Integrating the Polish language into the MULTEXT-East family: morphosyntactic specifications, converter, lexicon and corpus

Natalia Kotsyba[1], Adam Radziszewski[2], Ivan Derzhanski[3]

[1] Institute of Interdisciplinary Studies, Warsaw University
[2] Institute of Informatics, Wrocław University of Technology
[3] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

Abstract. In this article we discuss the theoretical background, the resources employed and the process of integrating the Polish language into MULTEXT-East (version 4) including: 1) specifying a MTE-compliant tagset for it with an indication of the restrictions on combinations of attributes; 2) creating, or rather converting, a representative lexicon of word forms with tags; 3) tagging a sample text using the prepared resources.

1 Introduction

The Polish language forms, together with Kashubian and Silesian, the Lechitic subgroup of the Western group of the Slavic branch of the Indo-European language family [Ethnologue 2009]. In terms of grammar, it is a typical Slavic language. It shares with several other Slavic languages (Slovak, Upper and Lower Sorbian) a complex category of noun class including three varieties of the masculine gender (human, animal, and thing), with a peculiar subvariety of depreciative (derogative) nouns. The most unusual feature of Polish is the cliticised present tense forms of the copula along with the newly formed synthetic past tense and conditional mood of the verb, which use the cliticised copula as a subject marker.

There is no single generally accepted standard for encoding Polish corpora. The most widely used tagset for Polish is that of the Institute of Computer Science's Corpus (IPIC, http://korpus.pl). Other standards exist, however, such as the ones used in the PELCRA Corpus of Polish (http://korpus.ia.uni.lodz.pl/) or the PWN Corpus of Polish (http://korpus.pwn.pl/). In the National Corpus of Polish, which is currently being compiled by a consortium consisting of the contributors above (http://nkjp.pl/), it is anticipated that a tagset which is slightly different from IPIC [Przepiórkowski, 2009] will be employed.

The multiplicity of encoding systems makes it difficult to match existing resources for Polish and hinders the reuse of resources available for other languages and the interoperability between processing tools. Mapping them on existing international recommendations like MULTEXT could facilitate the situation.

The MULTEXT-EAST project (MTE, http://nl.ijs.si/ME) produced a family of morphosyntactic tagsets for various languages (primarily of Central and Eastern Europe) based on a common formalism. With the addition of Russian in 2008 [Sharoff et al., 2008] it already covers 13 languages. As it expands, it is becoming more and more diversified, from the point of view of both language typology and linguistic description. The former direction of diversification has objective reasons, the latter is due to the differences between the traditions of grammatical description in the various countries. An attempt to analyse the representation of the hitherto encoded Slavic languages in MTE and the possibility of its extension to Polish was made in [Derzhanski, Kotsyba 2009]. A number of discrepancies were identified, mostly resulting from the inconsistent use of terminology in the description of phenomena found in more than one language. What is more, some solutions already applied in MTE appear not to be extensible.

In this article we discuss the theoretical background, the resources employed and the process of integrating the Polish language into MTE including: 1) specifying a MTE-compliant tagset for it with an indication of the restrictions on combinations of attributes; 2) creating, or rather converting, a representative lexicon consisting of word forms with tags; 3) tagging a sample text basing on the prepared resources.

2 Design of the tagset

In this section the particularities of the MTE morphosyntactic specifications for Polish are explained, the new categories and their attributes with values are presented.

General considerations

Morphological tagging means endowing every word in a text with a tag identifying its grammatical form and the lemma (citation form). The grammatical form includes classificatory, inflectional and occasionally subcategorisation features.

Generally speaking, a word in this context means a graphical unit. Some special cases call for special attention: clitics that can't be conveniently treated as affixes but are written together with their hosts (the tagging process may treat them as separate words); hyphenated compounds (these may or may not be treated as a whole); 'burkinostki'[1] (forms which only occur in a certain context, essentially forming a whole with another form across blank space).

Typically forms that are superficially identical but are perceived as different in grammar get different tags (in Slavic languages such are, for example, the 2nd and 3rd person singular aorist or imperfect forms of the verb, the dative and locative singular of *a*-declension nouns). However, different uses of the same form (for instance, within analytical forms) are not normally distinguished, although this is one of the tasks of morphosyntactic tagging.

MTE is an offshoot of, and builds upon, the MULTEXT project, which was oriented primarily towards the processing of the languages of Western Europe. It recognises 14 categories, 10 of which correspond to the traditional parts of speech. A list of features is associated with each category, and a set of values with each feature. Each word form pertaining to a given category must have all features, though some values may be marked as undefined (for example, verbs normally have person, but non-finite forms do not). On the whole, MTE tagsets have tended to adhere to the national grammatical traditions. As a side effect, the same phenomena in different languages have often been treated differently, especially in the absence of a precedent in the MULTEXT languages. Contrariwise, IPIC strays away from tradition. It classifies word forms into flexemes, which correspond to parts of speech only very roughly. Characteristically, the IPIC formalism is meant expressly for Polish.

The proposed specifications are based on a modified version of the flexemic tagset developed by Marcin Woliński and Adam Przepiórkowski for IPIC, for which a converter was written to bring that categorization closer to the MTE one. Some parts of speech (*flexemes* in IPIC terminology) were decomposed into finer categories (e.g., *qubliks*—into particles, interjections and adverbs), some were presented as combinations of selected values and attributes of existing parts of speech (derogative nouns, participles, etc.).

Thus, as in the case of Russian MTE tagset [Sharoff et al. 2008], our proposal takes into account the following:

- the consistency of MTE specifications,
- the specific features of the language,
- the possibility of automatic disambiguation of feature values,
- the de-facto standard—in our case, the IPIC tagset [Wolinski, Przepiórkowski 2003].

We shall now list and briefly discuss the categories in the tagset and the associated features. The possible values of a feature are listed after its name in brackets.

[1] The term was devised by Magdalena Derwojedowa to refer to dependent words which can be encountered and identified only in a fixed combination (as *Burkina* in *Burkina Faso*).

Noun (N)

The main classificatory feature of nouns is Type (common, proper, gerund). Gerunds (*bieganie* ‘running’) are considered a Type of Nouns (strictly speaking, they are a subtype of common nouns, but are treated as a type parallel to both common and proper nouns for convenience). The features Aspect (progressive, perfective) and Negation (no, yes) are added to characterize gerunds.

The complex category of noun class that Polish shares with with Slovak and both Sorbian languages is implemented through the three features Gender (masculine, feminine, neuter), Animate (no, yes) and Human (no, yes). The values of the latter two distinguish between the masculine-human (m1), masculine-animal (m2) and masculine-thing (m3) genders of traditional grammar and of IPIC ([+Animate, +Human], [+Animate, −Human], [−Animate, −Human], respectively). This allows the relevant morphological generalisations to be captured: the feature Human is neutralised in the singular, Animate in the plural. The attribute Human also expresses what the IPIC calls derogativity (derogatives in Polish are a class of plural noun forms which are [−Human] in the nominative/vocative but [+Human] in the accusative). As both Animacy and Humanity are justified semantically and the information about them is already recorded in the morphological analyser Morfeusz[2], the source of grammatical information, these data are retained in the MTE tags. To technically differentiate between derogative forms of lexically [+Human] nouns and those originally marked [+Animate, −Human], the nominative/vocative plural of derogatives is encoded using the fourth theoretically possible combination, [−Animate, +Human].

The features Number (singular, plural) and Case (nominative, genitive, dative, accusative, instrumental, locative, vocative) have their traditional interpretation.

Verb (V)

Verbs are classified by Type (main, auxiliary) and Aspect (progressive, perfective).

The non-finite verb forms and the mood of the finite verb are identified by the feature VForm (indicative, imperative, infinitive, impersonal, gerund). Note that gerund as a VForm means an adverbial participle (*imiesłów przysłówkowy*), not to be confused with gerund as a Noun Type (*gerundium*).

Verbs are further tagged for Tense (present, future, past). Finite verb forms have the features Person (first, second, third), Number (singular, plural), Gender (masculine, feminine, neuter) and Human (no, yes). Animacy is not relevant for verbs.

The feature Definiteness (full-art, short-art), recycled from the MTE tagset for Bulgarian, encodes here the Vocalicity of agglutinated clitics (e.g., *-em* vs *-m*). Since these are not articles, the names of the feature and both values are misnomers, but the phenomenon is similar to the Bulgarian one (essentially, allomorphy).

The feature Clitic (no, yes, agglutinant, demanding) encodes the agglutination phenomenon, which in Polish is similar to what the MTE tagset for Czech models through the feature Clitic_s for verbs and pronouns, but has a wider scope and affects more parts of speech, thus calling for a more general attribute. It is specified, e.g., for the indicative past tense form (corresponding to IPIC’s flexeme praet, the so-called pseudoparticiple) to differentiate between forms such as *gniótł* (value ‘no’) and *gniotł-* (‘demanding’), where the latter not only requires a clitic but also has different form. An ‘agglutinant’ is the clitic itself, e.g., *-em* ‘1sg’ in *gniotłem*. The value ‘yes’ is left to allow showing that a graphical word is a combination of a demanding (or free) segment and an agglutinant in case the word segmentation should be revised in the future.

No Voice feature need be defined for Polish verbs, as all verbal forms are active (adjectival/attributive participles are treated as adjectives).

[2] http://nlp.ipipan.waw.pl/~wolinski/morfeusz/

Adjective (A)

Adjectives are classified by Type (qualificative, participle). Qualificative adjectives have Degree (positive, comparative, superlative). Aspect (progressive, perfective), Voice (active, passive) and Negation (no, yes) are used for further differentiation of adjectival participles.

Gender (masculine, feminine, neuter), Animate (no, yes), Human (no, yes), Number (singular, plural) and Case (nominative, genitive, dative, accusative, instrumental, locative) work as for nouns, except that adjectives, like all other nominal categories other than nouns, have no vocative case forms.

The feature Definiteness (short-art, full-art) serves to label the IPIC flexeme *winien* 'obliged' and predicatives like *rad* 'glad' as short adjectives and to separate them from the bulk of full adjectives.

In contrast to the IPIC, ordinal numerals were extracted from adjectives and moved to numerals, and pronominal adjectives were moved to pronouns. Post-prepositional adjectives like *(po) polsku* 'in Polish' are treated as adverbs.

Pronoun (P)

Pronouns are subjected to the traditional classification through the feature Type (personal, demonstrative, indefinite, possessive, interrogative, relative, reflexive, negative, general). The IPIC tagset does not have pronoun types, so this information had to be supplied by hand. Further division is achieved by the features Referent_Type (personal, possessive) and Syntactic_Type (nominal, adjectival, adverbial).

Pronouns of the personal (but not the possessive) type are distinguished by Person (first, second, third). Gender (masculine, feminine, neuter), Animate (no, yes), Human (no, yes), Number (singular, plural), Case (nominative, genitive, dative, accusative, instrumental, locative) have the same interpretation as for the other nominal categories.

The feature Clitic (yes, no, agglutinant) distinguishes postprepositional forms (*nią*, *niego*) from regular ones (*ją*, *go*) and bound (agglutinating) clitics (*-ń*).[3]

The feature Definiteness (full-art, short-art) serves to separate full forms of pronouns (*jego*, *niego*) from short ones (*go*, *-ń*). Again, the names of the feature and both values should not to be understood literally; this attribute was used in order to avoid multiplication of attributes.

Adverb (R)

Two features are defined for adverbs: Degree (positive, comparative, superlative), as for adjectives, and Clitic (no, yes, agglutinant, burkinostka).

The IPIC tagset has a special treatment for 'adjectival' forms that are used to form composite adjectives (e.g., *polsko* in *polsko-ukraiński* 'Polish–Ukrainian'). These are considered agglutinating adverbs here.

Forms which can only be used in a fixed context (e.g., *polsku* in *po polsku* 'in Polish') are likewise classified as special kinds of adjectives in the IPIC. In this proposal such a form is labelled as a burkinostka.

Adposition (S)

Two features are defined for adpositions: Type with a single value (preposition; there are no postopositions in Polish) and Case (genitive, dative, accusative, instrumental, locative), which encodes the preposition's subcategorisation.

[3] Cf. the value 'bound' of the feature Clitic for Slovene pronouns like *zame* 'for me' which refers to the whole cluster of a preposition and a pronoun. This coding can be used for similar phenomena in Polish, e.g., *dlań* 'for him', provided the word segmentation is revised towards a more traditional one.

Numeral (M)

Numerals are classified by Form (digit, roman, letter) and Type (cardinal, ordinal, collect[ive]).

Gender (masculine, feminine, neuter), Animate (no, yes), Human (no, yes), Number (singular, plural) and Case (nominative, genitive, dative, accusative, instrumental, locative) are interpreted as expected.

The feature Class (definite34, definite), introduced in the MTE tagset for Czech, does what IPIC achieves through the accommodability (congr, rec) feature: ‘agreeing’ (congr) numerals such as *dwa*, *dwaj*, *trzy*, *trzej*, *cztery* have the value ‘definite 34’, whereas ‘governing’ (rec) numerals such as *pięć*, *pięciu*, *dwóch* are ‘definite’. The numeral *jeden* ‘1’ is left with the indefinite pronouns.

Particle (Q)

Particles were extracted by hand from IPIC’s qublik category along with adverbs, pronouns and interjections and a few conjunctions. The only feature associated with them is Clitic (no, yes, agglutinant, demanding). An agglutinant is a particle which is joined to another word (*by*, *że*). The value ‘yes’ labels a composite particle such as *niechby* when treated as one word; alternatively it may be encoded as a sequence of two particles, the optionally demanding *niech* and the agglutinant *by* (at the moment, the IPIC uses both approaches).

Conjunction (C), Interjection (I), Abbreviation (Y), Residual (X)

No features are associated with these categories.

The data associated with the proposed tagset are presented in the morphological specifications, a lexicon and a sample tagged corpus.

3 Mapping the tagsets and tags

To obtain corpora tagged with the proposed scheme, a conversion procedure was developed. It allows for conversion between the IPIC tagset and our MTE-based scheme. As the differences between tagsets are significant, the procedure is not trivial (it is discussed in the next section).

It is rather difficult to map the IPIC tagset on the MTE one without providing large lists of exceptions and conditions with lengthy explanations. Moreover, the available corpora use grammatical information coming from Morfeusz, which is not an open-source product. This is why the task of collecting the list of tags was approached empirically rather than theoretically and the mapping was basically conducted at the level of tags using information coming from already tagged corpora. For this purpose we have extracted a list of tags from the IPIC corpus.

3.1 Preparing data for conversion

3.1.1 The source corpora

In order to extract as complete as possible a set of morphosyntactic tags for Polish we used two sources: a manually disambiguated mini-IPIC consisting of 1 mln tokens and the large IPIC itself, which amounts to approx. 250 mln tokens. The first corpus was supposed to give us relatively reliable information about the number of tags and lemmas. Theoretically, there should be no such situation when two possible disambiguations are checked manually (this happens more often in the automatically disambiguated corpus, when disambiguation criteria are not sufficient for the tagger and several options are identified as correct).[4] The whole IPIC has been disambiguated using an automatic tagger, therefore the tag count statistics may be biased. Nevertheless, as it is 264 times larger than the manually-disambiguated one (as measured in tokens), we decided to employ both. Surprisingly, not only do the numbers of tag types in the two corpora not coincide, but there is a large group in each that is not present in the other. This is explained in part by differences in notation between corpora. For example, ppron12 receives the additional value of accentability in the large IPIC and this is reflected in the tags. So, both tags with and without this feature are available and used for the same forms in texts, which unnecessarily doubles their quantity.

3.1.2 Lemmatization

One of the problems of using the two corpora together as one source of information is that the lemmatization strategy differs slightly. This does not affect the list of tags but influences the lexicon and converter.

Most discrepancies in the lemmatization concern personal pronouns. In the small corpus there are three different lemmas for ppron3 (3rd person personal pronouns): *on*, *ona*, *ono* 'he, she, it'. In the large one they are all represented by the lexeme *on* 'he'[5]. For the purposes of both the taglist and the lexicon all such tags were relemmatized back to the small corpus pattern with three lemmas.

Gerunds are treated differently in the two corpora: in the small one they are lemmatized as their nominative forms, in the big one as the infinitive they are derived from. For the purpose of the lexicon, as well as in the converter, the lemmas were restored to the nominative case of the noun form. Also, negation has a more morphological status in the big corpus and lemmas are presented there without the negative prefix *nie-*. This was retained in the MTE version, where nouns possess negation (because gerunds are one of the types of noun).

3.1.3 The problem of disambiguation

Some disambiguation issues had to be dealt with also in the smaller, manually disambiguated corpus. This is connected, first of all, with truly ambiguous cases, when a word and the whole phrase can be interpreted in different ways. This is unavoidable but also extremely rare. Most of the other situations concern cases where two or more IPIC tags are mapped to a single MTE tag because in IPIC personal pronouns of the first and second person are tagged for gender, or past tense masculine verb forms for animacy and humanity, which is not done in MTE. For example, the verb *dał* '(he) gave' has three tags selected as correct (praet:sg:m1:perf, praet:sg:m2:perf, praet:sg:m3:perf), all of them corresponding to a single MTE one: Vmeis-sm (i.e., Verb Type=main Aspect=perfective VForm=indicative Tense=past Number=singular Gender=masculine). This naturally simplified the task of counting tags and their usage.

3.2 Conversion of tags

The collected tags amounted to 1295, including 898 tags from the small corpus and 397 tags from the big corpus that were absent in the small one. The tags were further processed and transformed into their closest MTE correspondents. They were split into their minimal values and recorded in a relational database with each value taking a separate column. Then the notation of values was replaced by the MTE one and their order was rearranged to fit the new tagset.

A large part of the original tags were mapped unconditionally. The rest had to be mapped on several MTE tags and the conditions of mapping were defined by special lists of lexemes that had to be treated as separate groups. For example, IPIC adjectives are mapped onto adjectives proper, adjectival pronouns and ordinal numerals. As the latter two are closed groups, their sets were defined in the lists of lexemes. In the

[4] We explain such cases and their origin below.

[5] In the MTE-3 Slavic languages whose lexicons are available for exploration there is no agreement either on how these forms should be lemmatized. Czech *my* 'we', *vy* 'you (pl)' are lemmatized as *já* 'I' and *ty* 'you (sg)', respectively. The situation in Slovene is the same. In Serbian and Bulgarian all four are different lemmas.

remaining cases a lexeme was referred to the adjectives proper.

In some cases MTE demands a more detailed description of categories than the IPIC; such divisions were introduced manually and recorded as lists of lemmas to be assigned specific tags. On the other hand, some original tags were simplified, which significantly reduced their number. The tags in the IPIC column[6] can be divided into the following groups:

- those that are mapped to exactly one tag in the MTE map (1192 tags): comparative and superlative degree forms of adjectives, verbs, adjectival participles, gerunds, cardinal numerals, depreciative nouns, personal and reflexive pronouns, plural forms of nouns, prepositions.
- those subjected to additional division into MTE groups, first of all qubliks and non-personal pronouns.
- new tags: collective numerals, some missing pronoun forms that where deduced.
- tags that were combined into one.

We discuss some of those cases in more detail below, and the distribution of tags according to categories and source corpora is summarized in Figure 7.

3.2.2 Expanding the IPIC tags

The overall number of IPIC tags, the arithmetic sum from both corpora, that we have managed to extract amounts to 1298.[7] 101 of them have received more then one projection in the MTE tags. Those are grouped in the following way: 60 tags for adjectives in the positive (neutral) degree of comparison were projected to 13 tags each; 18 substantive tags, to 2–7 tags each; qubliks were split into 7 categories with 27 unique tags, cf. Figure 1; predicatives were split into 3 categories with 4 tags. Such a large expansion of adjectival tags is connected first of all with separating ordinal numerals and adjectival pronouns from adjectives proper. Secondly, adjectival pronouns were split into semantic types (basically, 11 combinations of the Type and Referent_Type features in MTE), as practised in the MTE tradition. Similarly, subst tags for nouns were split into nouns proper and pro-nouns, the latter also having eight semantic types. The qublik class[8] contained adverbs that do not inflect for degree. Those were manually marked as such and relegated to adverbs (R). Apart from this, qubliks include all interjections (I) and pronouns, mostly adverbial but also a few adjectival ones, and the short reflexive *się* (P). A few conjunctions (C) and prepositions (S) were also redirected from qubliks to corresponding classes. Figure 1 below shows the distribution of qubliks into MTE classes with number:[9]

Figure 1. Distribution of qubliks in MTE projection.

Category	Example	MTE tags	Tokens
C	alboż	1	11
I	hej	1	179
P	jakoś, się	16	85
Q	że	2	74
R	wczoraj	4	233

[6] They cannot be called IPIC tags as some of them were added by us.

[7] 45 tags for numerals arising from permutation of attributes but not realized in the Polish language are not included into this list. They are present, however, among the tags rejected by the TaKIPI tagger during disambiguation of corpus texts. Along with the closedness of Morfeusz this is another reason for taking tagged corpora as the starting point for extraction of tags.

[8] The name of the category originates from the Polish word *kubło* ‘waste-paper basket’, which explains well the concept behind it.

[9] We are thankful to the participants of the Slavic Corpora discussion group who, with their comments and advice, helped to resolve some doubtful issues concerning the division of qubliks.

Category	Example	MTE tags	Tokens
S	ponad	2	7
X	mocium	1	8

In the treatment of predicatives we followed the approach explicated in [Derzhanski, Kotsyba 2008]: copulative *to* is classified as a pronoun, the items with the morphological properties of verbs (infinitives of verbs of perception), adjectives (short forms) or nouns (citation forms) as these same parts of speech, and all others as adverbs.

3.2.3 New tags

New interpretations were added very sparingly. Figure 2 below shows two new IPIC tags (those with no entries for quantity of tokens) for short feminine forms of personal pronouns in the genitive and accusative.

Figure 2. Example of added IPIC tags and their MTE correspondents.

IPIC tag	MTE tag	MTE extended	Tokens	Example
ppron3:sg:gen:f:ter:nakc:praep	Pp-3f--sgy-n	Pronoun Type=personal Person=third Gender=feminine Number=singular Case=genitive Clitic=yes Syntactic_Type=nominal	44	*niej*
ppron3:sg:gen:f:ter:nakc:praep	Pp-3f--sgasn	Pronoun Type=personal Person=third Gender=feminine Number=singular Case=genitive Clitic=agglutinant Definiteness=short-art Syntactic_Type=nominal		*ń*
ppron3:sg:acc:f:ter:nakc:praep	Pp-3f--say-n	Pronoun Type=personal Person=third Gender=feminine Number=singular Case=accusative Clitic=yes Syntactic_Type=nominal	11	*nią*
ppron3:sg:acc:f:ter:nakc:praep	Pp-3f--saasn	Pronoun Type=personal Person=third Gender=feminine Number=singular Case=accusative Clitic=agglutinant Definiteness=short-art Syntactic_Type=nominal		*ń*

Differentiating collective numerals from cardinal ones is theoretically foreseen in the IPIC (there is a special tag for this subcategory) but not implemented in the corpus. We have added 12 new tags for such forms (masculine and neuter times six cases). Neither animacy nor humanity were relevant there. The forms are the same for the masculine and the neuter, but the gender distinction was preserved as they cannot be used with feminine nouns.

3.2.4 Collapsing the IPIC tags

Preserving all possible information was our priority, so in fact collapsing means a more economic way of recording information. This is why decisions about rejecting some tags only seemingly led to losing data, as they were superfluous in practically all cases. For example, the three masculine genders differentiated in IPIC (m1, m2, m3) were replaced by a single masculine gender (m), but the information about peculiarities of inflexion encoded by m2 and m3, provided it is relevant in a particular case, is still stored in an MTE tag, being expressed by the categories of animacy and humanity. Numerous tags were simplified in this way in the following categories: adjectives, ordinal numerals, adjectival participles, verbal *l*-participles, numerals, and, most of all, personal pronouns.

Morfeusz presents a very detailed characteristics of word forms, often retaining attributes useless for differentiation. This leads to many tags that are never found in texts and have no theoretical justification. Moreover, they make disambiguation more difficult. For example, 3rd person personal pronouns (ppron3 flexeme in the IPIC) in general foresees 287 different IPIC tags that serve to describe 5 lemmas and their 23 forms. They are expressed by 65 MTE tags.

A similar situation is with the 1st and 2nd person personal tags (flexeme ppron12). There 146 such original IPIC tags map on 30 MTE ones.

All in all, there are 42 forms of personal pronouns in the IPIC and 433 tags for them, which were collapsed to 95 in the MTE version. The distribution of quantity of tags per word form is unequal, starting from the form *nim* with 53 interpretations in IPIC, followed by *nich* 33 and *nimi* 25 (16 forms with 10 or more interpretations) to *mu, jemu, ją* with 3 or 4 interpretations.

IPIC tags possess such attributes as accentability and prepositionality which are realized only in some forms. The extra two genders (m2 and m3) also unnecessarily increased the number of tags.

Figure 3. Tags for the 3rd person singular feminine personal pronouns' forms.

IPIC tag	MTE tag	Word form
ppron3:sg:acc:f:ter:akc:npraep	Pp-3f--san-n	*ją*
ppron3:sg:acc:f:ter:akc:praep	Pp-3f--say-n	*nią*
ppron3:sg:acc:f:ter:nakc:npraep	Pp-3f--san-n	*ją*
ppron3:sg:acc:f:ter:nakc:praep	Pp-3f--say-n	*nią*
ppron3:sg:acc:f:ter:npraep	Pp-3f--san-n	*ją*
ppron3:sg:acc:f:ter:praep	Pp-3f--say-n	*nią*

Legend:

Pp-3f--san-n: Pronoun Type=personal Person=third Gender=feminine Number=singular Case=accusative Clitic=no Syntactic_Type=nominal

Pp-3f--say-n: Pronoun Type=personal Person=third Gender=feminine Number=singular Case=accusative Clitic=yes Syntactic_Type=nominal

Figure 3 shows the situation with the two singular accusative forms of the personal pronoun *ona* ‘she’, which differ only in their prepositionality feature (the last two tags are from the mini-IPIC). The large IPIC adds the accentability attribute (short and full form in MTE-Polish specifications) that is not realized in the accusative, increasing the general quantity of tags to six. In the MTE tagset they were reduced again to two.

Let us have a look at some examples of disposing of the gender value in adjectivals. First the feature of gender as understood in the IPIC corpus was recast into 3 values: Gender proper, Animacy and Humanity. This gave the same number of combinations as the IPIC tagset. Further, Animacy and Humanity never have to be set simultaneously: every combination needs to contain only Gender and Humanity (66 original IPIC tags are represented by 22 MTE ones with with no Animacy value and Human=yes to differentiate between forms of nominative and accusative plural), or only Gender and Animacy (33 original IPIC tags are represented by 22 with no Humanity value and Animate=yes to differentiate between forms of accusative singular), or Gender alone. This led to a significant decrease in the number of target tags from 660 IPIC-based ones[10] for adjectival pronouns to 429 MTE ones and 629 IPIC tags grouped together as

[10] Originally 110 but multiplied by 6 for each semantic type.

adjectives to 425 MTE ones (including 439 active and passive adjectival participles mapped on 301 MTE ones), and finally 60 ordinal numerals split from the IPIC adjectives to 39 MTE ones.

Figure 4. Tags for ordinal numerals, the accusative case.

IPIC tag	MTE direct correspondent	MTE revised	MTE tag expanded	Example
adj:pl:acc:f:pos	Mlof--pa	Mlof--pa	Numeral Form=letter Type=ordinal Gender=feminine Number=plural Case=accusative	*pierwsze*
adj:pl:acc:m1:pos	Mlomyypa	Mlom-ypa	Numeral Form=letter Type=ordinal Gender=masculine Human=yes Number=singular Case=accusative	*pierwszych*
adj:pl:acc:m2:pos	Mlomynpa	Mlom-npa	Numeral Form=letter Type=ordinal Gender=masculine Human=no Number=singular Case=accusative	*pierwsze*
adj:pl:acc:m3:pos	Mlomnnpa	Mlom-npa	Numeral Form=letter Type=ordinal Gender=masculine Human=no Number=singular Case=accusative	*pierwsze*
adj:pl:acc:n:pos	Mlon--pa	Mlon--pa	Numeral Form=letter Type=ordinal Gender=neuter Number=plural Case=accusative	*pierwsze*
adj:sg:acc:f:pos	Mlof--sa	Mlof--sa	Numeral Form=letter Type=ordinal Gender=feminine Number=singular Case=accusative	*pierwszą*
adj:sg:acc:m1:pos	Mlomyysa	Mlomy-sa	Numeral Form=letter Type=ordinal Gender=masculine Animate=yes Number=singular Case=accusative	*pierwszego*
adj:sg:acc:m2:pos	Mlomynsa	Mlomy-sa	Numeral Form=letter Type=ordinal Gender=masculine Animate=yes Number=singular Case=accusative	*pierwszego*
adj:sg:acc:m3:pos	Mlomnnsa	Mlomn-sa	Numeral Form=letter Type=ordinal Gender=masculine Animate=no Number=singular Case=accusative	*pierwszy*
adj:sg:acc:n:pos	Mlon--sa	Mlon--sa	Numeral Form=letter Type=ordinal Gender=neuter Number=singular Case=accusative	*pierwsze*

The combinations of gender, animacy and humanity corresponding to the meanings of m1, m2 and m3 are shown in the second column. In the plural the forms *pierwszych* and *pierwsze* are differentiated only by the feature of humanity, this is why the values for animacy were removed. In the singular, the forms *pierwszego* and *pierwszy* are differentiated only by animacy, so the values for humanity were removed. This spares us 2 extra tags. Thus, only 9 out of 60 original IPIC tags retain features differentiated originally by the three masculine genders.

Another example of collapsing tags can be seen in verbal stem forms. The category of animacy was removed from this group, while humanity was left to differentiate such cases as *były* ‘were (non-m. human)’ and *byli* ‘were (m. human)’. However, this feature is important only for the plural forms. In the singular we get 6 tags out of the original 18: the ones in figure 5 plus the same combinations for the imperfective (progressive) aspect.

Figure 5. Tags for the *l*-participle.

IPIC tag	MTE tag	MTE tag expanded	Word form
praet:sg:m1:perf	Vmeis-sm	Verb Type=main Aspect=perfective VForm=indicative Tense=past Number=singular Gender=masculine	*został*
praet:sg:m2:perf	Vmeis-sm	Verb Type=main Aspect=perfective VForm=indicative Tense=past Number=singular Gender=masculine	*został*
praet:sg:m3:perf	Vmeis-sm	Verb Type=main Aspect=perfective VForm=indicative Tense=past Number=singular Gender=masculine	*został*
praet:sg:m1:perf:agl	Vmeis-sm--d	Verb Type=main Aspect=perfective VForm=indicative Tense=past Number=singular Gender=masculine Clitic=demanding	*odniosł*
praet:sg:m2:perf:agl	Vmeis-sm--d	Verb Type=main Aspect=perfective VForm=indicative Tense=past Number=singular Gender=masculine Clitic=demanding	*odniosł*
praet:sg:m3:perf:agl	Vmeis-sm--d	Verb Type=main Aspect=perfective VForm=indicative Tense=past Number=singular Gender=masculine Clitic=demanding	*odniosł*
praet:sg:m1:perf:nagl	Vmeis-sm--n	Verb Type=main Aspect=perfective VForm=indicative Tense=past Number=singular Gender=masculine Clitic=no	*poniósł*
praet:sg:m2:perf:nagl	Vmeis-sm--n	Verb Type=main Aspect=perfective VForm=indicative Tense=past Number=singular Gender=masculine Clitic=no	*poniósł*
praet:sg:m3:perf:nagl	Vmeis-sm--n	Verb Type=main Aspect=perfective VForm=indicative Tense=past Number=singular Gender=masculine Clitic=no	*poniósł*

Figure 6 shows a very rough correspondence of categories in the MTE and IPIC.

Figure 6. Projection of MTE basic categories on IPIC ones.

MTE category	Closest IPIC flexeme
Noun (N)	subst(−) ger
Verb (V)	verb(−)*
Adjective (A)	adj(−)
Adverb (R)	adv(+)

Pronoun (P)	subst(−)
	adj(−)

Numeral (M)	num (+)
Particle (Q)	qub(−)
Adposition (S)	prep (−)**
Conjunction (C)	conj (+) **
Residual (X)	ign(−)
Abbreviation (Y)	ign(−)
Interjection (I)	qub(−)

Legend:
* understood as IPIC alias for verbal flexemes, without the gerund (-*nie* form)
** slight modifications
(+) as well as from other categories
(−) but not all of them

We tried to present the main corresponding flexeme.

We can see from the table and the legend that conjunctions and prepositions are the only parts of speech in the IPIC whose interpretation coincides with MTE. Among the few exceptions are such words as *niby*, *jak* ‘as, like’ that are classified in IPIC as prepositions governing the nominative case. They are treated as conjunctions in MTE, where the specifications for prepositions do not allow them to subcategorise for the nominative. Also, a few conjunctions were found in the qublik class.

3.3 Statistics of tags

The quantities of tag types in the original (both IPIC corpora) and the target tagsets are very close: 1295 in the IPIC and 1266 in the MTE. Their content and informativity, however, differs greatly. (On their way to the final number, while being converted, they passed through a reduction of a nearly twice larger overall quantity.)

The MTE tag list contains 1266 tags, 102 of them have been obtained from more than one IPIC tag.

Figure 7. Correspondence of tags depending on the category and the source corpus.

	Original IPIC tags (M)*	Original IPIC tags (A) **	Expanded IPIC tags (M)	Collapsed IPIC tags (M) MTE
Noun (N)			95	95
subst	69			71
depr	2			
ger	21	3	24	24
Verb (V)		-	71	56
aglut	6			
bedzie	6			
fin	12			
imps	2			
impt	6			
inf	2			
praet	32			

Adjective (A)		203	629	425
adj	171	11 (comp/sup degree)		
pact	82	125		
ppas	167	65		
pcon	1	1		
pant	1	1		
winien	10	-		
Adverb (R)	3		7	6
pred	1			
Pronoun (P)		182 (only personal)	1167 (with new ones)	597
ppron12	140		146	30
ppron3	107		287	65
siebie	5		5	5
Numeral (M)			114	75
cardinal	33	3[11]	34 (transfer from adj)	22
ordinal			60 + 2	39
collective			18 (newly added)	12
Particle (Q)	1	-	1	1
Adposition (S)	14	-	15 (transfer from qub)	6
Conjunction (C)	1	-	1	1
Residual (X)	1	5[12]	8	1
Abbreviation (Y)	-	-	1	1
Interjection (I)	1	-	1	1
Total	898	397	2157 (without 45 theoretically impossible)	1266

* M – manually disambiguated corpus
** A – automatically disambiguated corpus, only the new tags that were absent from M.

3.4 Word segmentation

One of the major differences between the IPIC approach and the MTE one is in the word segmentation principles. This is not a trivial issue and calls for the development of an optimal strategy for dealing with such situations in the future. The IPIC approach is a highly practical and economic one but it deviates from the traditional understanding of what a word is, which is realized in the MTE records of language material. A typical example of token representation in the IPIC:[13]

```
<orth>mogli</orth><lex disamb="1"><base>móc</base><ctag>praet:pl:m1:imperf</ctag></lex>
<ns/>
<orth>by</orth><lex disamb="1"><base>by</base><ctag>qub</ctag></lex>
<ns/>
<orth>ście</orth><lex disamb="1"><base>być</base><ctag>aglt:pl:sec:imperf:nwok</ctag></lex>
```

Here one graphical word *moglibyście* 'you(pl) could' is presented by three segments with their own lemmas. The same word in the MTE notation (before revising its segmentation):

```
<w lemma="móc" ana="Vmpis-pmy">mogli</w>
<w lemma="by" ana="Q">by</w>
<w lemma="być" ana="Vapip2p--sa">ście</w>
```

Legend:

[11] Including two tags for digits added by the TaKIPI tagger, whereas in IPIC digits would be classified as residuals (ign).

[12] As in the case with numerals, these are the TaKIPI tagger tags that are "ignorable" for both the IPIC and the MTE. Examples: tdate, tmail, turi, tdate, tsym.

[13] The <tok> tags were removed here to simplify the representation.

Vmpis-pmy: Verb Type=main Aspect=progressive VForm=indicative Tense=past Number=plural Gender=masculine Human=yes

Q: Particle

Vapip2p--sa: Verb Type=auxiliary Aspect=progressive VForm=indicative Tense=present Person=second Number=plural Definiteness=short-art Clitic=agglutinant

The IPIC notation includes a “no space” tag <ns/> to signal cases when a segment of a word is presented as a separate lemma in the corpus. This allows several problems to be solved: the floating ending of the past indicative verb forms (a remnant from the old analytical perfective form) which can be attached practically to everything (nouns: *swiniaś* (*świnia jesteś*) ‘pig (you) are’, pronouns: *tyś* (*ty jesteś*) ‘you are’, conjunctions: *żebyście* (technically: *że by jesteście*) ‘in order for you(pl) to (be)’, adverbs: *wcaleś* (technically: *wcale jesteś*) ‘at all (you) are’, etc.) and the multiplication of verbal forms that can created according to strict agglutinating rules: *myślał-by-m* ‘I would think’, *znalazł-by-ś* ‘you would find’. If we wanted to treat all such clusters as single words, we would frequently be at a loss for a way to name them or would have to introduce a bulky category of predicativity for nouns, adverbs, etc., and further complicate the interpretation of their morphology. These cases are treated as technically combined independent words. Combinations of prepositions and pronouns like *dlań* (*dla niego*) ‘for him’ are marked in the MTE tagset with the help of the Clitic feature for pronouns. The value a(gglutinant) shows that the string is technically part of an orthographic word, cf. Figure 8.

Figure 8. Morphological tagging for strings like *dlań*.

dlań	dla	Spg	Adposition Type=preposition Case=genitive
	ń	Pp-3m--sgasn	Pronoun Type=personal Person=third Gender=masculine Number=singular Case=genitive Clitic=agglutinant Definiteness=short-art Syntactic_Type=nominal

This, however, means that each segment receives an independent morphosyntactic interpretation, including tense etc. information (cf. the interpretation of *moglibyście* above), which is at variance with traditional grammatical description and speakers’ intuitions. We believe that the problem can be solved and a more truthful picture can be achieved by the partial use of a secondary grouping. However, not all of these cases can and need to be treated as whole words (let us remember that orthographic rules are often a matter of convention).

We will distinguish cases when the agglutinant rambles away (*bym mógł*, *świniaś*, *dlań*) and when it accompanies its master participle. The former will have to await further analysis using syntactic parsing, as it is not always possible to technically differentiate between situations when it is originally an ending of the past verbal form that carries the information about the category of person and when it represents an independent verb in present tense. The latter was modified by combining both segments’ forms and their grammatical information to generate a single tag for the whole.

Thus a two-segment word *mogliście* after revising its segmentation looks in the MTE notation as follows (cf. with a three-segment word above):

<w lemma="móc" ana="Vmpis2pmy-y">mogliście</w>

A similar situation obtains with the clitic *-by*, which introduces the conditional mood. This clitic can be a standalone word form (when it precedes the verb) or a part of the verb form. In the latter case, the verb stem and the clitic are combined into a single token with a new grammatical information. The Tense value is changed into “present” and the Form acquires the value “conditional” instead of the former “indicative”. As well as in the example above, the clitic can also be followed by a floating ending—in such cases all the information is integrated into a single verb token.

Below are two examples of conversion: a third person plural conditional verb form, *mogliby* ‘they could’, and a second person plural conditional verb form, *moglibyście* ‘you(pl) could’.

```
<w lemma="móc" ana="Vmpcp3pmy-y">mogliby</w>
<w lemma="móc" ana="Vmpcp2pmy-y">moglibyście</w>
```

3.5 Tag converter

The discussed conversion method has been implemented in the Python programming language, the code and the data are available online at http://domeczek.pl/~polukr/mte-conv/. The converter consists of source code and separate files with conversion tables (tab-delimited lists). Each entry in the main conversion table may be a 1:1 tag correspondence or a reference to another conversion table (for lemma-based conversion rules). When running, all the tables are first read and indexed, which allows for faster performance. The converter reads IPIC XML files and produces TEI XML output compliant with other MTE sample corpus files.

As noted above, the conversion is conducted at the level of tags, i.e., the conversion tables provide a closed list of tags and rules for their conversion, with no generalisation. The obvious disadvantage is that we may encounter an unexpected tag. This solution still seemed preferable since it is not an easy task to capture a reasonable generalisation within a moderate set of rules while assuming that the employed list of tags is quite extensive. What is more, some well-formed IPIC tags are practically impossible, if not invalid —it may be desirable to get explicit information about such cases. The out-of-list tags are converted to residuals (X) and reported to the user.

4 Deliverables

In order to include a new language into MTE, the following package should be prepared: morphosyntactic specifications with a MSD index (representative list of possible tags), a lexicon and a sample of a tagged corpus.

4.1 Morphosyntactic specifications

The morphosyntactic specifications have been prepared in TEI XML format. The whole description is contained within one XML file with several sections. The file commences with a header containing metadata, followed by the main part which specifies each category, its attributes and their possible values. Every category is followed by optional notes/comments and a table which presents possible combinations of tags for this particular category. XML files can be transformed into HTML format, which is more convenient for the human reader, with the help of special XSLT writing scripts (stylesheets) provided by MTE V.4 developers, cf. [Erjavec 2009].

Figure 9 shows a fragment of the specifications as they look in HTML format (Polish adverb).

Figure 9. A fragment of the specifications in HTML (Polish adverb)

0	CATEGORY	Adverb		R
1	Degree	positive comparative superlative	p c s	
2	Clitic	yes no agglutinant	y n a	

		burkinostka u	

The last part of the specifications – the MSD index – consists of an extensive tag list, providing token occurrence count as well as example forms and lemmas. Both source corpora were fed through the converter. Employing both of them was significant, since there is a slight difference in the adopted tagging scheme: some categories are considered optional and omitted in the smaller corpus (we wanted to acquire all of the allowed tags). The resulting lists of tags are combined; the overlapping part is taken from the manually disambiguated corpus. To balance the acquired tag occurrence counts, we multiply the counts taken from the bigger corpus by an appropriate ratio.

Figure 10. A fragment of the MSD index.

MTE tag	MTE expanded	Tokens	Example
Vmeis2sf--y	Verb Type=main Aspect=perfective VForm=indicative Tense=past Person=second Number=singular Gender=feminine Clitic=yes	85	*powiedziałaś/powiedzieć, zrobiłaś/zrobić, przyszłaś/przyjść*
Vmeis2sm--y	Verb Type=main Aspect=perfective VForm=indicative Tense=past Person=second Number=singular Gender=masculine Clitic=yes	274	*przyszedłeś/przyjść, powiedziałeś/powiedzieć, zrobiłeś/zrobić,*
Vmeis2sn--y	Verb Type=main Aspect=perfective VForm=indicative Tense=past Person=second Number=singular Gender=neuter Clitic=yes	1	*pozostałoś/pozostać, przeszłoś/przejść*
Vmeis-pf	Verb Type=main Aspect=perfective VForm=indicative Tense=past Person=second Number=plural	619	*odbyły/odbyć, rozpoczęły/rozpocząć, zaszły/zajść*

Whenever possible, three examples of form/lemma pairs for a tag (some tags occur with one or two distinct forms only) are provided. To lower the number of repetitions, a simple heuristic for the selection of examples was employed. Some tokens in the corpus contain more than one candidate tag. Fortunately, many of these ambiguities disappeared after the conversion (as the proposed standard does not follow all the distinctions introduced in IPIC, which was a major cause of insoluble ambiguities). Nevertheless, some of them remained, resulting in troublesome situations, especially those coming from the big corpus submitted to an automatic disambiguation. We decided to count such candidate tags as fractions of occurrences (their counts adding up to 1 for a token).

4.2 The lexicon

The lexicon is meant to provide full inflection paradigms of the most frequent lemmas. As no extensive lexicographic resource with such information is available for Polish, we resorted to the corpus (IPIC). The 15 thousand most frequent lemmas were extracted from it with the help of Poliqarp.[14] Then the remaining forms for those lemmas were extracted from the large corpus. The lexicon includes a word form, its lemma, its tag and the number of token occurrences in the IPIC.

Figure 6. A fragment of the lexicon.

absurdami	absurd	N-mnnpi	17

[14] http://korpus.pl/index.php?page=poliqarp

absurdem	absurd	N-mnnsi	307
absurdom	absurd	N-mnnpd	6
absurdowi	absurd	N-mnnsd	4
absurdu	absurd	N-mnnsg	578
absurdy	absurd	N-mnnpa	59
absurdy	absurd	N-mnnpn	58
absurdzie	absurd	N-mnnsl	17
absurdów	absurd	N-mnnpg	163
aby	aby	C	201168
ac	ac	X	1099
ach	ach	I	1170

The total number of unique word forms in the lexicon is 175848 (roughly 11.72 per lemma), while the number of forms with all possible interpretations is 339031.

4.3 The corpus

The MTE-like tagged corpus in our case consists of one book, approx. 100000 words, namely George Orwell's *1984*. This book was chosen because it was used for the MTE multilingual parallel corpus for 11 languages, thus adding it was a natural way to extend the multilingual MTE parallel corpus for Polish and is intended to facilitate the validation of the specifications for Polish and the converter on sufficiently large language data.

The tagging was performed with the help of TaKIPI program, cf. [Broda et al. 2008], specially developed for tagging Polish using IPIC tagset. Afterwards the tag converter was used to bring it to MTE-style format. The resulting corpus contains 79807 word tokens and 17642 punctuation mark occurrences. The word tokens appear with 801 different MTE tags and 9480 different lemmas. Below we present a fragment of the corpus in the TEI XML format:

```
<p id="Opl.5">
<s id="Opl.5.1">
<w lemma="być" ana="Vmpis-sm">Był</w>
<w lemma="jasny" ana="A-pm--sn">jasny</w>
<c>,</c>
<w lemma="zimny" ana="A-pm--sn">zimny</w>
<w lemma="dzień" ana="N-mnnsa">dzień</w>
<w lemma="kwietniowy" ana="A-pmn-sa">kwietniowy</w>
<w lemma="i" ana="C">i</w>
<w lemma="zegar" ana="N-mnnpn">zegary</w>
<w lemma="bić" ana="Vmpis-pmn">biły</w>
<w lemma="trzynasty" ana="Mlof--si">trzynastą</w>
<c>.</c>
</s>
```

5 Conclusions and future work

An MTE-4 compliant package for the Polish language was prepared on the basis of existing resources and presented in this paper. This is an important step in integrating linguistic resources of Slavic languages, as it makes Polish much more comparable than it was before. Of course, this is only a first step and much remains to be done.

One point that received relatively little attention in [Derzhanski, Kotsyba 2009], but may be very important for comparative studies based on the common tagset and the parallel corpus, is that certain categories (or rather subcategories) existing in most MTE languages are only explicated in some of them. For example, the Russian MTE tagset introduces non-specific pronouns (*весь* 'all', *всякий* 'any, every', *сам* 'oneself', *самый* 'the very', *каждый* 'every, each', *иной* 'other', *любой* 'any', *другой* 'other'). This category, inspired by MAK Halliday's works, is not part of either Russian traditional grammar (the standard description of which is the Academic grammar), the theoretical premises of the Russian National Corpus, or the descriptions of other MTE languages. Nevertheless, items semantically and etymologically corresponding to the words in this group exist in all MTE Slavic languages, though classified as other types of pronouns or even other parts of speech. This issue deserves a separate investigation; here we just want to signal that both those who deal with language description and with searching through the parallel corpus have to be aware of different granulation levels for some grammatical categories. Likewise, participles are treated variously as adjectives or verb forms in MTE lexicons. The earlier mentioned lemmatization discrepancies need to be removed. And so on.

As for Polish itself, its specific word segmentation regarding clitics needs further syntactic analysis to correct grammatical information provided by tags about some agglutinated forms of *być* 'to be'.[15] Similarly, clustering analytical verb forms for Polish and other languages would give us a picture much closer to the traditional understanding of grammar and would facilitate further linguistic research and information retrieval.

All the described resources are very "fresh" and need validation to eliminate possible mistakes. It would be very useful if online search in the existing parallel corpus were provided. Presently, the resources from MTE-3 version are available for download upon registration. However, the absence of search tools does not allow linguists to use their full capacity. We would expect that giving such a possibility to a greater public could result in a feedback from which the general quality of corpora and the rest of the resources could only benefit.

Bibliography

[1] Broda B., Piasecki M. and Radziszewski A. (2008). Towards a Set of General Purpose Morphosyntactic Tools for Polish. In *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 *Proceedings of Lexicographic Tools and Techniques: Proceedings of the MONDILEX First Open Workshop*, pages 68–79, Moscow: IITP—RAS.

[3] Derzhanski I. and Kotsyba N. (2009). Towards a Consistent Morphological Tagset for Slavic Languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian. In Garabík 2009, pp. 9–26.

[4] 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.

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

[6] Erjavec, T. (2009). MULTEXT-East Morphosyntactic Specifications: Towards Version 4. In Garabík 2009, pp. 59–70.

[7] Garabík, R. (ed.) (2009). *Proceedings of Metalanguage and Encoding Scheme Design for Digital Lexicography: MONDILEX Third Open Workshop*, Bratislava, 15–16 April 2009.

[8] 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.

[9] Lewis M. P. (ed.), 2009. Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International. Online version: ‹http://www.ethnologue.com/›.

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

[11] Przepiórkowski A. (2009). A comparison of two morphosyntactic tagsets of Polish. Version of 15

[15] If the agglutinant is a floating ending and it is possible to identify its master, the information about the person should be added to the verbal form. Depending on whether *być* is an independent verb or the verbal ending, the same form carries different grammatical information. If it is independent the interpretation in the tag gives the truthful picture about its grammar but if it is an ending, its grammatical information is in conflict with the one of the master verbal form.

July 2009. To appear in *the proceedings of the MONDILEX workshop held in Warsaw*, 29–30 June 2009. URL: ‹http://nlp.ipipan.waw.pl/~adamp/Papers/2009-mondilex/›.

[12] Roszko, R. (2009). Morphosyntactic Specifications for Polish. Theoretical Foundations. Description of Morphosyntactic Markers for Polish Nouns within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004). In Garabík 2009, pp. 140–150.

[13] 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›.

[14] 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.

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

[16] Николаева Т. М. (2008). Непарадигматическая лингвистика. *История «блуждающих частиц»*. Москва. Studia Philologica.