Overview of the Ukrainian language resources within the multilingual European MULTEXT-East project, v.4

Natalia Kotsyba
faculty of „Artes Liberales”
University of Warsaw
29 May, 2013

Проект MULTEXT-East (MTE)

- Багатомовний і багаторічний міжнародний проект MULTEXT-East (MTE), версія 4
- http://nl.ijs.si/ME/V4
- Одна з небагатьох міжнародних ініціатив, спрямованих на інтегральний розвиток комп'ютерних ресурсів, котрі охоплюють також українську мову
- У відкритому доступі від травня 2010 року, безкоштовно для дослідницьких цілей

Морфосинтактичне кодування для української мови: теорія і практика

Kotsyba Natalia „Praktyczny przewodnik po korpusach języka ukraińskiego.”, розділ присвячений українським корпусам у колективній моногафії „Практичний довідник з українських корпусів” („Praktyczny przewodnik po korpusach języków słowiańskich”) під редакцією М. Хебаль-Єзерської (у друці, Варшава 2013)

Розвиток від концепції до реалізації:

- Чиста теорія:

Орися Демська-Кульчицька О. Основи національного корпусу української мови. (2005)

Демська О. Текстовий корпус: ідея іншої форми. (2011), включає опис MTE-2010

- Реалізована на практиці теорія:

Широков В. І ін. «Корпусна лінгвістика» 2005

383 синтетичні морфосинтактичні коди, описано в монографії,

Матеріальна база -- Український Граматичний Словник авторства Ігоря Вікторовича Шевченка 90-ті роки мин. століття

Тагсет створювався спеціально на потреби Корпусу УМІФ. На жаль, з огляду на незняту граматичну неоднозначність граматичні дані не є доступні для пошуку у корпусі

Морфосинтактичне кодування для української мови: теорія і практика

- Перевірка практикою:

Корпус Української Мови Лабораторії Компютерної Лінгвістики Київського Національного університету [mova.info]. Брак детального теоретичного опису. Омонімія знята автоматично, відсоток помилок досить високий

- Доступність граматичних даних: брак
- Специфічність:

Кожен з цих тагсетів, навіть ті, що існують тільки в теорії, створювалися для одного конкретного проекту – паралельно щонайменше три ініціативи створення національного корпусу української мови, реалізовані вони на різних рівнях

- Інформаційний вакуум

Матеріали створюються в багатьох місцях одночасно. Оскільки вони не потрапляють у вільний дослідницький обіг, інші дослідники змушені створювати все з нуля, що звичайно ж не дуже сприяє розвитку ресурсів і прогресу загалом.

MULTEXT-East

- Міжнародний проект, ціллю якого є однорідне, гармонізоване представлення ресурсів природніх мов на основі спільної концептуальної бази граматичних понять, яке робить можливим простий обмін мовними даними та їх повторне використання у різних проектах
- Перша версія MULTEXT, 1995 охоплювала шість західноєвропейських мов
- MULTEXT-East, 1998 додано шість центрально- і східноєвропейських мов
- Чергові версії (кер. проф. Томаж Ер’явец, Любляна) виходили під назвою MULTEXT-East, нові мови (зі слов’янських немає білоруської та лужицьких) і вдосконалення якості лінгвістичних даних

Загальний опис MTE
http://nl.ijs.si/ME/V4

- Ліцензія Creative Commons licence Attribution-ShareAlike 3.0
- **Морфоситактичні специфікації** для 16 мов у можна переглядати на сайті і/або завантажити.
- **Лексикон** (~ граматичний словник) можна завантажити після реєстрації і використовувати безкоштовно для дослідницьких цілей.
- Завдяки технологіям сім'ї **XML** (поичнаючи від версії 3) та однорідному кодуванню різні типи ресурсів (специфікації, лексикон, корпус) є добре інтегровані "making it possible to easily move between different representations of the same data" [3:1].
- **Корпус**: George Orwell „1984", не існує перекладу українською

Структура **специфікацій**

- Документ, згідний із вимогами TEI P5, що є визначенням атрибутів та їх вартостей, що використовуються у різних мовах для синтаксичного опису на рівні словоформи, тобто опис формальної граматики морфосинтаксичних властивостей мов, що ввійшли до проекту
- Вступ, спільна частина – опис властивостей, що є спільними для всіх мов, мовноспецифічні розділи

Спільна частина специфікацій

- 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”
- There are 12 main morphosyntactic categories which correspond to traditional linguistic parts of speech. 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. One of the categories, residual (R), is used for technical reasons, such as for unknown wordforms.
- While most categories are the same for the 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.

Мовноспецифічні розділи специфікацій

- 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
- Наступний слайд – фрагмент теоретичної частини специфікацій

-
- <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>

Матеріал для специфікацій

- 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.

Деякі специфічні для української граматики властивості 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*).

Деякі специфічні для української граматики властивості MTE

- 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, pro-nouns are a class of nouns. The Syntactic_Type shows further POS distribution.

Фрагмент індексу MSD для української

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

Що від української мови було додано до «загального котла»

- 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.

Комбінації атрибутів, які є легітимними тагами для української мови (іменники)

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

Технічні особливості проекту

- 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.

Лексикон

а		а	І	0
а		а	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	

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

Фрагмент уоднозначненого українського тексту в форматі XML з тагами MTE (корпус)

- <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>

Розповсюдження, використання та апробація ресурсів

- Українські МСС досить детально описані у монографії Орисі Демської-Кульчицької 2011 р., ст. 197—230
- Тагсет використовується єдиним вільнодоступним тагером для української мови UGTag
- Тексти, отаговані цією програмою, використовуються у Польсько-українському паралельному корпусі (Варшавський університет, проект 2005-2009 рр., керівник проекту Наталя Коциба), корпус є доступний для пошуку без потреби реєстрації на сайті http://www.domeczek.pl/~polukr/
- Використовується у корпусі текстів Івана Франка (Соломія Бук)
- Низка менших за розмірами студентських та аспірантських проектів, присвячених побудові порівняльних або паралельних корпусів (Женя Мудрак, Оля Предко, Іванка Кушнірук, Роксолана Перхач – Львівська Політехніка).
- Експеримент, проведений автором та доц. Андрієм Романюком – ручне зняття омонімії, роман «Депеш Мод» Сергія Жадана. Переваги XML -- можна було застосувати схему XSD, яка позволяла контролювати вписування тільки дозволених у специфікаціях тагів (далі питання natalia@al.uw.edu.pl)
- Конверсія до онтології OWL: “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 features (concepts, or classes) and making logical inferences based on the relationships between themєм , useful in mediating between different tagsets and tools” [TE].

Підсумки: ресурси MTE-4 для української мови

- Доступні безкоштовно та добре описані
- Досить детальні (1239 унікальних тагів)
- Інтуітивні та в міру мнемонічні
- Використовують міжнародні стандарти
- Узгоджені з іншими 15-ма мовами проекту, що робить можливим використання їх у багатомовних проектах, напр. автоматичний переклад, при збереженні спільного концептуального грунту
- Вже застосовуються у корпусах та експериментах дезамбігуації
- Мають популяний для збереження та обміну даних формат XML
- Додатково зручні та безкоштовні знаряддя XSLT і XSD для швидкої конверсії даних, представлення їх у інших форматах та валідації
- Використовуються єдиним вільнодоступним тагером для української мови (UGTag)

<w lemma="дякувати" disamb="1" ana="Vmpip1s">**Дякую**</w>

<w_>

<w lemma="за" disamb="**1**" ana="Q">**за**</w>

<w lemma="за" disamb="0" ana="Rp">**за**</w>

<w lemma="за" disamb="0" ana="Spsa">**за**</w>

<w lemma="за" disamb="0" ana="Spsi">**за**</w>

</w_>

<w lemma="увага" disamb="1" ana="Ncfsan">**увагу**</w>

<c>**!**</c>