active adjectival participles are considered ungrammatical, being a consequence of russification in Ukrainian, they still can be found in the language use. Thus, they are not generated by the Ukrainian grammatical dictionary but codes for them are foreseen in the UGTag and the MTE MSD index.

Many pronouns can be assigned to more than one Type. The Referent_Type feature is used to show the additional feature, like possessiveness or personality. The main type is defined according to the grammatical tradition. Note: there is no PRONOUN as POS in the Ukrainian Grammatical Dictionary, pronouns are a class of nouns. The Syntactic_Type shows further POS distribution.

Table 1

Fragment of Ukrainian MSD index

MSD tag	English description (verbose)	Ukrainian description (verbose)	Example wordform/ lemma/ quantity

Ncnpdy	Noun Type=Common
	Gender=Neutral
	Number=Plural
	Case=Dative
	Animacy=Yes
		Іменник
		тип=загальний
		рід=середній
		число=множина
		відмінок=давальний
		істота=так
			ангелятам/
			ангеля
			451

Vmpip3s	Verb Type=Main
	Aspect=Progressive
	Verb Form=Indicative
	Tense=Present
	Person=Third
	Number=Singular
		Дієслово
		тип=основне
		вид=недоконаний
		тип=дійсна
		час=теперішній
		особа=третя
		число=однина
			абеткує/
			абеткувати
			24694

Afcnsdf	Adjective
	Type=Qualificative
	Degree=Comparative
	Gender=Neutral
	Number=Singular
	Case=Dative
	Definiteness=Full-Art
		Прикметник
		тип=якісний
		ступінь=вищий
		рід=середній
		число=однина
		відмінок=давальний
		форма=нестягнена
			азартнішому/
			азартніший
			554

Interrelation between the Ukrainian linguistic data and the common part of the MTE specifications is twofold. There are several codes that were introduced to MTE especially for the needs of Ukrainian.

The impersonal VForm (o) is characterized by the ending *-то*/*-но*. It exists in other Slavic languages as well, although in most of them it coincides with the neutral form of the passive adjectival participle and is classified as such. In Ukrainian, as well as in Polish, the attributive form is different from the predicative one, cf. in Ukrainian *писане правило* (“a written rule”) vs *писано правило* (“a rule was/is written”).

The emphatic (h) type of pronouns is also used only for Ukrainian, for predicative words like “нікому, нікого” with the stress at the first syllable and complex meanings like "there is nobody/nothing (to do sth/to use for doing sth, etc.)" that are classified in traditional grammars as either predicatives or pronouns.

Possible combinations of attribute-value pairs showing feature concurrence restrictions have also been deduced and included into the MTE specifications for Ukrainian. Some notes and examples were added where necessary.

Table 2

Combinations Showing Legitimate for Ukrainian Nouns

OS	Type	Gender	Number	Case	Animate	Example
N	p	N	p	ngdailv	N	Азовське
N	p	C	s	ngdailv	N	Самоа
N	p	Mf	sp	ngdailv	Ny	Марія, Ігор, Білинські
N	c	Cfmn	sp	ngdailv	Yn	лікар, ялина, вікно, сирота, роки
N	p	-	p	ngdailv	Yn	сани, Бережани
N	p	N	s	ngdailv	Yn	Здвиження

Version 4 distribution of MSD specifications includes associated XSLT stylesheets, available at http://nl.ijs.si/ME/V4/msd/xslt/, that can be used for different transformations of the specifications' data. The output is either in XML, HTML, or text format. There are three types of transformations: those for adding a new language to the specifications themselves, those transforming the specifications into HTML, and those validating and transforming a list of MSDs.

Specific XSL stylesheets and XSD schemes have been developed specifically for Ukrainian as well, mainly for the purposes of converting ready, annotated texts into others formats/grammars or validation purposes.

Morphosyntactic lexicon

Table 3 below presents the basic statistics about the morphosyntactic lexicon for Ukrainian.

Table 3

Ukrainian Morphosyntactic MTE Lexicon

Entries	Wordforms	Lemmas	MSD
318,547	205,348	15,162	1,239

The lexicon has the following format: wordform, lemma, tag, frequency (for the lemma). Below is the fragment of the lexicon:

a	a	I	0
a	a	Ccs	0
абзац	абзац	Ncmsnn	2
абзацу	абзац	Ncmsgn	
абзаца	абзац	Ncmsgn	
абзацу	абзац	Ncmsdn	
абзацові	абзац	Ncmsdn	
абзац	абзац	Ncmsan	
абзацом	абзац	Ncmsin	
абзаці	абзац	Ncmsln	
абзаце	абзац	Ncmsvn	
абзаци	абзац	Ncmpnn	
абзаців	абзац	Ncmpgn	
абзацам	абзац	Ncmpdn	
абзаци	абзац	Ncmpan	
абзацами	абзац	Ncmpin	
абзацах	абзац	Ncmpln	

абзаци	абзац	Ncmpvn	
аби	аби	Css	3
або	або	Ccs	4
або	або	Q	4

The lexicon is available for download and use for academic purposes due to the courtesy of its main author Igor V. Shevchenko.

Dissemination, use and approbation of the resources

As mentioned earlier, apart from the project’s website, a fairly detailed general description of the MTE Ukrainian tagset can be found in [11:197—230].

The MTE style tagset is used in UGTag, the only freely available morphosyntactic tagger for Ukrainian. It is also used in PolUKR project for the Polish-Ukrainian Parallel Corpus and in the presently developed experimental corpus of Ukrainian language. It has been applied for the corpus of Ivan Franko developed by S. Buk, as well as in a number of smaller student projects dedicated to building parallel or comparative corpora (by Ye. Mudrak, O. Predko, I. Kushniruk, R. Perkhach).

To approbate the tagset from the human user perspective, it was decided to use in an experimental disambiguation project at the Chair of Applied Linguistics of Lviv Polytechnical University, co-directed by A. Romanyuk and the author of this paper, where students had the task to manually disambiguate annotated pieces of texts and insert proper tags for words that were not in the original dictionary. A format restricting XSD scheme which allowed only legitimate tags was used with the tagged XML-formatted texts.

Fragment of a disambiguated, XML-formatted Ukrainian text with MTE tags:

```
<w_>
<w lemma="із" disamb="0" ana="Spsg">із</w>
<w lemma="із" disamb="1" ana="Spsi">із</w>
</w_>
<w lemma="нез'ясований" disamb="1" ana="Afpfsif">нез'ясованою </w>
<w lemma="ти" disamb="1" ana="Pp-2-ysin">тобою</w>
<w lemma="регулярність" disamb="1" ana="Ncfsin">регулярністю </w>
<w lemma="з'являтися" disamb="1" ana="Vmpip3s">з'являється</w>
<w lemma="янгол" disamb="1" ana="Ncmsny">янгол</w>
<w_>
<w lemma="із" disamb="0" ana="Spsg">із</w>
<w lemma="із" disamb="1" ana="Spsi">із</w>
</w_>
<w lemma="чорний" disamb="1" ana="Afp-pif">чорними</w>
<w lemma="бухгалтерський" disamb="1" ana="Afp-pif"> бухгалтерськими</w>
<w lemma="нарукавник" disamb="1" ana="Ncmpin">нарукавниками </w>
<w_>
<w lemma="й" disamb="1" ana="Ccs">й</w>
<w lemma="й" disamb="0" ana="Q">й</w>
</w_>
<w lemma="лупа" disamb="1" ana="Ncfsin">лупою</w>
```

The results of the experiment were satisfactory, which shows that the MSD tags are reasonably intuitive and mnemonic, and can be applied in a wide range of researcher and learner oriented corpora of Ukrainian[4].

MTE specifications have also been converted into an OWL ontology: “While TEI is more appropriate for authoring the specifications and displaying them in a book-oriented format, the OWL/DL encoding has the advantages of enabling formally specifying interrelationships between the various

[4] It is desirable to continue philologists’ disambiguation practice both for didactic purposes and for creating a manually approved disambiguated corpus of Ukrainian. Teachers and/or students willing to participate in manual disambiguation practice are invited to address the author of this paper via natalia@al.uw.edu.pl.

features (concepts, or classes) and making logical inferences based on the relationships between them, useful in mediating between different tagsets and tools" [1].

Conclusions

Morphosyntactic specifications for Ukrainian that were developed within the multilingual European MULTEXT-East project were presented concisely in this paper. Summarizing the advantages of the specifications that will hopefully be acknowledged by their future users, they:

- are freely available and well-documented;
- are detailed enough (1239 unique tags);
- are intuitive and relatively easy to remember for human users;
- follow the international standards;
- are consistent with 15 other languages, which enables a high degree of interoperability and use in multilingual projects such as machine translation, while preserving the common conceptual ground;
- have been approbated in corpora and disambiguation experiments;
- are stored in the popular for data exchange XML format;
- possess handy, and also freely available, XSLT and XSD tools for data conversion, rearrangement and validation;
- are supported by the only freely available tagger for Ukrainian (UGTag), allowing to encode textual data in standard XML-based corpus formats;
- are accompanied by an extensive, freely available morphosyntactic lexicon.

The importance of the resources from the multilingual perspective can be demonstrated by citing T. Erjavec, the leader of the MTE project: "The resources now cover most Slavic languages, which is esp. important as a) for a number of them, language resources are otherwise still hard to find and b) these languages have many common characteristics, i.e., they exhibit complex behaviour on the morphosyntactic level, and this is the first dataset that enables a qualitative and quantitative comparison between them." [3:4].

The author hopes that the Ukrainian MTE resources will open new perspectives for development of Ukrainian corpus and computational linguistics and its fast international integration.

1. Chiarcos, Christian and Tomaz Erjavec. OWL/DL formalization of the MULTEXT-East morphosyntactic specifications. / Proceedings of the 5th Linguistic Annotation Workshop (LAW-V), held in conjunction with the ACL—HLT 2011, June 2011 Portland, Oregon, USA — p. 11—20. 2. Derzhanski, Ivan and Natalia Kotsyba. Towards a Consistent Morphological Tagset for Slavic Languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian. / Proceedings of "Metalanguage and Encoding Scheme Design for Digital Lexicography: MONDILEX Third Open Workshop" Bratislava, Slovakia, 15—16 April 2009 — Bratislava — 2009. 3. Erjavec, Tomaž. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. / Proceedings of the LREC 2010, Malta, 19—21 May — 2010. — pp. 131—142. 4. Kotsyba, Natalia, Praktyczny przewodnik po korpusach języka ukraińskiego. / Praktyczny przewodnik po korpusach języków słowiańskich, ed. by Hebal-Jezierska M. — Warsaw — 2013, (forthcoming). 5. Kotsyba, Natalia, Andriy Mykulyak, Igor V. Shevchenko. UGTag: morphological analyzer and tagger for Ukrainian language. / Explorations across Languages and Corpora, Łódź Studies in Language, ed. by Goźdź-Roszkowski S. — 2011. 6. Kotsyba, Natalia, Adam Radziszewski and Ivan Derzhanski. Integrating the Polish language into the MULTEXT-East family: morphosyntactic specifications, converter, lexicon and corpus. / Proceedings of Research Infrastructure for Digital Lexicography: MONDILEX Fifth Open Workshop, October 14, 2009, Ljubljana, Slovenia. — Ljubljana — 2009. 7. Kotsyba, Natalia, Olha Shypnivska and Magdalena Turska. Linguistic principles of organizing a common morphological tagset for PolUKR (Polish-Ukrainian Parallel Corpus). / Proceedings of the international conference "Intelligent Information Systems, 16—18 June 2008, Zakopane, Poland". — Warsaw — 2008. 8. MULTEXT-East Morphosyntactic Specifications, Version 4. Ukrainian Specifications. / http://nl.ijs.si/ME/V4/msd/html/msd-uk.html 9. PolUKR (Polish-Ukrainian Parallel Corpus) / http://www.domeczek.pl/~polukr/index.php?option=welcome 10. Демська-Кульчицька

О. Основи національного корпусу української мови. — Київ — 2005. 11. Демська О. Текстовий корпус: ідея іншої форми. — Київ: ВПЦ НаУКМА, 2011. — 282 р. 12. Корпус української мови. / [*http://mova.info*](http://mova.info) *13. Коциба Н. Морфосинтаксичне тагування польсько-українського паралельного корпусу (PolUKR). / Proceedings of the International Conference "MegaLing'2008. Horizons of Applied Linguistics and Linguistic Technologies" Parthenit – Crimea, Ukraine, 20—27 September 2008 — Kyiv — 2009. 14. Перебийніс, Валентина і Тетяна Бобкова. Історія лабораторії комп'ютерної лінгвістики КНЛУ. Комп'ютерна лінгвістика: сучасне та майбутнє. Матеріали Міжнародної науково-практичної конференції — К.: КНЛУ, 2012. — 52 с. 15. Широков В.А та ін. Корпусна лінгвістика. — Київ — Довіра. 2005.*