Ivan Derzhanski, Natalia Kotsyba

Towards a consistent morphological tagset for Slavic languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian

Ivan Derzhanski, Natalia Kotsyba

Towards a consistent morphological tagset for Slavic languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian **(… as well as Upper and Lower Sorbian)**

The main idea

- Morphosyntactic annotation is needed by theoretical, computational and applied linguistics alike.
- Morphosyntactic annotation requires a **tagset** (ideally one consistent with linguistic theory and *then* with grammatical tradition).

The main idea *(continued)*

- Comparative theoretical studies, morphosyntactic annotation in parallel corpora, bi- and multilingual dictionary making all require a **common, crosslinguistically consistent tagset**.

… that is, a tagset which

- treats **like** phenomena in **like** ways,
- treats **unlike** phenomena in **unlike** ways,
- reflects the structural, etymological and semantic unity of grammatical categories to the greatest extent (especially in the case of closely related languages).

MULTEXT-East

11 tagsets developed in v.3, with 3 more added in v.4:

- Indo-European
  - Slavic
    - East (1): RU
    - West (2): CS, SK
    - South (4): SL, SL-R, HR, SR, BG, MK
  - non-Slavic (3): EN, RO, FA
- Uralic (2): HU, ET

MULTEXT-East: virtues

- Intended to be a multilanguage tagset from the beginning.
- Already *de facto* standard for several languages.

Thus a natural starting point for further work in this field.

MULTEXT-East:

same phenomenon, different treatment

- attributive participles
  - verb forms (BG, RU),
  - adjectives (CS, SK);
- adverbial participles
  - V, Vform=gerund (BG),
  - V, Vform=transgressive (CS, SK),
  - R, Type=verbal|causal (HU).

MULTEXT-East:

same phenomenon, different treatment

- virile (masculine human) forms of numerals
  - BG *двама*: Form=m_form
  - SK *dvaja*: Form=letter

MULTEXT-East:
similar phenomenon, different treatment

- short:long forms of adjectives
  - Formation=nominal:compound (CS),
  - Definiteness=no:yes (SL),
  - Definiteness=short_art:full_art (RU),
  - ?? (BG).

MULTEXT-East:
same term, different content

- M, Type=multipl[icative]
  - adverbial: *dvakrát* (CS),
  - adjectival: *dvojen* (SL)
- V, Definiteness: HU *vs* BG
- N, Case=genitive: FA *vs* everything else

MULTEXT-East:
language-specific solutions

- Clitic_s (CS)
- extra cases (RU)
  RU *цвет чая ~ чашка чаю,*
  шишка в снегу ~ вдохновение в снеге;
  UA *муха в меді ~ зварено на меду,*
  краснопера (individual) ~ *краснопery* (species);
  BE *пераезда* (place) ~ *пераезду* (act);
  CS *bratrovi ~ bratru Janovi*

That national grammatical traditions have often been followed is understandable.

But comparative work requires a common theoretical ground, the lack of which defeats the purpose of a common tagset.

So some traditional propositions will have to be sacrificed.

Moreover, traditional grammar can be inconsistent.

Bulgarian:

- *водата **ми*** ‘my water, το νερό **μου**’
- *дай **ми*** ‘give me, δώς **μου**’

Slovak:	2nd singular	reflexive
personal	*tebou*	*sebou*
possessive	*tvoj*	*svoj*

Priorities for a pan-Slavic tagset

- crosslinguistic consistency,
- linguistic adequacy,
- compactness.

Agglutination

A mechanism definitely needed for the **floating copula** (as well as other clitics) in Polish.

- ***powinniście*** *znać*
- ***słyszeliście***
- ***gdzieżeście*** *słyszeli …?*
- ***czybyście*** *uwierzyli …?*
- ***po coście*** *mnie tu przyniosły?*

Agglutination *(continued)*

- Will also do for:
  - Czech
    - floating *-s* < *jsi* (currently Clitic_s for verbs and pronouns only),
    - *aby*, *kdeby* + *-ch(om)*, *-s(te)* (currently inflecting particles);
  - Czech/Slovak *-že*;
  - adposition+pronoun compounds: PL ***przezeń***, HS ***tohodla***, etc.

Additional features needed

- N, V, A, P, M: Virile
  (for SK, PL, UA, BE, HS, DS, BG)
- N: CaseForm (first, second)
  (currently Case2=p|l)
- V: Agglutinativity, Vocalicity

Additional features needed

- A: Voice, Negation (for participles)
- A: Owner_Gender (for Sorbian)
  HS *stareje žoniny syn*
  DS *našogo nanowe crjeje*
- P: Post-prepositional (by any name)
  HS *jón ~ njón, što ~ čo*
  RU *ниже них ~ ниже их*
- S: Vocalicity

Additional values needed

- N|Type: gerund (for Polish at least)
  - Aspect, Negation
- N|Gender: common
- V|Aspect: biaspectual
- V|Person: inclusive (for Russian)
- A|Type: participle (etc.), pre-adjectival
- P|Person: reflexive

Conversion of existing formats for Polish and Ukrainian to an MTE-like format

Resources for morphological processing of Polish and Ukrainian have been developed independently from the project MTE in Poland (IPI—PAS corpus) and Ukraine (ULIF NASU corpus), respectively.

Morphological information is encoded in the form of grammatical dictionaries that allow for both analysing and synthesising word forms.

The granulation of grammatical information there and the formats of recording it differ considerably from the core MTE tagset.

Grammatical categories and values overlap (are one-to-one relations) only in part; some of them have to be decomposed into finer ones, and new categories/values need to be assigned to all relevant lexemes in a grammatical dictionary.

On the other hand, grammatical dictionaries contain information that is not necessary for MTE-like tagging.

Conversion of existing formats for Polish and Ukrainian to an MTE-like format

Two possible levels of introducing changes into Polish and Ukrainian grammatical sources: level of conversion of tagged texts, or directly in the dictionary source files.

Polish source files are not available for processing and development.

Ukrainian: additional grouping of lexemes is done within UGTag, morphological tagger with the possibility of adding new words from tagged texts, unrecognised by the tagger. One possible output format of UGTag will be an MTE-like tagged text.

Belarusian: a grammatical dictionary is under development now on the basis of an extensive orthographic dictionary; suggestions concerning its design and compatibility with MTE-like tagging format can be taken into account, no further conversion will be required.

Conversion of existing formats for Polish and Ukrainian to an MTE-like format

The tagsets for Polish (IPIC) and Ukrainian (UGD) were brought together within the PolUKR project with the aim of creating a common tagset for the parallel corpus of those languages.

The criterion of minimal information loss was used, although the common tagset is not a pure arithmetic sum of the two tagsets.

it was based on the pattern of IPIC, as it was easier this way to adjust the search program Poliqarp for the needs of PolUKR.

Since MTE-like tagging is becoming a standard now, it was decided to bring the PolUKR tagset to conformity with it.

Fragment of the conversion table IPIC/PolUKR → MTE v.3/4 (111 dictionary positions):

English term	PolUKR tag	MTE tag (fragment)	example
particle-adverb	qub	Q	*niech*
discourse markers	dsc	Q	*властиво*
infinitive	inf	V, VForm=n	*спатоньки*
impersonal form	imps	V, VForm=t	*rozpoczęto, robiono*
adverbial participle	part	V, VForm=r	
simultaneous adverbial participle	pcon	V, VForm=r, Tense=p	*роблячи, robiąc*
anterior adverbial participle	pant	V, VForm=r, Tense=a, Aspect=e	*зробивши, zrobiwszy*
simultaneous past participle	ppast	V, VForm=r, Tense=a, Aspect=p	*робивши, *robiwszy* (rare)
common (general) noun	gnoun	N, Type=c	*шахи*
proper name	propnoun	N, Type=p	*Сколе*
disparaging (depreciative) noun	depr	N, Animate=y, Human=n	*profesory*
1st- or 2nd-person pro-noun	ppron12	P, Type=p, Person=(1|2)	*я, ти*
gerund	ger	N, Type=g	*robienie, nierobienie niezrobienie*
3rd-person pro-noun	ppron3	P, Type=p, Person=3	*він, вони*

And a fragment of the correspondence table MTE v.3/4 → IPIC/PolUKR (332 positions):

category	attribute	value code	value name	IPIC/PolUKR equivalent
Adjective(A)	Aspect	E	perfective	(pact | pass)&aspect=perfective
Adjective(A)	Aspect	P	progressive	(pact | pass)&aspect=imperfective
Adjective(A)	Voice	A	active	pact&aspect=perfective
Adjective(A)	Voice	P	passive	pass&aspect=perfective
Adverb (R)		R		adv | adjp | pred
Verb(V)	VForm	I	indicative	fin | praet | bedzie
Verb(V)	Tense	P	present	fin&aspect=imperf
Verb(V)	Tense	F	future	bedzie | (fin&aspect=perf)

A fragment of the XML specification file for Ukrainian compatible with the MTE-4 proposal for Russian:

```
<row role="attribute">
  <cell xml:lang="en" role="position">6</cell>
  <cell role="name" xml:lang="en">Case2</cell>
  <cell xml:lang="en" role="values">
    <table>
      <row role="value">
        <cell role="name" xml:lang="en">genitive</cell>
        <cell role="code" xml:lang="en">g</cell>
      </row>
      <row role="value">
        <cell role="name" xml:lang="en">dative</cell>
        <cell role="code" xml:lang="en">d</cell>
      </row>
      <row role="value">
        <cell role="name" xml:lang="en">locative</cell>
        <cell role="code" xml:lang="en">l</cell>
      </row>
    </table>
  </cell>
</row>
```

The same fragment for Ukrainian according to our proposals:

```
<row role="attribute">
  <cell xml:lang="en" role="position">6</cell>
  <cell role="name" xml:lang="en">CaseForm</cell>
  <cell xml:lang="en" role="values">
    <table>
      <row role="value">
        <cell role="name" xml:lang="en">first</cell>
        <cell role="code" xml:lang="en">1</cell>
      </row>
      <row role="value">
        <cell role="name" xml:lang="en">second</cell>
        <cell role="code" xml:lang="en">2</cell>
      </row>
      <row role="value">
        <cell role="name" xml:lang="en">third</cell>
        <cell role="code" xml:lang="en">3</cell>
      </row>
    </table>
  </cell>
</row>
```

Conclusions and recommendations

General agreement on the tagset to be achieved among its developers; a common ground must be found.

In its current state the MTE tagset includes information from different levels of language description: purely morphological, derivational, syntactic and semantic.

Syntactic and semantic analysis and tagging are further necessary steps in language description, and principles of tagging for them should be developed.

The layer of derivation is significant for (semi)automatic lexicon development.

Information currently encoded about levels other than the morphological one (such as valency for prepositions or classification of pronoun types) should also be redistributed in the future.

Дякуємо за увагу

Благодарим за внимание

Дзякуем за ўвагу

Благодарим за вниманието

Dziękujemy za uwagę

Благодараме за вниманието

Dźakujemoj so za kedźbnosć

Bug lunej za to atencjun

Źěkujomej se za zajmowanosć

Mulțumim pentru atenție

Ďakujeme za pozornosť

Thank you for your attention

Děkujeme za pozornost

Täname tähelepanu eest

Zahvaliva se za pozornost

Köszönjük a figyelmet

Zahvaljujeme na pažnju

از دقتتان سپاسگزاریم