Mediating between Incompatible Tagsets*

Alexandr Rosen

Charles University
Faculty of Arts, Prague
E-mail: alexandr.rosen@ff.cuni.cz

Abstract

The issue of incompatible morphosyntactic tagsets in multilingual corpora could be solved by an abstract hierarchy of concepts, mapped to language-specific tagsets. The hierarchy supports the user and tools by resolving categories that do not match the relevant tagset in queries, by providing links between language-specific tagsets, and by displaying responses using a preferred tagset. The hierarchy, built using the methods of Formal Concept Analysis, can also help to refine morphosyntactic annotation in one language by using word-to-word alignments to parallel texts tagged by a different tagset.

1 Introduction

Users of multilingual corpora are often confronted with a variety of language-specific morphosyntactic tagsets. To use tags in a query or to understand its results requires cheat sheets or even lengthy manuals. Without the benefit of intuitive understanding of distinctions and similarities between notationally different or similar tags, multilingual applications drawing on linguistic knowledge and more abstract (syntactic and semantic) annotation schemes built on top of morphosyntactic annotation stumble over an even harder problem.

The ideal solution could be a single consistent standardised annotation scheme in the spirit of *MULTEXT-East* [1]. However, to build a multilingual corpus using such a scheme seems unrealistic, especially when more than a handful of languages are involved.[1] Available taggers are trained on different tagsets, and consistently annotated training data are seldom available even for typologically close languages.

*This work was supported by grant no. MSM0021620823 of the Czech Ministry of Education, Youth and Sports, as a contribution to the parallel corpus project *InterCorp*.

[1]The parallel corpus *InterCorp* currently offers on-line concordances in 23 languages, 14 of them tagged with different morphosyntactic tagsets. The corpus can be queried at korpus.cz/Park after registration at http://ucnk.ff.cuni.cz/english/dohody.php. For more information about the project see http://korpus.cz/intercorp/.

Confronted with texts already tagged in different ways, the user may still believe that tagsets can be translated into a common standard. But a given tag may be too specific or too general to be expressed by a tag from a different tagset. Fig. 1 illustrates the tagset variety using comparable examples of prepositional phrases in 11 languages, tagged by available tools.[2] While some corresponding tags used in the examples are indeed notational equivalents, other tags are not related 1:1. The English tag `IN`, unlike all its prepositional counterparts, is used also for subordinating conjunctions, the German tag `ADJA` covers attributive adjectives (including ordinal numerals) irrespective of degree, while its English counterpart `JJS` is used for superlative adjectives, ignoring the attributive/predicative distinction. The Czech and Polish words *těch* and *tym* are members of the same class, yet the Czech form is tagged as demonstrative pronoun, undistinguished between attributive or substantive use, while the Polish form is tagged on a par with all forms of adjectival declension, including some other types of pronouns and numerals. The partial overlaps in the meaning of corresponding tags are reminiscent of translational mismatches in bilingual dictionaries, including phenomena such as false friends.

en	in `IN`	the `DT`	remotest `JJS`	exurbs `NNS`
de	in `APPR`	den `ART`	abgelegensten `ADJA`	Außenbezirken `NN`
nl	in `600`	dit `370`	schitterende `103`	appartement `000`
fr	dans `PRP`	les `DET:ART`	plus lointaines `ADV ADJ`	banlieues `NOM`
sp	en `PREP`	las `ART`	zonas `NC`	más remotas `ADV ADJ`
it	da `PRE`	queste `PRO:demo`	lingue `NOM`	babeliche `ADJ`
ru	v `Sp-l`	samych `P--pl`	otdaljonnych `Afp-plf`	rajonach `Ncmpln`
cs	v `RR-6`	těch `PDXP6`	nejodlehlejších `AAFP6---3A`	zástavbách `NNFP6---A`
bg	na `R`	tova `Pde-os-n`	prijatelsko `Ansi`	dviženie `Ncnsi`
pl	w `prep:loc:nwok`	tym `adj:sg:loc:m3:pos`	wspaniałym `adj:sg:loc:m3:pos`	apartamencie `subst:sg:loc:m3`
hu	a `ART`	szép `ADJ`	katalán `ADJ`	lányba `NOUN(CAS(ILL))`

Figure 1: Differences in tagging: prepositional phrases

[2]Bulgarian, Dutch, English, French, German, Italian, Russian and Spanish are tagged by *TreeTagger* (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/), Czech by *Morče* (http://ufal.mff.cuni.cz/morce/), Polish by *TaKIPI* and *Morfeusz* (http://nlp.ipipan.waw.pl/TaKIPI/), Hungarian by *HunPOS* (http://code.google.com/p/hunpos/). The tags used here and below are often truncated for brevity.

When the problem of converting between incompatible tags and tagsets concerns only closed-class items (pronouns, function words), it can be solved by using lexeme-specific information corresponding to the source tag (see [6]). In cases involving open word classes we could use an intermediate representation that allows for underspecification at the cost of leaving the target tagset with a potentially imprecise translation of the source tag, as in *Interset* [9]. In the context of many different languages and tagsets, the latter option is more appealing, provided that the language-specific tagsets are correctly linked with the abstract interlingual categories and the representation allows for an arbitrary level of specificity. Both of these features, not inherent to *Interset*, are important for using the representation as the common tagset, and for deriving the most appropriate target tag, which may be too general or too specific, but the extent of the residual part is always known.

Our goal is to delegate the task of dealing with multiple tagsets in a corpus to such an abstract interlingual hierarchy of linguistic categories, where each language-specific tag is mapped onto a node, positioned appropriately with respect to the interpretation of other tags. Because the differences between tagsets often reflect different linguistic perspectives rather than typological distinctions between the relevant languages, a specific word class is seen as an intersection of classification along several dimensions. Following [5] and others, the hierarchy takes three different views of the concept of word class. Thus, the tag for the Czech relative pronoun *který* 'which' is decoded as a category with the properties of lexical pronoun, inflectional adjective and syntactic noun, each with its appropriate morphological characteristics.

Rather than adopting or attempting to design a universal typology of linguistic categories, we prefer to base the hierarchy on distinctions present in our language-specific tagsets and stay open to future extensions. The hierarchy can be built and mismatches between tagsets partially resolved using Formal Concept Analysis [2]. In a parallel corpus with word-to-word alignment and the definition of language-specific domains of the hierarchy, morphosyntactic annotation can be refined by adding information from corresponding tags in other languages, even when the individual tagsets do not make that distinction.

2 Word Classes in 3D

The traditional list of eight word classes is defined by a mix of morphological, syntactic and semantic criteria. For nouns or adjectives the three criteria agree. Nouns refer to entities and decline independently in typical nominal positions; attributive or predicative adjectives represent properties and agree with nouns. On the other hand, numerals and pronouns are defined solely by semantic criteria, while their syntactic and morphological behaviour is rather like that of nouns (cardinals and personal pronouns) or adjectives (ordinals and possessive pronouns). For such cases, the option of a cross-classification along several dimensions seems attractive. Distinctions between the three aspects are borne out also by tagsets. The

Czech tagset has a preference for lexically-based classification [3], the Polish tagset [8] for inflectional word classes, the German tagset distinguishes pronouns by their syntactic function.

A comparison of tags in closely related languages is illustrative. An item tagged as adjective in the Polish tagset (`adj`) can be tagged in the Czech tagset also as an ordinal numeral (`Cr`), possessive (`P8`), demonstrative (`PD`) or relative pronoun (`P4`). A Polish tag for non-inflected words (`qub`) may correspond to a Czech tag for particles (`TT`), non-gradable adverbs (`Db`), reflexive pronouns (`P7`), subordinating (`J,`), or coordinating conjunctions (`J^`).

The 3D space helps to sort out such differences in tagsets. Using the tagset specification, properties of each tag can be identified and related to similar tags in other tagsets. The properties translate into categories in the abstract hierarchy, as in Fig. 2, where the topmost node *wcl* stands for nouns, adjectives and relative pronouns. Its daughters are labelled by a word-class aspect: *lexical* (for 'semantic'), *inflectional* (for 'morphological') and *syntactic*.[3] The other nodes stand for word classes in the three respective dimensions, distinguished in their labels by the initial letter. The seven nodes share only three daughters. Each of the three objects inherits the property of being a word class according to the three criteria.

Each node denotes a set of objects – language-specific tags. The topmost node denotes all tags in all tagsets. Immediate subnodes of a node denote its subsets. A tag denoted by a node must be denoted by at least one of its subnodes. A node can be a subnode of more than one node. In this case, the subnode denotes a subset of the intersection of the sets denoted by its supernodes.

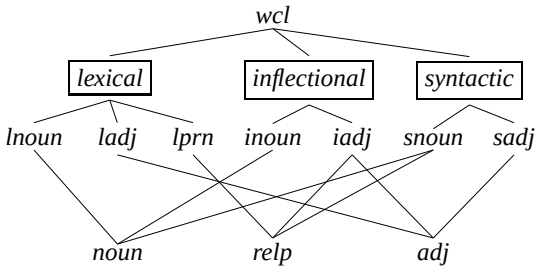

Figure 2: A hierarchy for nouns, adjectives and relative pronouns

Nouns and adjectives are members of their respective classes along all the three dimensions. On the other hand, a Czech *wh-* form *který* 'which' in its use as a relative (rather than interrogative) pronoun (1) is a *syntactic* noun as the subject of the relative clause, a *lexical* pronoun with "dog" as its antecedent, and – due to its adjectival declension – an *inflectional* adjective.

[3] We use *lexical* rather than *semantic* – *lexical* word classes have their properties specified in the lexicon. The boxes around the labels suggest that the sets of objects denoted by the sister nodes are identical.

(1) Psa, který nemá náhubek, do vlaku nepustí.
dog$_{\text{ACC}}$ which$_{\text{NOM}}$ has$_{\text{NEG}}$ muzzle$_{\text{ACC}}$ into train let in$_{\text{NEG,PL,3RD}}$
'An unmuzzled dog won't be allowed on the train.'

The hierarchy in Fig. 3 focuses on Czech numerals and pronouns. ordinals such as *pátý* 'fifth' are treated as *lexical* numeral and adjective – both *inflectional* and *syntactic*. Possessive pronouns differ in being *lexical* pronouns. Personal pronouns are inflectional and syntactic nouns, similarly as cardinal numerals. The interrogative homonym of the relative *který* can be used as a syntactic adjective or noun. The node *intp* inherits from *snom*, representing syntactic nouns *or* adjectives, while *relp* can only be a syntactic noun, due to its ancestor *snoun*.

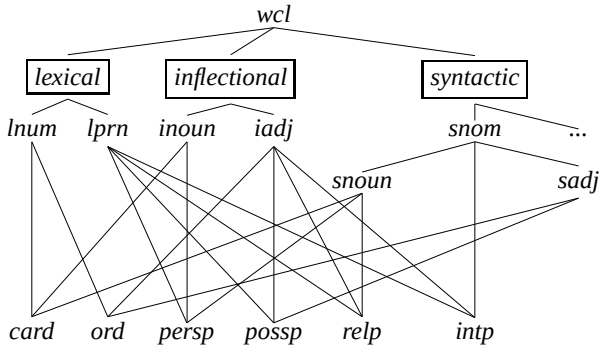

Figure 3: Distinguishing types of numerals and pronouns in a hierarchy

Který in its relative and interrogative use shares a single tag (P4), corresponding to a category ambiguous between relative pronoun and syntactic noun on the one hand and interrogative pronoun and syntactic adjective or noun on the other. The modified hierarchy in Fig. 4 captures this ambiguity. The Czech tag P4 corresponds to a node labelled *lprn* $\wedge$ *iadj* $\wedge$ *snom*.

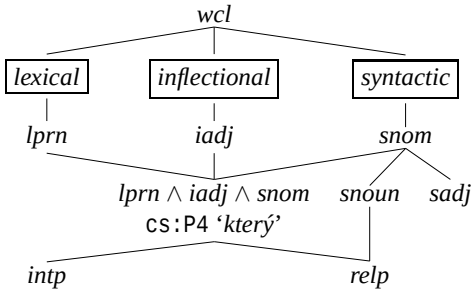

Figure 4: A single node for interrogative and relative pronouns

The concept of three-dimensional word class allows for proper mapping between language-specific tagsets. The tag for adjective in English, German, French,

Italian and Polish covers also ordinal numerals. If all these tags are represented as *syntactic* adjectives, they end up correctly in the same class as Czech, Spanish, Russian or Bulgarian adjectives, ordinal numerals and possessive pronouns. Their *lexical* word class is unknown, although it is not arbitrary. Fig. 5 shows a fragment of the hierarchy with a node representing exactly ordinal numerals and adjectives, labelled $(lord \vee ladj) \wedge iadj \wedge sadj$ and corresponding to the German tag ADJA.

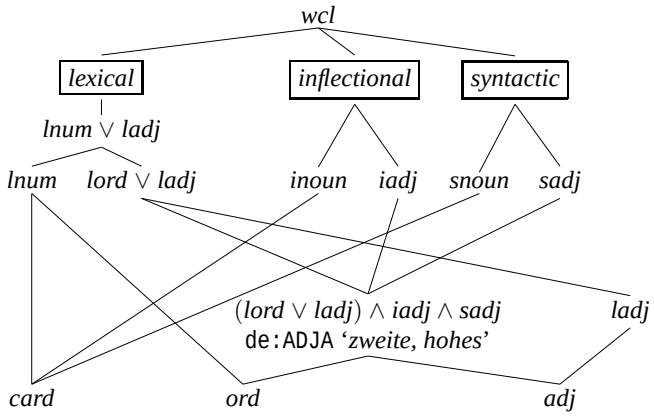

Figure 5: A single node for ordinal numerals and adjectives

The German ordinal number *zweite*, tagged as adjective (similarly as *hohes*), is a subtype of inflectional and syntactic adjective (*iadj* and *sadj*), and also a subtype of a general type covering lexical adjectives and ordinal numerals (*ladj* $\vee$ *lord*).

Word class of any flavour may be required to co-occur with a set of morphological categories: personal and possessive pronouns with the *lexical* categories of person, number and gender, inflectional adjectives with the *inflectional* categories of gender, number and case. A Czech possessive pronoun such as *jejího* 'her' is *lexically* 3rd person, singular and feminine, while *inflectionally* it is masculine or neuter, singular, genitive or accusative.[4] This is an additional motivation for the three-dimensional approach to word classes.

3 Building and Using the Hierarchy

The hierarchies are equivalent to concept lattices of Formal Concept Analysis (FCA).[5] FCA relates objects according to their attributes with *concept*s, each consisting of a set of objects and attributes as its extension and intension, respectively.

The first step is to identify objects and their attributes in a *formal context*. Table 1 is the formal context for our previous example of adjectives and numerals

[4]Czech personal and possessive pronouns share the same *lexical* categories and are distinguished by their *inflectional* category.

[5]For an overview of linguistic applications of FCA see [7]. [4] is concerned with a lexical interlingua, similar to our hierarchy of linguistic categories.

(Fig. 5). Attributes corresponding to the boxed labels in Fig. 5 are omitted: they would be specified for all objects and would not make the resulting lattice more informative. Next, a set of formal concepts is built. Objects belonging to a concept belong also to its superconcept and the concepts are partially ordered by specificity (roughly: the more attributes, the more specific). Finally, the concept lattice can be drawn (Fig. 6). Its geometry is significantly simpler than the hierarchy constructed intuitively (as in Fig. 5), but the concept ambiguous between adjectives and cardinal numerals is still there. The latter two steps can be done automatically.[6]

	ladj	*lnum*	*iadj*	*inoun*	*sadj*	*snoun*
adj	✓		✓		✓	
ord		✓	✓		✓	
card		✓		✓		✓

Table 1: Formal context for adjectives and ordinal numerals

1	⟨{adj,ord,card},	{}⟩
2	⟨{ord,card},	{*lnum*}⟩
2	⟨{adj,ord},	{*iadj*,*sadj*}⟩
3	⟨{adj},	{*ladj*,*iadj*,*sadj*}⟩
3	⟨{ord},	{*lnum*,*iadj*,*sadj*}⟩
3	⟨{card},	{*lnum*,*inoun*,*snoun*}⟩
4	⟨{},	{*ladj*,*lnum*,*iadj*,*inoun*,*sadj*,*snoun*}⟩

Table 2: Formal concepts derived from Table 1

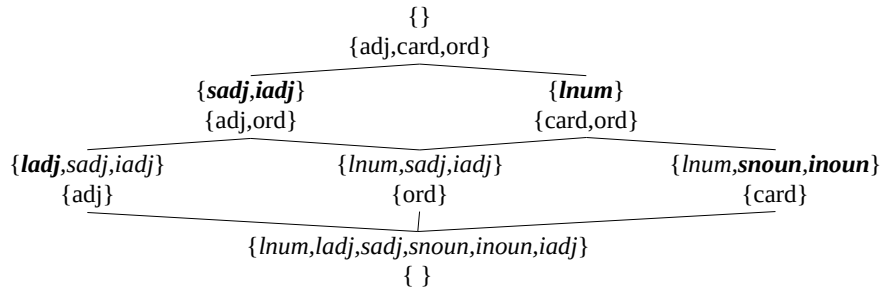

Figure 6: Concept lattice for adjectives and ordinal numerals

Attributes specified for an object in a formal context are interpreted in conjuction. Thus, specifying both *snoun* and *sadj* as attributes of interrogative pronoun (*intp*) would mean that it is syntactic noun and syntactic adjective at the same time. To model disjunction of attributes we have to introduce a more general attribute covering the two options. The formal context for numerals and pronouns is shown below in Table 3 and the corresponding lattice in Fig. 7.

[6] See http://www.fcahome.org.uk/fca.html.

	lnum	*lprn*	*inoun*	*iadj*	*snoun*	*sadj*	*snom*
card	✓		✓		✓		✓
ord	✓			✓		✓	✓
persp		✓	✓		✓		✓
possp		✓		✓		✓	✓
relp		✓		✓	✓		✓
intp		✓		✓			✓

Table 3: Formal context for numerals and pronouns

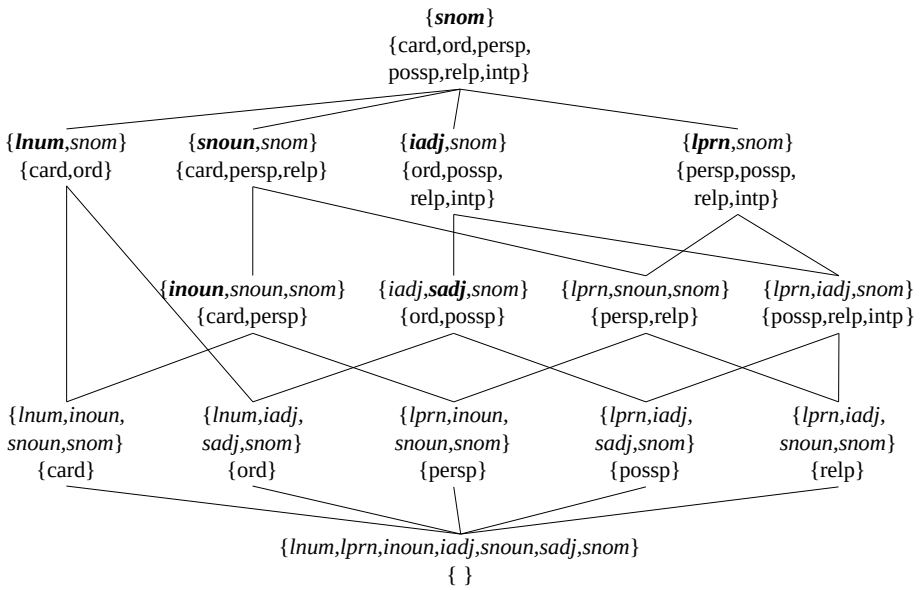

Figure 7: Concept lattice for numerals and pronouns

Lattices can be used for reasoning about attributes, as in the implications *ladj* ⇒ *sadj* or *snoun* ⇒ *lnum*, refering to Fig. 6. Such statements may help the user with language-independent category labels, or to match incompatible language-specific tags. The concept with the extension {ord} corresponds to `Nr`, the Czech tag for ordinal numerals, while the concept with the extension {adj,ord} corresponds to `ADJA`, the German tag covering adjectives and ordinal numerals. Its optimal Czech equivalent would be a Czech tag corresponding to the {adj,ord} concept. In the absence of such a tag, the more specific concepts are traversed and the disjuction of Czech tags corresponding to {adj} and {ord} is the result. Looking up a German equivalent of `Nr` is similar to the scenario when the user asks for "ord" in a German text. It is easy in a Czech text, because the appropriate tag `Nr` is available. For German, there is no tag corresponding to "ord". There are also no concepts more specific than {ord} that would correspond to German tags. The only option is to resort to a more general concept {adj,ord}, with the corresponding German tag.

The extensions of the two concepts can be compared and the user warned that she would have to filter out concordances including categories corresponding to “adj”.

This is a chance for a more data-driven approach to step in. If at least some of the word tokens tagged in the German corpus as ADJA are aligned with their Czech counterparts, the Czech word’s tag may decide whether the German word is a regular adjective or an ordinal numeral. In a multilingual corpus, multiple alignments can be used and a voting scenario applied. Then the hierarchy should decide what kinds of distinctions (i.e. what categories) are relevant for a given language, independently of its tagset.

It seems that incompatible tagsets may actually be useful; there are quite a few cases where projecting morphosyntactic tags in a language pair may bring mutual benefit. In 1.5 million word-to-word alignments extracted from the Czech-English part of *InterCorp*, more than 16.2% of 357 thousand Czech tokens tagged as nouns have their English equivalent tagged as proper noun, which is a category missing on the Czech side. Switching the direction, 85.3% of the total of 95 thousand Czech prepositions have as their English equivalent a token tagged by one of the two highly ambiguous tags: `IN` as preposition/subordinating conjuction or `TO` as preposition/infinitival particle *to*. In 2 million Czech-Polish pairs, 67.2% of 197 thousand Czech tokens tagged as pronouns of different types are likely to have pronominal Polish equivalents, tagged by their *inflectional* class, mostly adjectival or nominal. This opens up the option to project their Czech *lexical* class, although pronouns as a closed class category could be identified as lexemes. The other direction may be more attractive – some Czech pronominal tags are underspecified along the inflectional and syntactic dimensions, which is precisely the information offered by their Polish counterparts. Czech demonstrative and indefinite pronouns (about 31.9% of the total number of Czech pronouns) can thus be identified as attributive or substantive.

4 Conclusion

As a solution to the issue of tagset variety in multilingual corpora we have proposed an abstract interlingual hierarchy of categories, based on a three-way distinction in the system of word classes. In addition to intuitive and underspecified queries and principled mappings between different language-specific tagsets, the hierarchy can be used to refine morphosyntactic annotation in word-aligned parallel corpora by learning from more specifically tagged word tokens in other languages.

If corpus data include only original, language-specific tags, the system can be easily modified and extended without touching the corpus data and the abstract categories can be mapped to tags in any format. Formal Concept Analysis is the answer to concerns about the costs of designing the hierarchy.

The abstract hierarchy is currently built for languages equipped with morphosyntactic annotation and represented in the *InterCorp* project. The work is based on available documentation, annotations actually produced by the taggers,

and previous work, mainly the results of the *Intertag* project. Experiments aiming at the refinement of morphosyntactic annotation by projecting information using word-to-word alignment bring positive results and may be useful even for untagged texts. Although a proper evaluation has not been done yet, it is obvious that incompatible tagsets can actually complement each other and have synergic effects.

References

[1] Tomaž Erjavec. MULTEXT-East Morphosyntactic Specifications: Towards Version 4. In Radovan Garabík, editor, *Metalanguage and Encoding Scheme Design for Digital Lexicography*, pages 59–70, Bratislava, April 2009. Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences.

[2] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis. Mathematical Foundations*. Springer, Berlin/Heidelberg, 1999.

[3] Jan Hajič. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Charles University Press, Prague, 2004.

[4] Maarten Janssen. Multilingual Lexical Databases, Lexical Gaps, and SIMuLLDA. *International Journal of Lexicography*, 17(2), 2004.

[5] Miroslav Komárek. Autosemantic Parts of Speech in Czech. In *Travaux du Cercle linguistique de Prague*, volume 3, pages 195–210. 1999.

[6] Natalia Kotsyba, Adam Radziszewski, and Ivan Derzhanski. Integrating the Polish Language into the MULTEXT-East Family: Morphosyntactic Specifications, Converter, Lexicon and Corpus. In *Proceedings of Research Infrastructure for Digital Lexicography: MONDILEX Fifth Open Workshop*, pages 37–55, Ljubjana, Slovenia, 2009.

[7] Uta Priss. Linguistic Applications of Formal Concept Analysis. In Bernhard Ganter, editor, *Formal Concept Analysis. Foundations and Applications*, volume 3626 of *Lecture Notes in Artificial Intelligence*, pages 149–160. Springer, Berlin/Heidelberg, 2005.

[8] Adam Przepiórkowski and Marcin Woliński. A flexemic tagset for Polish. In *Proceedings of* Morphological Processing of Slavic Languages, *EACL 2003*, 2003.

[9] Daniel Zeman. Hard Problems of Tagset Conversion. In Alex Fang, Nancy Ide, and Jonathan Webster, editors, *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, pages 181–185, Hong Kong, China, 2010.