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

Ivan A Derzhanski[1] and Natalia Kotsyba[2]

[1] Institute for Mathematics and Informatics, Bulgarian Academy of Sciences
[2] Institute of Slavic Studies, Polish Academy of Sciences

Abstract. Comparative studies in theoretical linguistics and the production of bi- and multilingual dictionaries and tagged corpora, especially of closely related languages, can benefit from the use of a common, crosslinguistically consistent tagset which reflects the unity of grammatical categories to the greatest extent. As a case in point, the project MULTEXT-East developed tagsets for several Slavic languages and laid the foundations of the creation of a common Slavic tagset. Close scrutiny reveals, however, that it suffers from a number of inconsistencies and design flaws, which can have an adverse effect on its use in comparative work. In this paper we will suggest some amendments to MULTEXT-East v.3 (and v.4), and discuss what will have to be done in order for the remaining Slavic languages to be covered as well, with a focus on Polish, Ukrainian and Belarusian.

1 Introduction

Comparative studies in theoretical linguistics and the production of bi- and multilingual dictionaries and tagged corpora, particularly digital ones, can benefit from the use of a common, crosslinguistically consistent morphological tagset reflecting the structural, etymological and semantic unity of grammatical categories to the greatest extent. This is especially desirable in the case of closely related languages.

The project MULTEXT-East (MTE [3]) housed a classic endeavour to construct a foundation for creating tagsets for Eastern European languages (as well as one Western European language, namely English, which served as the hub language of the project). Version 3.0 covers 11 languages, with three more added in Version 4, to wit [4]:

- Indo-European:
    - Slavic:
        - East: (v. 4) RUSSIAN
        - West: CZECH, SLOVAK
        - South:
            - Western:
                - Slovenian: SLOVENE, RESIAN[1]
                - Serbo-Croat: CROAT, SERBIAN
            - Eastern: BULGARIAN, (v. 4) MACEDONIAN
    - non-Slavic: ENGLISH, ROUMANIAN, (v. 4) PERSIAN
- Uralic: ESTONIAN, HUNGARIAN

The seven Slavic tagsets in v.3 use 13 of the 14 parts of speech defined in the common tagset, with a total of 72 features and 263 values.

The project is generally acknowledged as having been very successful, and some of the MTE tagsets have become *de facto* standard for the respective languages. It is therefore a natural starting point for further

[*] The study and preparation of these results have received partial funding from the EC's 7th Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX.

[1] This is the Resian sub-dialect of the Slovene language of Bela/San Giorgio, Italy. Resian and standard Slovenian are mutually unintelligible due to archaisms preserved in Resian but not in contemporary Slovenian and to Italian-induced innovations in Resian grammar (including prepositive definite and indefinite articles).

work in this field.

Close scrutiny reveals, however, that the MTE system of tagsets for Slavic languages has a number of shortcomings which can have an adverse effect on its use in comparative work and its potential for extension to cover the remaining languages of the branch:

- On several occasions the same phenomenon in different languages is handled in different ways. For example, attributive participles are classified as verb forms in Bulgarian, but as adjectives in the other six Slavic languages in v.3, although there is no structural, semantic or etymological reason for such a discrepancy.[2] The four tagsets for Czech, Slovene, Russian and Bulgarian assume four different attitudes to the treatment of short and full forms of adjectives, where the actual semantic divergence might justify two.
- There are redundant values, such as 'transgressive' and 'gerund' (values of the feature VForm of the part of speech Verb), which refer to the same category, but the former is used in the tagsets for Czech and Slovak and the latter for Bulgarian and Serbian.
- Some terms are interpreted in unlike ways in different tagsets. Within the part of speech Numeral the type multipl[icativ]e is defined, but to the Czech tagset a multiple numeral is an adverbial one (*dvakrát* 'twice'), whereas to the Slovene tagset it is adjectival (*dvojen* 'double').
- Some solutions are not extensible. In Czech the 2nd person singular present tense form of the copula *jsi* can be cliticised as *-s* on certain non-finite verb forms and pronouns, and its presence is indicated by the positive value of the binary feature Clitic_s of the parts of speech Verb and Pronoun. Essentially the same phenomenon exists in Polish, but it involves four cliticised forms of the copula (1sg *-m*, 1pl *-śmy*, 2sg *-ś*, 2pl *-ście*), and they float more freely (the host can be any content word, e.g. *świniaś* 'thou art a pig', *dobryś* 'thou art good'), so the solution chosen in MTE for Czech can't be applied to Polish.

Excessively faithful adherence to grammatical tradition creates more awkwardness in the marking. This is especially conspicuous in the part of speech Pronoun. According to the traditional classification, personal and possessive pronouns are separate types, but reflexive pronouns are a single type. Thus in Czech *tobě* 'to thee' and *tvůj* 'thine' have different values of the feature Type (personal and possessive, respectively), whereas *sobě* 'to oneself' and *svůj* 'one's' are of the same Type (reflexive) and differentiated through the additional feature Referent_Type, although the relation is obviously the same in the two cases.

Some peculiarities can be explained by the need to keep the system compact because of the limitations of computing power a decade ago, a likely motivation for the designers to reuse the features as much as possible, even at the cost of linguistic adequacy. Now these concerns are no longer relevant.

In this paper we will examine MTE's treatment of the Slavic languages already covered and discuss what will have to be done in order for the rest of the branch, especially Polish, Ukrainian and Belarusian, to be treated as well.[3] In so doing we will focus on linguistic adequacy and crosslinguistic consistency, but will also aim for a concise tagset.

[2] Some of this is rooted in differences between national grammatical traditions. That they have often been followed is understandable, but comparative work requires a theoretical common ground, the lack of which defeats the purpose of a common tagset, so some traditional propositions will have to be sacrificed. (If the information is retained in whatever form, it will be a straightforward matter to convert it to the traditional form.) We are not aware of any post-MTE work aimed at bringing the various MTE tagsets closer to one another.

[3] We will not be concerned here with non-Slavic languages. Their coverage is particularly problematic, because so is the question of identifying matching grammatical categories when the languages aren't (closely) related. One of MTE v.3's most perplexing choices is that it uses the same binary feature Definiteness of the part of speech Verb to indicate, in Bulgarian, that a participle bears a definite article (*говорилите* 'the ones who talked'), and in Hungarian, that a finite form of a transitive verb has a definite 3rd person direct object (*tanulom* 'I learn it'). Thus two totally dissimilar (not to mention unrelated) phenomena are handled alike merely because their names in the respective grammatical traditions happen to mean the same. In MTE v.4 the tagset for Persian encodes izafet as Case=genitive (i.e., practically the opposite!) in an effort to avoid introducing a language-specific feature.

2 General remarks

The working definition that a word is a maximal uninterrupted sequence of letters stands in good stead most of the time, but there are several morphemes and clitics which form a graphic whole with their hosts in the standard orthographies (forms of the copula, the emphatic particle *-że* in Polish and *-ž* in Czech, prepositional markers of degrees of comparison), and some multi-word sequences might count as lexical units, but this technique should be used sparingly, and the matter relegated to syntax wherever possible.

2.1 Definiteness

Bulgarian has developed a synthetic definite article through the fusion of a form of a word belonging to one of the nominal parts of speech and a postpositive demonstrative pronoun. It is a peculiarity of the written norm that with singular masculine nouns ending in a consonant (as well as singular masculine forms of words of the other parts of speech) the article has two forms, full and short, originally stemming from different dialects but coexisting in the standard, being artificially assigned to different functions (according to the current norm, the full form is nominative and the short form oblique[4]).

The MTE tagset for Bulgarian maintains the feature Definiteness with the four values no (no article), yes (unique form of the definite article), full_art (full form of the definite article) and short_art (short form of the definite article). This makes it appear as though the distinction between the two forms of the article were on a par with its presence or absence. In fact these are features of different orders: the short and the full forms are varieties of the article, not its alternatives. We would propose two features, Article (no, yes) and DefForm (full, short).

Most Slavic languages (including Bulgarian) preserve the distinction between the full and the short form of the adjective, though typically only in a small part of the paradigm.[5] This can also be encoded through the feature DefForm (rather than Definiteness or Formation, as in MTE v.3 for the South Slavic languages and Czech respectively). The system would then look as follows:

Article	DefForm	Bulgarian (*як* m. ‘yak’)	Bulgarian (*яка* f. ‘collar’)	Bulgarian (*як* adj. ‘strong, sturdy’)	Ukrainian (*ярий* ‘violent’)
–	–				*ярий* (m.)
–	short				*яра* (f.); *ярі* (pl.)
–	full				*ярая* (f.); *ярії* (pl.)
no	–	*як; якове*	*яка; яки*	*яка* (f.)*; яки* (pl.)	
no	short			*як* (m.)	
no	full			*яки(й)* (m.)	
yes	–	*яковете*	*яката; яките*	*яката; яките*	
yes	short	*яка*		*якия*	
yes	full	*якът*		*якият*	

[4] Another norm existed during the rule of the Bulgarian Agrarian Popular Union (1921–23), when the choice of the full or short form of the article was based on euphonic rather than syntactic grounds (it depended on whether the following word began with a vowel or a consonant).

[5] In Serbo-Croat and Slovene the long forms are used as definite in all genders, numbers and cases, which justifies their encoding through a positive value of the feature Definiteness (or Article).
In Russian only the short nominative case forms are productive; they are used predicatively, as a general rule to express a temporary rather than permanent quality (*он весел* ‘he is in a cheerful mood’ vs *он весёлый* ‘he has a cheerful character’). However, short oblique case forms survive in numerous collocations (*среди бела дня* amidst white:GEN[SHORT] day:GEN ‘in broad daylight’). The situation is similar in Czech.
In Bulgarian only the masculine singular has a long form in *-и* (archaic *-ий*), used as a vocative (*драги съседе* ‘dear neighbour!’), appellative (*Петър Велики* ‘Peter the Great’), or (in archaic and poetic usage) definite (*равнините, набраздени с наший плуг* ‘the plains furrowed by our plough’). The MTE v.3 tagset for Bulgarian does not account for this form.
Ukrainian has lost the short masculine singular forms of all but 31 adjectives (an exhaustive list is given in [20]) and restricted the full feminine, neuter and plural forms to poetic speech.

In Macedonian the norm supports three forms of the article distinguished by distance, and in MTE v.4 they are encoded as values of Definiteness (proximal, yes, distal). Strictly speaking, they call for a separate feature, Distance (proximate, neutral, distal), since the presence of any article should be opposed to indefiniteness, but DefForm and Distance can be unified for practical convenience.

Article	DefForm	Distance	Bulgarian (*як* m. 'yak')	Macedonian (*јак* m. 'yak')
no	–	–	*як*	*јак*
yes	short	–	*яка*	
yes	full	–	*якът*	
yes	–	proximal		*јаков*
yes	–	neutral		*јакот*
yes	–	distal		*јакон*

2.2 Clitic_s

This feature is only defined for verbs and pronouns in Czech. As said before, it should be eliminated, because it is too specific, and can't be extended to the parallel phenomenon in Polish.

3 Noun

3.1 Type

Currently gerunds (deverbal nouns) are encoded as common nouns. Since they are very frequent in Polish, it seems expedient to add a type for them, with the additional features Aspect and Negation relevant only to gerunds. The latter would enable *celebrowanie* 'celebrating' and *niecelebrowanie* 'not celebrating' to count as forms of the same lexeme [15:46].

3.2 Class

Noun class in Slavic is an interplay of gender and animacy. All Slavic languages have the same system of three genders (masculine, feminine and neuter). In addition, inflexion and agreement often draw a line between live beings and everything else or between human beings and everything else. In Polish and Sorbian both distinctions are relevant (the former in the singular and – in Sorbian – the dual, the latter in the plural); many accounts of Polish grammar handle them by distinguishing three masculine genders (human, animal and inanimate), but this leads to massive syncretism, because in fact the differences only affect a few forms each, and is not readily extensible to other languages (in Russian, for example, animacy is orthogonal to gender in the plural). It seems more advantageous to maintain three features: Gender (m, f, n), Human (yes, no) and Animate (yes, no).[6] Here is how the forms of the Polish cardinal numerals '1' and '2' in all genders and cases can be encoded. Note especially the rows where either Human or Animate is neutralised, but not both.

[6] The idea of encoding the Slavic generalised gender category through a combination of gender and animacy features was also expressed in [13–14], though stipulating a feature with further subdivisions ('animacy' includes 'inhumanity' and 'humanity' with two values). In our proposal there are a total of four values, including the contradictory combination of 'human and inanimate', but this is a low price to pay for the simplification of the general feature structure of the tagset, and it actually saves rules: in [9] it is shown that the entire paradigm of the Polish demonstrative pronoun *ten* 'this' can be described by 34 rules in a five-gender system, but in ours only 31 are needed.

Gender	Human	Animate	Case	Polish
m	–	–	n	*jeden*
m	no	no	a	
m	–	yes	a	*jednego*
mn	–	–	g	
mn	–	–	d	*jednemu*
mn	–	–	i, l	*jednym*
n	–	–	n, a	*jedno*
f	–	–	n	*jedna*
f	–	–	a, i	*jedną*
f	–	–	g, d, l	*jednej*
m	yes	yes	n	*dwaj*
m	yes	yes	n, a	*dwóch, dwu*
–	–	–	g, l	
–	–	–	d	*dwom, dwu*
–	–	–	i	*dwoma*
m	no	–	n, a	*dwa*
n	–	–	n, a	
f	–	–	n, a	*dwie*
f	–	–	i	*dwiema*

In Polish some masculine human nouns are formally demoted to non-human to express derogation (*te*/**ci pijaki* 'these:NONHUM/*HUM drunkards'); these can be encoded as masculine animal.[7] With other nouns of the same class occasional conversion to the wrong class is used to express a certain attitude. Some authors have suggested introducing Disparagement as a formal feature of the noun [7]. This is unworkable, however, because which form is neutral and which is disparaging depends on the lexeme, and agreement is with humanness, not with disparagement (cf. neutral *ci profesorowie* 'these professors', *te chłopaki* 'these lads', disparaging *te profesory*, *ci chłopacy*).

A common gender is also expedient for words that can be masculine as well as feminine whilst retaining the same inflexion (Bulgarian *роднина* 'relative, kins[wo]man', Russian *сирота* 'orphan'). On the other hand, if a noun inflects in different ways (or not at all when feminine, as Polish *doktor* 'doctor'), this should be considered a pair of homonymous lemmata, with the homonymy resolved in the oblique cases.

3.3 Case

The original Slavic case system, preserved intact in most languages, contains seven cases (nominative, accusative, dative, genitive, instrumental, locative, vocative).

In Russian some nouns have two genitive or two locative forms with different meanings. Since these nouns are few, and the distinctions appear nowhere else in the grammar, introducing extra cases seems counterproductive. It is better to have an extra feature, CaseForm (first, second), whose value will select the correct subcase when needed, and be undefined most of the time.[8]

[7] When such a word is a subject, the predicate is masculine human (*Te pijaki przyszli* 'These:NONHUM drunkards came:HUMAN'). This is merely an instance of semantic agreement, which occurs in other Slavic languages also (Russian *Последний человек уволилась* 'The last:M person [= woman] resigned:F'), has an occasional character, and is outwith the scope of tagging.

[8] The proposed Russian tagset for MTE v.4 introduces the feature Case2 (p 'partitive', l 'locative'). This confines the choice to two possibilities with necessarily pre-defined cases, which is too restrictive, especially given that the locative in Ukrainian can even have three forms for the same word (*на водії, на водію, на водієві* 'on the driver'), cf. [19].

Case	CaseForm	Russian
n	–	*чай* ‘tea’, *молоко* ‘milk’, *снег* ‘snow’, *вода* ‘water’
g	–	*молока: цвет, чашка* ~ ‘the colour, a cup of milk’
g	first	*чая: цвет* ~ ‘the colour of tea’
g	second	*чаю: чашка* ~ ‘a cup of tea’
l	–	*воде: увидеть кольцо, красоту в* ~ ‘see beauty, a ring in the water’
l	first	*снеге: увидеть красоту в* ~ ‘see beauty in the snow’
l	second	*снегу: увидеть кольцо в* ~ ‘see a ring in the snow’

The same technique can be used for other instances of forms of the same case distinguished by usage, e.g.:

- the dative and locative singular of masculine nouns in Czech, which have the ending *-ovi* if the word is last in its phrase and *-u* otherwise (*bratrovi* ‘to the brother’, *bratru Janovi* ‘to Brother John’), and the similar alternation *-ові* ~ *-у* in Ukrainian, partly motivated by euphony (*панові Карпові Микитовичу Ковалеві* ‘to Mr Karp Mykytovych Kovalev’ [21:190]);
- the locative of monosyllabic Ukrainian nouns, where the ending *-у* tends to render a more specific meaning than *-і* (*муха в меді* ‘a fly is in the honey’, *зварено на меду* ‘cooked with honey’ [21:192]);
- the genitive of masculine nouns in Belarusian and Ukrainian, which has the ending *-а* for count nouns and *-у* for mass nouns, with some nouns assuming either depending on the interpretation (Bel.[9] *пераезда* ‘of the [place for] crossing’, *пераезду* ‘of the [act of] crossing’; Ukr. *краснопера* ‘of the [individual] redeye’, *красноперу* ‘of the redeye [as a species]’).

This phenomenon is not to be confused with variability in the use of case, which is not restricted to the noun form, e.g., accusative in Ukrainian: *пасти (чорні) бики*$_{\text{ACC=NOM}}$, *пасти (чорних) биків*$_{\text{ACC=GEN}}$ ‘herd (black) bulls’ or *писати (довгий) лист*$_{\text{ACC=NOM}}$, *писати (довгого) листа*$_{\text{ACC=GEN}}$ ‘write a (long) letter’.

Russian, Slovak, Slovene and Lower Sorbian have lost the vocative case except for a few fossilised forms (*боже*, *bože* ‘god!’), which may be encoded as vocative forms of the nouns, as can Russian colloquial vocatives formed by truncation (*мам* ‘mum!’, *Вань* ‘Vanya!’). Categorising concordant adjectives etc. as vocative case forms (as *môj* in Slovak *môj bože* ‘my god!’), however, appears superfluous.

3.4 Additional features

All Slavic languages have pluralia tantum nouns (Bulgarian, Russian *клещи* ‘pliers’), consequently the tagset needs a way of marking this, as they have some syntactic peculiarities, such as cooccurrence with collective numerals (Russian *двое часов* ‘two clocks’ vs *два часа* ‘two hours’). It might be possible to do this by an additional value of the feature Gender, but for those languages that don’t collapse all genders in the plural, gender features (possibly reduced[10]) for pluralia tantum nouns are also essential (Serbian *маказе* f. pl.t. ‘scissors’, *клешта* n. pl.t. ‘pliers’; Slovene *anali* m. pl.t. ‘annals’, *gosli* f. pl.t. ‘fiddle’, *vrata* n. pl.t. ‘door’), which means that a separate feature will be needed.

As said earlier, the features Aspect (imperfective, perfective) and Negation (no, yes) should be added at least for Polish, where gerunds are especially frequent and *nie-* ‘non-’ is productively prefixed to them.

[9] In Belarusian this is actually an innovation, an effect of the incursion of the Russian genitive ending *-а* into the language in the second third of the 20th century and its rivalry with the originally ubiquitous *-у*, although the ensuing opposition of count and mass nouns is different from the distribution of the two genitives in Russian ([18:53–54]).

[10] Or conventional: e.g., in the IPI—PAS corpus of Polish pluralia tantum nouns that are not masculine human (and thus are fully ambiguous between masculine non-human, neuter and feminine) are labelled as neuter.

4 Verb

4.1 Verb form

Verb forms include the following:

- Original finite forms, typically inflecting within each tense only for person and (verbal) number, although Upper Sorbian also distinguishes gender in the dual, Slovene does likewise (although the feminine/neuter forms are considered obsolete), and Resian has a distinction of courtesy in the 2nd person plural.

  The following three tables display forms of the verb ‘be’.

Person	Number	Gender	Human	Courtesy	Resian	Slovene	U Sorbian
1	dual	–	–	–	*swa*	*sva*	*smój*
1	dual	f, n	–	–		*sve	
2, 3	dual	–	–	–	*sta*	*sta*	*stej*
2, 3	dual	m	yes	–			*staj*
2, 3	dual	f, n	–	–		*ste	
2	plural	–	–	–		*ste	*sće*
2	plural	–	–	no	*sta*		
2	plural	–	–	yes	*stë*		

- Erstwhile perfect participles that are only used predicatively and have effectively become finite past-tense indicative forms. They only inflect for number and gender.

Number	Gender	Russian
singular	m	*был*
singular	f	*была*
singular	n	*было*
plural	–	*были*

- Past participles (termed pseudoparticiples in [15]) used mostly as complements of an occasionally omitted copula in analytic forms of perfect tenses, the conditional mood or the passive voice, inflecting for (nominal) number (including collective in Resian) and nominal class. These are encoded as VForm=participle.

Number	Gender	Human	Animate	Resian	Czech	Polish	U Sorbian
singular	m	–	–	*bil*	*byl*	*był*	*był*
singular	f	–	–	*bila*	*byla*	*była*	*była*
singular	n	–	–	*bilu*	*bylo*	*było*	*było*
dual	–	–	–				*byłoj*
dual	m	–	–	*bila*			
dual	f, n	–	–	*bili*			
plural	–	–	–				*byli*
plural	m	–	–	*bili*			
plural	m	–	yes		*byli*		
plural	m	yes	–			*byli*	
plural	m	–	no		*byly*		
plural	m	no	–			*były*	*byłe*
plural	f	–	–	*bile*	*byly*	*były*	*byłe*
plural	n	–	–	*bile*	*byla*	*były*	*byłe*
collective	m	–	–	*bile*			

- Adverbial participles (gerunds as they are called in MTE’s tagset for Bulgarian, or transgressives by the name used in the West Slavic tradition), uninflecting except in Czech, where they have retained number and gender: *nesa* (sg. m.), *nesouc* (sg. f./n.), *nesouce* (pl.) 'carrying'. These two values of the feature VForm should be unified; we would propose the label 'r' (because the part of speech Adverb is marked 'R').

- An invariable impersonal, originally an adverbial form of the past passive participle (in Polish, Ukrainian and Belarusian). For this we would propose the label 't', reminiscent of one of the suffixes.
- Finite forms of moods other than the indicative.
- Infinitive, invariable.[11]
- Supine, ditto (only in Slovenian, Resian and Lower Sorbian, though formerly in Czech as well).

Attributive participles, inflecting for number, gender and case or definiteness, are considered adjectives in several but not all tagsets in MTE. We believe this is right, and should be followed for all languages. The assumption that fully inflected participles are verb forms entails that the entire paradigm of the adjective is a proper part of the paradigm of the verb. This runs afoul of the proposition that the adjective and the verb are entities of the same order (parts of speech). Intuitively, too, Russian *читающего* 'reading:SG.M.GEN' is a form of the lemma *читающий* 'reading (present participle)', not of the lemma *читать* 'read'. And the argument (of a syntactic nature) that clause-forming participles have verbal government should not be considered relevant to morphological analysis.[12]

The tagset for Resian includes a subjunctive, but this category contains merely the 2nd person imperative forms, which are used as a subjunctive mood for all persons.

The tagsets for the other languages except Bulgarian include a conditional marker, inflecting for person and number in Czech and Serbo-Croat as in Polish and Upper Sorbian, uninflecting in Slovak, Slovene, Macedonian and Russian as in Ukrainian, Belarusian and Lower Sorbian.[13]

The IPI—PAS corpus of Polish (IPIC [7]) introduces a separate subcategory within the part of speech Verb for the so-called agglutinants, i.e., bound cliticised forms of the copula. The form *-s* of Czech *jsi* (2nd person singular form of the copula) calls for the same treatment.

VForm	Tense	Person	Number	Polish	Czech
indicative	present	1	singular	*jestem*	*jsem*
indicative	present	2	singular	*jesteś*	*jsi*
indicative	present	1	plural	*jesteśmy*	*jsme*
indicative	present	2	plural	*jesteście*	*jste*
bound	–	1	singular	*-m*	*-ch*
bound	–	2	singular	*-ś*	*-s*
bound	–	1	plural	*-śmy*	*-chom*
bound	–	2	plural	*-ście*	*-ste*

4.2 Aspect

Aspect is a category common to all Slavic languages, although not reflected in all tagsets in MTE. It would be desirable for the aspect called progressive to regain its usual name, imperfective. An ambivalent aspect might be more widely recognised (biaspectual verbs are numerous in Bulgarian, for example).

[11] The Bulgarian (truncated) infinitive has recently become obsolete, but can occur in texts: *недей казва* 'don't say', *можете ли каза* 'can you say' (now more commonly *недей да казваш*, *можете ли да кажете*).

[12] Neither is it consistently appealed to: Czech and Slovak attributive participles are clause-forming, but are encoded in MTE as qualificative adjectives; Bulgarian or Russian participles are no different.

[13] The Bulgarian conditional *бих*, *би* etc. are encoded in MTE as aorist tense forms of the verb *бъда* – a perfective counterpart of the imperfective copula *съм* –, although the forms *бидох*, *биде* etc. are better candidates for such encoding; in the contemporary language *бих*, *би* have no perceivable relation to the aorist.

4.3 Tense

MTE v.3 supports present, future, past, aorist, imperfect and pluperfect. The undifferentiated past tense is based on participles in the East Slavic languages or on the collapse of the aorist of perfective verbs and the imperfect of imperfective verbs into a single so-called preterite tense in Sorbian (a pronounced tendency in Macedonian as well).

Aspect	Tense	Person	Number	Gender	Bulgarian	Russian	U Sorbian
imperfective	imperfect	2, 3	singular	–	*ядеше*		
imperfective	past	2, 3	singular	–			*jědźeše*
imperfective	past	–	singular	masculine		*ел*	
imperfective	aorist	2, 3	singular	–	*яде*		
perfective	imperfect	2, 3	singular	–	*изядеше*		
perfective	past	–	singular	masculine		*съел*	
perfective	past	2, 3	singular	–			*zjě*
perfective	aorist	2, 3	singular	–	*изяде*		

The pluperfect is only introduced in the tagsets for Croat and Serbian, for no evident reason, as no Slavic language has a synthetic pluperfect.

4.4 Other features

Many (though not all) Russian verbs have a 1st person plural inclusive, formally present tense, form with hortative semantics: *идёмте* (imperfective), *пойдёмте* (perfective) ‘let us (you:PL and I) go’. This could be encoded as a 1st person plural form of a special mood (verb form, e.g. 2nd imperative, as in the National Corpus of the Russian Language); however, structurally it is not the mood but the person (a combination of *-м* ‘1st pl.’ and *-те* ‘2nd pl.’) that makes it exceptional. Such a form should either have a special value (inclusive) of the feature Person or be treated as an agglutinative compound of a 1st person plural verb form and the bound particle *-те* (also found in *нате* ‘here you are!’, *нуте* ‘well!’ with an addressee for whom the 2nd person plural is used).

For Polish the feature Vocalicity (voc, nvoc) has been added in IPIC to separate the cliticised forms of the copula with a buffering vowel (*-em*, *-eś*) or without one (*-m*, *-ś*).

IPIC also introduces the feature Agglutinativity (agl, nagl) for accounting for some problems of wordhood [15].[14] It has a positive value for past tense forms of verbs (pseudoparticiples) that require a bound clitic (*gniotł-em* ‘I kneaded’) and a negative one for their self-sufficient counterparts (*gniótł* ‘he kneaded’). The same technique might be used for Czech singular imperatives which have a bound form before the particle *-ž* (*bud’* ‘be!’, but *budi-ž* ‘be thou now’).

5 Adjective

5.1 Type

MTE v.3 recognises adjectives of three types: qualificative, possessive and ordinal (actually relative, a mistranslation of the Slovenian term *vrstni*). All attributive participles in all languages except Bulgarian are categorised as qualificative adjectives, ignoring voice and tense. However, it would be desirable to preserve this information by introducing a new type of adjective, participle, and voice, tense and aspect as features relevant only to participles. The table below displays the Bulgarian adjective *дъвчащ* ‘chewing (of sweets)’ as well as all participles formed from the verb *дъвча* ‘chew’:

[14] In the formalism used in the IPIC tagset [7] binary features typically have values of the type (‹value›, n‹value›); in MTE’s notation these can always be rendered as (yes, no).

PoS	Type	Aspect	Tense	Voice	Bulgarian
Adjective	qualificative	–	–	–	*дъвчащ*
Adjective	participle	imperfective	present	active	*дъвчещ*
Adjective	participle	imperfective	aorist	active	*дъвкал*
Adjective	participle	imperfective	aorist	passive	*дъвкан*
PoS	VForm	Aspect	Tense	Voice	Bulgarian
Verb	participle	imperfective	imperfect	active	*дъвчел*

Furthermore, since exclusively predicative adjectives (e.g., Slovak *dlžen* 'obliged') are treated as regular adjectives, predicative participles (including such as are used as past tense forms of verbs, alone or with conjugated forms of a copula) should be too.

It would be advantageous to also move ordinal (and other adjective-like) numerals and some types of pronouns to the part of speech Adjective, again distinguishing them by type, so as to relieve the other parts of speech of the strictly adjectival features.[15]

Type	Czech
qualificative	*dobrý* 'good'
possessive	*matčin* 'mother's'
ordinal numeral	*pátý* 'fifth'
specific numeral	*dvojí* 'double, twofold'

IPIC distinguishes two further types of adjectives: preadjectival (the first halves of compounds such as *biało-czerwony* 'white-and-red') and postprepositional (the content words in expressions of the type *po polsku* 'in Polish', only used following the preposition *po*). The former is advisable since it would be impractical to provide all compounds in the dictionary; the latter are better classified as adverbs.

5.2 Degree

Degree (positive, comparative and superlative[16]) is defined for all Slavic languages except Bulgarian, where it has been decreed that the degree markers *по-* (comparative) and *най-* (superlative), both linked to the adjective or adverb by a hyphen in the current orthography, might better be treated as separate words (Particles of type comparative). While fully functional, this decision separates the Bulgarian superlative *най-* from its counterparts in the other languages (*nej-* in Czech, *naj-* elsewhere, all prefixed to the comparative form and written as one word); then again, this may be justified by the fact that in Bulgarian both degree markers can also be used with other parts of speech and expressions, although then separated by a space in writing (*пò юнак* 'more of a hero', *най ми е жал* 'I regret most'). In Macedonian the same markers are written as a solid word together with the adjective or adverb (*подолг* 'longer', *најмногу* 'most'), and MTE v.4 treats the whole as a form inflected for degree.

In the Ukrainian Grammatical Dictionary [20], the source of morphological information for Ukrainian, degree was disposed of, comparative and superlative adjectives and adverbs are recorded as separate lexemes with corresponding lemmata. Rules for extracting information on degree and redirecting non-positive units to their lemma were designed and implemented in the project UGTag [6], enabling information on degree to be encoded for Ukrainian.

4.3 Additional features

The feature Negation (no, yes) should be added at least for Polish with its regularly formed participles.

[15] Some national traditions actually call for this: 'Numerals in Slovene can function as nouns, adjectives or adverbs, and are in grammars described as subtypes of these categories. The above classification runs counter to the established practice and is missing an important syntactic distinction' [4:205].

[16] Also elative for Slovene, Resian and Serbian and diminutive for Resian, though no examples are provided.

For Sorbian the feature Owner_Gender would have to be borrowed from the part of speech Pronoun, to encode the gender of the noun from which a possessive adjective is derived, as such a noun can have concordant modifiers (Upper Sorbian *stareje žoniny syn* ‘the old woman’s son’, Lower Sorbian *našogo nanowe crjeje* ‘our father’s shoes’ [8]).

PoS	Type	Owner_Gender	Gender	Number	Case	Upper Sorbian
Adjective	qualificative	–	feminine	singular	genitive	*stareje*
Adjective	possessive	feminine	masculine	singular	nominative	*žoniny*
Noun	common	–	masculine	singular	nominative	*syn*

6 Pronoun

6.1 Type

Traditional Slavic grammars acknowledge nine types of pronouns (personal, possessive, reflexive, demonstrative, interrogative, relative, indefinite, negative and general). The system is partly inconsistent: some pairs of pronouns of the same type (both reflexive, interrogative, etc.) stand in the same relation with one another as a personal and a possessive pronoun, and many pronouns fit the criteria for membership in more than one class (Ukrainian *свій* ‘one’s [own]’ could be classified as both reflexive and possessive, *хтозна-чий* ‘who knows whose’ as indefinite and possessive, *хтозна-який* ‘heaven knows what kind of’ as indefinite and demonstrative, etc.).

It appears that personal and possessive pronouns can be conflated (because there have to be other means for handling this kind of opposition anyway, as between ‘who’ and ‘whose’), and reflexive pronouns can be unified with them (as a special value of Person[17]).

MTE v.3				Our proposal		
Type	Person	Referent_type	Czech	Type	Person	Referent_type
p	2	–	*tobě*	p	2	p
s	2	–	*tvůj*	p	2	s
x	–	p	*sobě*	p	x	p
x	–	s	*svůj*	p	x	s
q	–	(p)	*kdo*	q	–	p
q	–	(s)	*čí*	q	–	s

In general these features refer to the meaning of pronouns and should be dealt with at the level of semantics. The developers of UGD [20] divide traditional pronouns into pro-nouns and pro-adjectives (pro-adverbs, too, in Russian National Corpus project); the designers of IPIC [7] refer to pro-adjectives as ordinary adjectives, while pro-nouns are singled out as a class. We would favour encoding pro-adjectives as several types of adjectives and preserving pro-nouns as a separate class.

6.2 Referent_Type and Syntactic_Type

These two features appear redundant, as a personal (possessive) value of Referent_Type correlates with a nominal (adjectival) value of Syntactic_Type.

The Bulgarian tagset doesn’t use Syntactic_Type at all, but employs two unique values of Referent_Type: attributive and quantitative. The first of these allows distinguishing, e.g., attributive *какъв* ‘what kind of’ from possessive *чий* ‘whose’. The words categorised as quantitative pronouns (*колко* ‘how many/much’, *няколко* ‘several’, *толкова* ‘this many/much’) correspond to numerals distinguished by values of the feature Class (interrogative, indefinite, demonstrative) in Czech and Slovak, and the Slovene and Resian tagsets don’t identify them in any way. The choice seems to be a matter of economy. Handling these

[17] This would not work, obviously, if English with its person-marked reflexives were restored to the system.

words as pronouns takes advantage of the numerous types of pronouns already defined, and treating them as numerals facilitates their classification by type of numeral (e.g., Czech cardinal *kolik* ‘how many’, ordinal *kolikátý* ‘number what’, multiplicative *kolikrát* ‘how many times’; Bulgarian has fewer such types, but it needs a way of distinguishing *колцина* ‘how many [people]’ from *колко* ‘how many/much’, although MTE v.3 provides none).

6.2 Additional features

In all East and West Slavic languages personal pronouns of the 3rd person have forms starting with /n/ instead of /j/, typically employed when the pronouns are objects of prepositions. For this phenomenon IPIC uses the feature Postprepositionality (praep, npraep), a practice which should be emulated. Also, in Upper Sorbian the pronoun *što* ‘what?’ has the same form in the accusative except after a preposition, where *čo* substitutes; this can be encoded in the same way.

Type	Gender	Human	Number	Case	Postprep	Upper Sorbian
personal	masculine	no	singular	accusative	no	*jón*
					yes	*njón*
interrogative	neuter	no	singular	accusative	no	*što*
					yes	*čo*

It should be noted, however, that the condition of the use of these forms vary somewhat across languages: in Russian they are optionally used after comparative degree forms (*ниже них* ~ *ниже их* ‘below them, lower than they’), in Ukrainian the conditions depend on the dialect. For this reason it may be advisable to give the feature a less binding name (one motivated by the form rather than the function).

7 Numeral

7.1 Type and Form

All languages distinguish cardinal and ordinal numerals; also, in MTE v.3 collect[ive]s are introduced for Serbian, and multipl[icativ]es and special[18] numerals for all seven languages except Resian and Bulgarian. On the whole the systems of numerals are made to look more different than most of them actually are.

The Bulgarian masculine personal numerals are handled as Type=cardinal Form=m_form in MTE v.3. In a common tagset this language-specific value would be superfluous, thanks to the feature Human.

Gender	Human	Bulgarian	
m	yes	*двама*	‘2’
m	no	*два*	
fn	–	*две*	

7.2 Class

For Polish the feature Accomodability (congr ‘agreeing’, rec ‘governing’) has been added in IPIC to identify the structural relation between the cardinal numeral and the noun (attribute–head or head–complement, respectively): *Przyszli dwaj chłopcy* ‘Two:CONGR boys:PL.NOM came:PL.HUM’, *Przyszło dwóch/dwu chłopców* ‘Two:REC boys:PL.GEN came:SG.N’. This can be encoded here through the feature Class, introduced in MTE v.3 in order to account for the different syntactic distribution of the cardinal numerals (esp. in Czech):

[18] Or specific, denoting a number of kinds of substances.

Gender	Human	Class	Polish	
m	yes	definite	*dwóch*, *dwu*	‘2’
		definite2	*dwaj*	
m	no	definite2	*dwa*	
n	–	definite2		
f	–	definite2	*dwie*	
m	yes	definite	*trzech*	‘3’
		definite34	*trzej*	
m	no	definite34	*trzy*	
f, n	–	definite34		
m	yes	definite	*pięciu*	‘5’
m	no	definite	*pięć*	
f, n	–	definite		

8 Adposition

8.1 Type

Slavic languages tend to only have prepositions. In Russian a few prepositions (*вопреки* ‘contrary to, notwithstanding’, *назло* ‘to spite’, *ради* ‘for the sake of’, *спустя* ‘after, later’) can be used postpositively; Sorbian *dla* ‘because of’ is more often a postposition than a preposition (Upper Sorbian *špatneho wjedra dla ~ dla špatneho wjedra* ‘because of the bad weather’; Lower Sorbian *chórosći dla ~ dla chórosći* ‘due to illness’, cf. German *krankheitshalber*). These should be undefined as to Type.

8.2 Case

In linguistic theory an adposition’s subcategorisation of an object in a certain case is no different from the subcategorisation of a verb. Tagsets don’t usually encode transitivity features for verbs, so introducing such a feature for prepositions amounts to an inconsistency. In practice, too, since in Slavic languages many prepositions can govern more than one case, the case syncretism common in nouns entails massive ambiguity in the tagging of prepositions.

We contend that no such feature ought to have been introduced into the morphological tagset. We would keep it only for the reason that its use is a widespread practice.

8.3 Additional features

Typically the object of a preposition, if a pronoun, must be a full (stressed) form. But there are exceptions. In Bulgarian the object of a few prepositions can be expressed as a dative (possessive) clitic[19] as well as a full accusative form (*помежду им* or *помежду тях* ‘between them’, but only *между тях* dto.). In Upper Sorbian the 1st person singular pronoun appears as a clitic after polysyllabic prepositions (*přećiwo mi* ‘against me’, *pola mje* ‘by me’, but *ku mni* ‘towards me’, *za mnje* ‘for me’). These peculiarities of the prepositions can be encoded by an additional feature.

It would be advisable to borrow the binary feature Vocalicity from the part of speech Verb for extended forms of prepositions (Bulgarian *във ~ в* ‘in’, Russian *передо ~ перед* ‘before’, Polish *ku ~ k* ‘towards’, Upper Sorbian *wote ~ wot* ‘from’, etc.), used in specific (morpho)phonological conditions.

[19] The MTE tagset for Bulgarian marks the short dative forms of the pronouns (*ми* ‘to me’, …, *им* ‘to them’) doubly as Type=personal Case=dative and Type=possessive, which is in conformity with the traditional descriptions, but redundant (especially since the use of a dative clitic as an adnominal possessive marker in Bulgarian is not an accident, but an areal feature shared with other languages of the Balkans).

In several languages adpositions optionally merge with some pronouns, yielding such compounds as Czech *zaň ~ za něho* ‘for him’, *proč ~ pro co* ‘for what’, Slovene *zate ~ za tebe* ‘for thee’, Polish *przezeń ~ przez niego* ‘because of him’, Upper Sorbian *mojedla ~ dla mnje* ‘because of me’, Lower Sorbian *mójogodla ~ dla mnjo* dto. (cf. German *meinetwegen*). It is best to treat these as agglutinative compounds, so as not to lose information about either the adposition or the pronoun.

9 Conjunction

Forms such as Czech *abych* ‘that I would’, *kdybyste* ‘if you would’ might also be treated as compounds (following the path suggested by their Polish counterparts *abym*, *gdybyście*) rather than as conjunctions inflected for person and number as in the MTE v.3 Czech tagset. (Conjunctions are, after all, supposed to be an invariable part of speech.) This would make for greater consistency across languages.

10 Predicative

Uninflecting words (and some collocations) which are restricted to being complements of copulative verbs are recognised as a separate part of speech in several reference grammars and tagsets of various Slavic languages. This appears superfluous: as we argued in [2], such items are adverbs no less than predicative adjectives (English *glad*, Russian *рад* dto.) are adjectives. However, attributivity/predicativity may be introduced as an additional feature for the purposes of syntactic analysis.

11 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 and Ukraine, 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. There are two possible levels of introducing changes into Polish and Ukrainian grammatical sources. This can be done at the level of conversion of tagged texts, or directly in the dictionary source files. The former option is chosen for Polish, since the source files are not available for processing and development. The latter option has been chosen for Ukrainian, and additional grouping of lexemes is done within UGTag [6], which foresees the creation of a morphological tagger for Ukrainian 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.

As for Belarusian, a grammatical dictionary for it is under development now on the basis of an extensive orthographic dictionary [11], and suggestions concerning its design and compatibility with MTE-like tagging format can be taken into account, so that no further conversion will be required.

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

Here is a fragment of the conversion table IPIC/PolUKR → MTE v.3/4 (111 dictionary positions):

Ukrainian term	Polish term	English term	PolUKR tag	MTE tag (fragment)	example
частка-вигук	partykuło-przysłówek	particle-adverb	qub	Q	*niech*
вставні слова	dyskursyw	discourse markers	dsc	Q	*властиво*
інфінітив	bezokolicznik	infinitive	inf	V, VForm=n	*спатоньки*
безособова форма	forma -no/-to	impersonal form	imps	V, VForm=t	*rozpoczęto, robiono*
дієприслівник	imiesłów przysłówkowy	adverbial participle	part	V, VForm=r	
недоконаний дієприслівник	imiesłów przysłówkowy współczesny	simultaneous adverbial participle	pcon	V, VForm=r, Tense=p	*роблячи, robiąc*
доконаний дієприслівник	imiesłów przysłówkowy uprzedni	anterior adverbial participle	pant	V, VForm=r, Tense=a, Aspect=e	*зробивши, zrobiwszy*
дієприслівник минулого часу	imiesłów czasu przeszłego	simultaneous past participle	ppast	V, VForm=r, Tense=a, Aspect=p	*робивши, *robiwszy* (rare)
загальний	ogólny	common (general) noun	gnoun	N, Type=c	*шахи*
власна назва	nazwa własna	proper name	propnoun	N, Type=p	*Сколе*
пейоративний іменник	rzeczownik deprecjatywny	disparaging (depreciative) noun	depr	N, Animate=y, Human=n	*profesory*
займенник-іменник 1-2 особа	zaimek 1-2 osoba	1st- or 2nd-person pro-noun	ppron12	P, Type=p, Person=(1|2)	*я, ти*
герундій	gerundium	gerund	ger	N, Type=g	*robienie, nierobienie niezrobienie*
займенник-іменник 3 особа	zaimek 3 osoba	3rd-person pro-noun	ppron3	P, Type=p, Person=3	*він, вони*
займенник себе	zaimek siebie	pronoun ‘self’	siebie	P, Type=x	*себе*

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)

Two sets of XML morphosyntactic specification files for Polish and Ukrainian have been prepared: specifications compatible with the most recent, still unreleased version of MTE (v.4), also based on [10][20], and specifications following from the suggestions formulated in this article.

[20] We would like to express our gratitude to Tomaž Erjavec for his advice and especially for directing us to the archives of the mailing list for MTE—Russian, which proved a valuable resource for our work on the XML specifications.

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>
```

12 Conclusions and recommendations

We realise that the suggested modifications entail a need of modifying, or even retagging, corresponding text files in various MTE languages. This should be undertaken only after general agreement on the tagset is achieved among its developers. We do hope that the proposed changes will evoke a wide discussion, and that a common ground will eventually 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. This is why the currently encoded information about levels other than the morphological one (such as valency for prepositions or classification of pronoun types) should also be redistributed in the future.

Bibliography

[1] Broda B., Piasecki M. and Radziszewski A. (2008). Towards a Set of General Purpose Morphosyntactic Tools for Polish. *Proceedings of Intelligent Information Systems, Zakopane, Poland, 2008*. Institute of Computer Science—PAS.

[2] Derzhanski I. and Kotsyba N. (2008). The category of predicatives in the light of the consistent morphosyntactic tagging of Slavic languages. In *Lexicographic Tools and Techniques: Proceedings of the MONDILEX First Open Workshop*, pages 68–79, Moscow: IITP—RAS.

[3] Dimitrova L., Erjavec T., Ide N., Kaalep H.-J., Petkevič V., Tufiş D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *Proceedings of COLING—ACL '98*, pages 315–319, Montréal, Québec, Canada.

[4] Erjavec, T. (ed.) (2004). *MULTEXT-East Morphosyntactic Specifications: Version 3.0*. Ljubljana.

[5] Kotsyba N., Shypnivska O. and Turska M. (2008). Linguistic principles of organizing a common morphological tagset for PolUKR (Polish-Ukrainian Parallel Corpus). In *Proceedings of Intelligent Information Systems, Zakopane, Poland, 2008*. Institute of Computer Science—PAS.

[6] Kotsyba N., Mykulyak A., Shevchenko I. (to appear). UGTag: morphological analyzer and tagger for Ukrainian language.

[7] Przepiórkowski A. and Woliński M. (2003). A Flexemic Tagset for Polish. In *Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003.*

[8] Sadock, J. (1985). Autolexical syntax: A proposal for the treatment of noun incorporation and similar phenomena. Natural Language and Linguistic Theory, 3, 379–439.

[9] Sauvet G., Włodarczyk A. and Włodarczyk H. (2007). Morphological data exploration using the SEMANA platform: Feature granularity problem in the definition of Polish gender. Lecture slides: ‹http://www.celta.paris-sorbonne.fr/anasem/papers/miscelanea/PolishGender.pps›.

[10] Sharoff S., Kopotev M., Erjavec T., Feldman A., and Divjak D. (2008). Designing and evaluating a Russian tagset. In *Sixth International Conference on Language Resources and Evaluation, LREC'08*, Paris, ELRA.

[11] Shevchenko I., Kotsyba N., Kurshuk K. (to appear). Towards the Creation of a Belarusian Grammatical Dictionary.

[12] Turska M. and Kotsyba N. (2007). Polish-Ukrainian Parallel Corpus and its Possible Applications. In *Proceedings of the International Conference 'Practical Applications in Language and Computers', 7–9 April 2005, Łódź*. Peter Lang GmbH.

[13] Włodarczyk, H. (2007). Relewantność cech HUM, ANIM i LOC w gramatyce języka polskiego. Presentation at *The 4th CASK Initiative—Workshop at the Jagiellonian University*, 17–21 April 2007.

[14] Włodarczyk, H. (2008). Pierwsze studium przypadku: problem ziarnistości definicji rodzaju w języku polskim. Presentation at the Institute for Slavic Studies—Polish Academy of Sciences, 14 April 2008.

[15] Woliński, M. (2004). System znaczników morfosyntaktycznych w korpusie IPI PAN. Polonica XII, 39–54.

[16] Бірала А. Я., Булахаў М. Г., Жыдовіч М. А., Жураўскі А. І., Карнеева-Петрулан М. І., Крыўчык В. Ф., Лапаў Б. С., Мацкевіч Ю. Ф. (1957). *Нарысы па гісторыі беларускай мовы. Дапаможнік для студэнтаў вышэйшых навучальных устаноў*. Мінск.

[17] Ломтев, Т. П. (1956). *Грамматика белорусского языка*. Минск.

[18] Сцяцко, П. (2002). *Культура мовы*. Мінск: Тэхналогія.

[19] Шевченко, І. В. (1996). Алгоритмічна словозмінна класифікація української лексики. Мовознавство №4–5, 40–44.

[20] Шевченко И. В., Широков В. А., Рабулець А. Г. (2005). Электронный грамматический словарь украинского языка. In *Труды международной конференции «Megaling'2005. Прикладная лингвистика в поиске новых путей», 27 июня–2 июля 2005 года, Меганом, Крым, Украина*, pages 124–129.

[21] Шерех, Ю. (1951). *Нарис сучасної української літературної мови*. Мюнхен: «Молоде життя».