Morphological Tags in Parallel Corpora*

Alexandr Rosen

1. Introduction

Multilingual parallel corpora can be annotated with morphosyntactic tags by monolingual tools, freely available for a number of different languages. However, each of the tools is typically bundled with a specific tagset and assumes a specific method of tokenization. The variety of tagging schemes and tag formats may be a problem for the user; a relatively simple tag query in a multilingual setting often means spending a while with tagset manuals.

The aim of the present contribution is to suggest a solution that would delegate the task of dealing with multiple tagsets to the system. The core component of the proposal can be viewed as an abstract interlingual tagset. It is actually a hierarchy of linguistic categories, partially ordered by their specificity, mapped to tags in language-specific tagsets. In order to capture different views of word classes, as seen by the tagsets, the common tagset takes three different perspectives of word class: lexical, inflectional and syntactic, each potentially coupled with its own set of morphological categories. 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.

The common tagset is formalized as a tangled hierarchy of types, each of the types corresponding to a linguistic category and some of the types to one or more language-specific tags. Tags in all tagsets can be described as objects with properties such as lexical, inflectional and syntactic word class, and the relevant morphological categories. Then the standard methods of Formal Concept Analysis (Ganter & Wille, 1999) can be used to construct the hierarchy automatically as a concept lattice and to (partially) resolve tag queries that do not quite match the tags used for the specific language in a way similar to that used by Janssen (2004) for dealing with lexical gaps in the multilingual lexical database.

Language-specific subsets of the abstract common tagset can be extracted using the links to tags in language-specific tagsets. Abstract language-specific tagsets can be used to generate or interpret tags in a format of the user's or a tool's preference. In addition, the modular setup allows for underspecified tag queries and for mappings between tagsets at the minimal information loss or distortion possible, even in cases of highly ambiguous and overlapping tags.

The rest of this section includes more motivation for this enterprise. In section 2, some related work is briefly reviewed, including a related *Interset* project, the leading contender among potential partners. Section 3 focuses on some problems arising in confrontation of multiple annotation schemes. In section 4, the proposed solution is presented in more detail, using a few examples. Some concluding remarks are included in section 5.

1.1. Why tags

Morphosyntactic tags provide information about part of speech and morphological categories for each word in the corpus according to lexical properties of the word, its form and syntactic

* The author is grateful for many helpful comments on the previous versions of this text to Patrick Corness, Jiří Hana, Daniel Zeman, Jarmila Panevová and Vladimír Petkevič, and for inspiring feedback to the kind audiences at the *InterCorp* conference on 17–19 September 2009 in Prague and at the Czech Day workshop on 26 November 2009 in Regensburg (at the Regensburg University's Institute for Slavic Studies). All remaining faults are the author's responsibility. Work on this project was supported by grant no. MSM0021620823 of the Czech Ministry of Education, Youth and Sports.

context. This information comes at a cost; even when assigned or checked manually, tags are not 100% reliable. This is even more true about tags assigned automatically by a *tagger*, a typical scenario for modern corpora of any practical size.

However, the error rate is acceptable for most uses. Some corpus queries could not even be made without tags; other queries can be specified more easily and most machine-learning tasks would not be possible without tags. Therefore, we assume that even imperfect tags are better than no tags.

1.2. Too many tagsets

Each tagger and each language usually comes with a tagset of its own due to the differences in languages, underlying theories, the authors' viewpoints and preferences, and intended usage. Conceptually different tagsets exist even for one language or closely related languages. For Czech, there are at least three tagsets that could be considered as candidates for a new project, each with its own set of tools and resources: the *MULTEXT* tagset, the "Brno" tagset (Osolsobě et al., 2006), and the "Prague" tagset (Hajič, 2004), used for tagging Czech in the *InterCorp* project.[1] The tagset variety is apparent also in the currently available set of tagged texts in the *InterCorp* project; texts in 11 out of the total of 22 currently accessible languages are tagged with 11 different tagsets (while texts in 8 of those 11 languages are lemmatized). Some tagsets obey at least a similar design principle (Bulgarian and Russian, French and Italian), but most of them present a strikingly different picture at first glance (Czech, English, Dutch, German, Hungarian, Polish and Spanish). The differences are not only formal; even when tags seem to be identical or similar across the languages, they have often mismatching or overlapping denotations, and the situation is bound to grow even more complex with more tagsets to come, often fairly extensive, such as those for South Slavic and Baltic languages. Fig. 1 illustrates the tagset variety using comparable examples of prepositional phrases in all of the 11 presently tagged languages.[2]

en	in	the	remotest	exurbs
	IN	**DT**	**JJS**	**NNS**
de	in	den	abgelegensten	Außenbezirken
	APPR	**ART**	**ADJA**	**NN**
nl	in	dit	schitterende	appartement
	600	**370**	**103**	**000**
fr	dans	les	plus lointaines	banlieues
	PRP	**DET:ART**	**ADV ADJ**	**NOM**
sp	en	las	zonas	más remotas
	PREP	**ART**	**NC**	**ADV ADJ**
it	da	queste	lingue	babeliche
	PRE	**PRO:demo**	**NOM**	**ADJ**
ru	v	samych	otdaljonnych	rajonach

[1] The three Czech tagsets are documented at http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html, http://nlp.fi.muni.cz/projekty/ajka/tags.pdf and http://www.korpus.cz/orwell_znacky.php. A help application for the Prague positional tagset is available at http://utkl.ff.cuni.cz/~skoumal/morfo/?lang=en. The corresponding morphological analysers are available on line at http://nlp.cs.jhu.edu/~hajic/morph.html and http://nlp.fi.muni.cz/projekty/wwwajka/WwwAjkaSkripty/morph.cgi?jazyk=0.

[2] Details about the tagsets are available with other information about the parallel corpus project *InterCorp* at http://korpus.cz/english/intercorp-info.php. 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/). Here and below, any unused positions at the end of a Czech tag, consisting of a sequence of 15 characters are truncated: **RR--6** stands for **RR--6----------** (tag for a preposition selecting local case).

	Sp-l	P---pl	Afp-plf	Ncmpln
cs	v	těch	nejodlehlejších	zástavbách
	RR--6	**PDXP6**	**AAFP6----3A**	**NNFP6-----A**
bg	na	tova	prijatelsko	dviženie
	R	**Pde-os-n**	**Ansi**	**Ncnsi**
pl	w	tym	wspaniałym	apartamencie
	prep:loc:nwok	**adj:sg:loc:m3:pos**	**adj:sg:loc:m3:pos**	**subst:sg:loc:m3**
hu	a	szép	katalán	lányba
	ART	**ADJ**	**ADJ**	**NOUN(CAS(ILL))**

Figure 1: Differences in tagging: prepositional phrases

1.3. Any solution?

The rest of this contribution is based on the assumption that tagset variety is a problem. This assumption may be questioned by users who are already familiar with the tagsets they need, find them easy to learn or look up, or who do not (intend to) use tags at all. On the other hand, others may object that tags should be easy to read and write, and that the user should not be expected to study lengthy manuals to make a simple query. In the following, we are going to explore options to satisfy the latter group of corpus users, while trying to make sure that the solution will not cause more problems than it is supposed to solve. Ideally, a common tagset should be used, although language-specific tags may be an option for very parochial categories present only in a single language. At the same time, the tagset should be well defined, both formally and conceptually; the same concept should be expressed in the same way and different concepts should be expressed in different ways across all languages.

2. Related work

There have been efforts to propose tagsets common to more languages, standards for tagset design, or common frameworks for representing grammatical categories. Such proposals include a set of standard abbreviations for linguistic terms *Eurotyp* (König et al., 1993); guidelines for morphosyntactic annotation of West European languages, including an *intermediate tagset*, one of the results of the *EAGLES* project (Leech & Wilson, 1996);[3] a common tagset for 8 languages from 2 language families of the Indian peninsula (Baskaran et al., 2008); General Ontology for Linguistic Description (*GOLD*), with a stress on endangered languages (Farrar & Langendoen, 2003); and an ontology derived from *EAGLES* and other tagsets, harmonized with *GOLD* and linked to language-specific tagsets (Chiarcos, 2008). Hughes et al. (1995) implemented mapping rules among different English tagsets as an interface to *AMALGAM*, an on-line English tagger with the choice of 8 tagsets.[4]

Common (or *harmonized*) tagsets have been designed to tag parallel corpora consisting of a number of languages, such as the *LE-PAROLE* project, a multilingual corpus of 14 European languages.[5] Another major project of this sort deserves special attention, because its spin-off included Czech and other highly inflectional languages – see 2.1 below on the *MULTEXT* project.

While considering options to make corpus searching easier, one should not miss *Poliqarp*, a (monolingual) corpus manager offering a very intuitive format for tag queries

[3] *EAGLES* stands for Expert Advisory Group on Language Engineering Standards, active in the 1990s. The tagset intentionally avoids mnemonics reminiscent of language-specific terminology: **V0002500100000** stands for main verb infinitive. See http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html.

[4] http://www.comp.leeds.ac.uk/amalgam/

[5] http://www.elda.org/catalogue/en/text/doc/parole.html

(Przepiórkowski et al., 2004).[6] Both positional and attributive specifications of a tag query are supported, and the option of specifying aliases makes querying even easier.[7] Furthermore, a sophisticated set of operators can be used to search for (un)ambiguously tagged tokens, including the values of their morphological categories before disambiguation. Given our concern about the ease of making tag queries, all these features would make *Poliqarp* an attractive choice, were it not for the fact that ours is a parallel corpus and our common tagset is designed as an abstract structure, not physically present in the data.

2.1. MULTEXT and MULTEXT-East

In *MULTEXT*, a project aimed at creating multilingual tools and resources, Ide & Véronis (1996) designed a common tagset for six West European languages (Dutch, English, French, German, Italian and Spanish). The tagset, based on recommendations of the *EAGLES* group, distinguishes general and specific features using positional tags, corresponding to sets of attribute-value pairs, see fig. 2.

$$\texttt{Ncms:} \begin{bmatrix} \text{category} & \text{noun} \\ \text{type} & \text{common} \\ \text{gender} & \text{masculine} \\ \text{number} & \text{singular} \end{bmatrix}$$

Figure 2: A *MULTEXT* tag and its corresponding set of attribute-value pairs

MULTEXT-East (*MTE*) followed the same basic concept as *MULTEXT* for East European languages. Its first results, published in 1998, included six languages in addition to English: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene, and in case of some of the languages it was the first project to introduce morphosyntactic tags (Erjavec, 2001; Erjavec et al., 2003). The common tagset consists of 12 major word classes and 100 attributes with 500 values. The project is still very much alive; its current release 4 includes 13 languages, conforms to the TEI P5 standard and offers XSLT stylesheets to derive simpler language-specific tags from the common tagset (Erjavec, 2009).

The *MTE* tagset has been criticized for an occasional lack of consistency (Przepiórkowski & Woliński, 2003; Derzhanski & Kotsyba, 2009; Feldman & Hana, 2010). Different tags are used for the same phenomenon: attributive participles are treated as verb forms in Bulgarian and as adjectives in six other Slavic languages, the tags for adverbial participles (transgressive) in Czech and Slovak are different from those annotating equivalent forms in Bulgarian and Serbian. Long and short forms of personal pronouns are tagged as such only in Romanian, although they do exist in other languages; similarly with negative adverbs. Some tags are too specific and hard to extend to cover similar phenomena in another language: Czech enclitic *s* is tagged as a binary feature **Clitic_s** on verbs and pronouns, without provision for the Polish agglutinative auxiliary that occurs also in first person (*em*).[8] More generally, the tagset misses some correspondences across the languages, such as the relation of

[6] http://poliqarp.sourceforge.net/, http://korpus.pl/

[7] Aliases are abbreviations for alternative values of an attribute; e.g. **masc** stands for any of the three masculine genders in Polish, **noun** for any of the six tags representing word classes in a nominal syntactic position (lexical nouns, deverbative nouns and pronouns), **verb** for any of the 13 verbal tags, both finite and non-finite.

[8] Polish was beyond the scope of the original *MTE* tagset and the project had no ambition to design a tagset that would cover more languages. (V. Petkevič, p.c.)

the two morphological cases in Romanian (direct and oblique) with their more specific counterparts in other languages.

Despite these drawbacks, the project provides an important reference point and a possible starting point.[9] A common tagset designed for Ukrainian and Polish, originally based on the Polish IPI PAN tagset (Kotsyba et al., 2008), was later brought in line with *MTE* (Derzhanski & Kotsyba, 2009); the Bulgarian *MTE* tagset has been compared with those for Slovak (Dimitrova et al., 2009b) and Polish (Dimitrova et al., 2009a), the treatment of predicatives in Russian, Ukrainian, Polish and Bulgarian has been examined vis-à-vis the *MTE* tagset (Derzhanski & Kotsyba, 2008) and modifications to the existing *MTE* tagset proposed (Derzhanski & Kotsyba, 2009).

2.2. Interset

Interset is an "interlingual" tagset (Zeman, 2008), designed primarily for translating tags from one tagset into another. Indeed, if the task is to convert between multiple tagsets, an intermediate tagset saves the effort of compiling a high number of pairwise mappings. Mappings to and from the intermediate tagset are implemented as freely available "decoding" and "encoding" modules. Users are invited to define new mappings by contributing to the project site.[10]

Some issues concerning the design of *Interset* are reminiscent of the question whether it is possible to design *interlingua*, an intermediary language for machine translation. Such a language is expected to capture all meanings expressed in all languages (or at least in all those involved, preserving distinctions at levels as general or as specific as needed. A major objection raised against *interlingua* concerns the fact that languages tend to view and structure the world in words from different, mutually incompatible angles. Luckily, *Interset* is faced with a much simpler task, where "languages" of tagsets, classifying grammatical categories, are much simpler and better defined, and the strategy of incorporating every distinction into the tagset interlingua is viable even with the individual tagsets taking different, but linguistically motivated, viewpoints.

Interset is constructed "bottom-up" by successively integrating distinctions present in newly added tagsets. The distinctions are represented as attribute-value pairs; a tag is typically mapped onto a set of such pairs, viewed as an abstract object rather than as a physical tag to be used for tagging real data. The number of integrated tagsets is currently 12 for 10 languages (Arabic, Bulgarian, Chinese, Czech, Danish, English, German, Polish, Portuguese and Swedish), plus 6 task-specific tagsets.[11]

Decoding (of a tag into *Interset*) is easier than encoding (into a tag); the abstract tagset can be extended in case a distinction present in the source tagset is missing in *Interset*, but a missing distinction or a combination of distinctions within a tag in the target tagset requires a non-trivial solution. Admissibility of a tag is checked by an exhaustive list of tags and problems are resolved by a precedence list of features and defaults for illicit values. Most common classification problems involve particles, pronouns, *wh*-words, determiners and participles. In the tag translation process, information can only be lost, not added; ordinal numerals tagged only as adjectives cannot be translated as numerals in the target set. Information relevant only within a single specific tagset is preserved in the original tag.

---

[9] *MTE* tagset is used in another multilingual project involving Slavic languages: http://www.mondilex.org/.

[10] https://wiki.ufal.ms.mff.cuni.cz/user:zeman:interset

[11] The task-specific tagsets are modifications of Czech, English and German tagsets, intended for "shared tasks" competitions in dependency parsing, organized by the Conference on Computational Natural Language Learning (CoNLL) in 2006 and 2009.

The design policy of *Interset* is based on recommendations rather than requirements (**verbform=participle** can be **pos=verb** or **pos=adj**), but some decisions should still be applied consistently; pronouns and determiners are not treated as major word classes. Instead, nouns, adjectives and adverbs can be specified as one of the pronominal types (personal, demonstrative, interrogative) and all determiners are treated as a type of adjective.[12]

Interset is unique among other projects in the variety of languages it handles, a fact that is reflected in the design and scope of the abstract tagset. Although *Interset* aims at a slightly different goal, we see this project as a welcome companion of the proposed *InterCorp* common tagset, mainly due to the possibility of integrating the available *Interset* mappings and the envisaged ease of linking categorial distinctions made in both systems.

Interset does not offer mapping between the Prague and Brno tagsets yet. This mapping is provided by a tool called 'morphological converter' – *MorphCon*[13] (Pořízka & Schäfer, 2009). The tool uses *Interset* and its drivers to convert tags between tagsets, including tags embedded in texts.

3. Problems

3.1. Arbitrary choices

The choice of a specific set of tags often involves a number of arbitrary decisions, as attested in incompatible tagsets even for a single language. English tagset designers and/or taggers have to decide whether *dancing* in an attributive position is a noun, adjective or gerund, whether *a*(*n*) is an article or – more generally – a determiner, whether *often* is adverb or qualifier. In some cases the decision preserves ambiguity; *to* is not distinguished as preposition or infinitival particle in the Penn Treebank tagset, a vertical bar is used to tag words in consistently ambiguous positions: JJ|VBG reflects the two readings of *entertaining* in *The duchess was entertaining last night*.[14]

In Czech, forms such as *udělána* 'done' are analysed differently by Ajka,[15] a morphological analyser based on the Brno tagset, and by a similar tool using the Prague tagset.[16] Ajka treats the form as either participle or adjective, assigning six possible tags, two for participle (feminine singular **k5eAaPmNgFnS**, or neuter plural **k5eAaPmNgNnP**) and four for adjective (feminine singular nominative **k2eAgFnSc1d1**, masculine singular accusative **k2eAgMnSc4d1**, neuter plural nominative or accusative – **k2eAgNnPc1d1** or **k2eAgNnPc1d1**). The Prague tool assigns just one tag for participle feminine singular or neuter plural (**VsQW---XX-AP**).[17]

---

[12] According to the list of "common problems" at https://wiki.ufal.ms.mff.cuni.cz/user:zeman:interset, tagsets often disagree on the status of pronouns and determiners: determiners themselves are sometimes treated as demonstrative pronouns. The *Interset* solution is not far from a consistent cross-classification of word classes, a solution adopted in 4.2 below.

[13] http://morphcon.webnode.cz/

[14] The vertical bar is also used to separate ambiguous lemmas. The taggers for German, French and Italian return the following results (shown as *form*/**lemma(s)**/**tag**): *maßt*/**maßen**|**messen**/**VVFIN** (finite content verb), *überdachte*/**überdachen**|**überdenken**/**VVFIN** or *überdachte*/**überdacht**/**ADJA** (attributive adjective), *Symposien*/**Symposion**|**Symposium**/**NN** (noun), *crus*/**croire**|**croître**/**VER:pper** (past participle), *crûmes*/**croire**|**croître**/**VER:simp** (simple past), *compiamo*/**compiere**|**compire**/**VER:cpre** (present subjunctive).

[15] http://nlp.fi.muni.cz/projekty/ajka/ajkacz.htm

[16] http://quest.ms.mff.cuni.cz/morph/

[17] Note the use of ambiguous gender and number specifications (**QW**), another example of ambiguity preservation. These cryptic values will be replaced by disambiguated specifications in the foreseeable future. (V. Petkevič, p.c.)

On the other hand, *udělánu* (also 'done') is analysed unambiguously by both tools, albeit in different ways, as adjective feminine singular accusative (**k2eAgFnSc4d1**) or as participle feminine singular accusative (**VsFS4--XX-AP**).

3.2. Different concepts of word classes

In contrast to the Czech tagsets, distinctions in the Polish IPI PAN tagset are based on inflectional classes (Przepiórkowski & Woliński, 2003). Thus the two tagsets, designed for the two closely related languages, have a very different concept of word class, with the Czech tagset closer to the traditional view and mostly more fine-grained and the Polish tagset better defined but lacking some distinctions.[18]

A Polish adjective (*dziewiąta*/**adj:sg:nom:f:pos** 'ninth') may correspond to a Czech ordinal numeral (*devátá*/**CrFS1** 'ninth'), possessive pronoun (*swoje*/**adj:pl:acc:m3:pos** – *svoje*/**P8XP4** 'his/her/its/their'), demonstrative pronoun (*temu*/**adj:sg:dat:m1:pos** – *tomu*/**PDZS3** 'that'), or relative pronoun (*który*/**adj:sg:nom:m1:pos** – *který*/**P4YS1** 'which'). A Polish tag for non-inflected words may correspond to a Czech tag for particles (*nie*/**qub** *tylko*/**qub** – *ne*/**TT** *jen*/**TT** 'not only'), non-gradable adverbs (*wtedy*/**qub** – *tenkrát*/**Db** 'then'), reflexive pronouns (*się*/**qub** – *se*/**P7-X4** 'himself/herself/itself/themselves'), subordinating conjunctions (*kiedy*/**qub** – *když*/**J,** 'when'), or coordinating conjunctions (*czy*/**qub** – *nebo*/**J^** 'or').

Some categorial distinctions are ignored or reflected only implicitly in the tagset. The Prague tagset implicitly marks reflexivity in personal pronouns such as *sobě* 'himself/herself/itself/themselves' (**P6-X3**) and reflexivity plus possessivity in possessive pronouns such as *svůj* 'his/her/its/their' (**P8IS1**), while the Polish IPI PAN tagset treats the corresponding forms either as a specific class – **siebie:dat** for *sobie* 'himself/herself/itself/themselves' – or as a syntactic word class – **adj:sg:nom:m1:pos** for *swój* 'his/her/its/their'.

3.3. Tokenization

The Spanish tagger tags and lemmatizes many multi-word units as a single item: *Estados Unidos*, *al mismo tiempo*, *en lugar de*, *tendrán que* (*tener que*/**VMfin**).[19]

Hyphenated compounds are treated as a single unit in Bulgarian (*Avstro-ungarski*/**A-pi**), Dutch (*Frans-Duitse*/**103**), English (*Franco-German*/**NP**), French (*franco-allemande*/**ADJ**), German (*deutsch-französisch*/**ADJA**), Italian (*franco-tedesco*/**ADJ**), and Spanish (*franco-alemana*/**NC**), but not in Czech (*francouzsko*/**A2--------A** + **-**/**Z:** + *německý*/**AANS3----1A**), Hungarian (*angol*/**ADJ** + **-**/**PUNCT** + *japán*/**ADJ**), Polish (*niemiecko*/**adja** + **-**/**interp** + *rosyjski*/**adj:sg:nom:m1:pos**) and Russian (*franko*/**Ncmsny** + **-**/**-** + *germanskij*/**Afpmsnf**).[20]

---

[18] The original Polish tagset has been slightly modified for the Polish National Corpus – see Przepiórkowski (2009).

[19] Unfortunately, the original orthographic words are ignored by the present version of the *InterCorp* search engine (http://korpus.cz/Park, available to registered users of the Czech National Corpus). A query specifying *al mismo tiempo* as a phrase returns zero hits: the three words are treated as a single word form, which also means that a naive search for the form *mismo* will not return concordances including the multi-word unit *al mismo tiempo*. Ignoring such cases in queries may seriously distort results; the currently available Spanish part of *InterCorp*, consisting of 8.4 million words, includes 83 thousand multi-word tokens of 323 different types. Here corpus annotation obscures the original text, which is certainly unfortunate.

[20] A caveat parallel to that in footnote 19 is due−a split compound can be searched using only its parts in the query, or as a phrase with the parts separated by blanks, i.e. "**česko - německý**". The present version of the corpus manager will not return any result when a split compound is queried as a form or a phrase without blanks in between. The opposite is true of compounds that are not split −they cannot be queried using only their parts in

Within a language, the treatment of hyphenation is fairly consistent. The German and French taggers prefer not to split: *Jelzin-Ära*/**NN**, *gut-ausgearbeiteten*/**ADJA**, *cure-dents*/**NOM**, unlike the Czech tagger: *padne*/**VB-S---3P-AA** + -/**Z:** + *li*/**TT**, *Tchaj*/**AAXXX----1A** + -/**Z:** + *wanu*/**NNIS2-----A**. Yet care must be taken in specific cases, as in the following German and French examples: *Rechts*-/**TRUNC** *und*/**KON** *Entwicklungsbewegung*/**NN**, *dit*/**VER:pres** + *-il*/**PRO:PER**. Fig. 3 gives more hints concerning French tokenization.

form	**lemma**	**tag**
n’	ne	ADV
avaient	avoir	VER:impf
-ils	il	PRO:PER
jusqu’	jusque	PRP
au	au	PRP:det
La	le	DET:ART
compassion	compassion	NOM
,	,	PUN
c’	ce	PRO:DEM
est	être	VER:pres
d’abord	d’abord	ADV
l’	le	DET:ART
oubli	oubli	NOM
de	de	PRP
soi	soi	PRO:PER
,	,	PUN
répliqua	répliquer	VER:simp
-t-il	il	PRO:PER
sèchement	sèchement	ADV
.	.	SENT
Ne	Ne	VER:futu
l’	la|le	PRO:PER
aurait	avoir	VER:cond
-il	il	PRO:PER
pas	pas	ADV

Figure 3: Examples of French tokenization

Tokenization of strings including an apostrophe may not be straightforward either: *children*/**NNS** + *’s*/**POS**, *parents*/**NNS** + *’*/**POS**, I/**PP** + *’m*/**VBP**, *ca*/**MD** + *n’t*/**RB**.

In some cases, even contiguous strings of alphabetic characters are split and each part is assigned a tag and lemma of its own. This is what happens to Polish (orthographic) words with the agglutinative auxiliary attached, as in *zrobiłeś* ‘(you) made’: *zrobił*/**zrobić**/**praet:sg:m1:perf** + *eś*/**być**/**aglt:sg:sec:imperf:wok**. A single orthographic word such as *żebyśmy* ‘that we would’ is split into three parts: *że*/**że**/**conj** + *by*/**by**/**qub** + *śmy*/**być**/**aglt:pl:pri:imperf:nwok**.[21]

On the other hand, Czech enclitic *s* as a second person singular auxiliary, spelt together with the preceding form, is treated in the Prague tagset on a par with inflectional endings. An

---

the query, but they will be found when specified as a single form (or a ‘single-word’ phrase without blanks in between).

[21] A single orthographic word can have different interpretations depending on the way it is tokenized. The form *miałem* can be tagged either as **miał**/**subst:sg:inst:m3** ‘dust’ or **mieć**/**praet:sg:m1:imperf** + **być**/**aglt:sg:pri:imperf:wok** ‘had’. Similarly with *gdzieś*: **gdzieś**/**qub** ‘somewhere’ or **gdzie**/**qub** + **być**/**aglt:sg:sec:imperf:nwok** ‘where have (you been)’. Unfortunately, the tagger’s choice is not reliable and the present version of the corpus manager cannot see the original orthographic words. This means that searching for such words may involve more than one attempt – a query for its non-split version and another one for its split version.

orthographic concatenation of an l-participle with enclitic auxiliary *udělals* '(you) made' is tagged as a single form of the l-participle **udělat/VpYS---2R-AA** (2nd person singular masculine, past tense, affirmative, active voice). The complementizer + enclitic auxiliary *žes* 'that (you) are' is tagged as subordinate conjunction in 2nd person singular (**J,-S---2**). However, the second person singular pronoun *ty* is specified for person even without the clitic (**PP-S1--2**), so the form with the clitic attached is distinguished by additional specifications for tense, polarity and voice, irrelevant for either the pronoun or the clitic auxiliary (**PP-S1--2P-AA**). German and French contractions of preposition and article (*zum*, *aux*) are similar examples of the same phenomenon.

In order to find as many equivalent tags among different tagsets as possible and to avoid postulating items such as a conjugated conjunction, our preference is to tag the least common denominators, the minimal tokens, i.e. parts of the compound or agglutinated forms, rather than design tags that would tag them as a whole, the whole often consisting of categorially distinct parts. Since tokenization is often part and parcel of the tagging procedure and related to the language-specific tagset, we need an option to virtually re-segment a token and assign a tag (and lemma) to each of its parts, or at least an option to assign a sequence of lemma/tag pairs to a single token. Ideally, both options should be available for each case of mismatch between orthographic and "syntactic" words at the same time, depending on the user's or tool's preference, or the form of the query, as in the concordancer *Poliqarp*, used in Polish and Portuguese corpus projects (http://korpus.pl and http://nkjp.pl).[22]

4. A proposal

4.1. Options

To ease the problem of many partially incompatible tagsets two alternative solutions are at hand. A "foreign" tagset can be converted into another, more familiar, already existing tagset. This approach has the advantage in that the user is not faced with an additional tagging scheme. However, for a larger number of tagsets of tagsets, conversion via a common tagset is a better solution anyway (see 2.2). Therefore we can see the problem as an opportunity to design a common tagset that would be useful not only for negotiating conversions between language-specific tagsets, but also to simplify tag-based corpus queries and understanding of concordances with tags displayed.

The first candidate to consider as our common tagset should be an existing tagset, such as *MTE* or *Interset*. However, they were created for somewhat different purposes, and adopting them without modifications would defeat some of our objectives, such as consistency, formal well-definedness and user-friendliness – see 2.1 and 2.2. Furthermore, restrictions inherent in the available language-specific tagsets prohibit the use of a ready-made linguistic ontology.

The present section deals with the question of an optimal design of the common tagset, both from a linguistic/conceptual and a formal/technical viewpoint. Technically, a common tagset may be designed as the union of all distinctions in all the language-specific tagsets involved, or as a mere intersection of such distinctions. With typologically distant languages, the latter option would produce a very restricted tagset, while the former approach runs the risk of overloading the tagset with many parochial distinctions. We assume that the common tagset should capture as many distinctions as possible, but may ignore a few exotic distinctions peculiar to a single tagset.

[22] A Polish agglutinated form such as *zrobiłeś* '(you) made' will be found no matter whether a word form query is specified as **zrobiłeś**, **zrobil**, or **eś**.

Another choice concerns a more conceptual aspect of the strategy. A common tagset can be built "bottom-up," in a purely formal fashion, merging explicitly identical tags and preserving explicitly distinct ones, or "top-down," in a linguistically motivated way, based on what underlies the tag distinctions. By making the conceptual structure an explicit design principle, the top-down approach runs a lower risk of different tags representing the same category and the same tag representing different categories.

Distinctions absent in a language-specific tagset cannot be reflected in the mapping to the common tagset,[23] but uncritical adherence to language-specific tagsets should be avoided, because they may treat equivalent linguistic categories in different, mutually incompatible ways. The common tagset should be built in a theoretically neutral, yet linguistically motivated way (as far as this is attainable), its distinctions precisely defined and realistic with a view of the available tagsets.

But how can we discover a correspondence between two tags from different tagsets, when their names offer no clue? For example, how can we identify direct case in Romanian as the equivalent of the disjunction of nominative and accusative cases in other languages? Unfortunately, this task requires understanding of the concepts underlying the tagsets. On the other hand, all tagsets deal with the same issue of classifying word forms, so the underlying concepts have some common denominator. They may differ in viewpoints and granularity, but they could be mapped onto an abstract hierarchy using cross-classification along different aspects. Any tag can be construed as an object having a number of properties potentially relevant outside the given tagset, and then the issue is to properly identify each tag as having the universal properties. More will be said about the process of designing the common tagset as an abstract hierarchy in 4.4.

There are also various types of tag format – positional (as in the Prague tagset), attributive (as in *Interset*, or – more compactly – in the Brno tagset), or type-dependent positional (as in the *MULTEXT* tagset). Our preference is to modularize the tagset specification in a way that would allow for any of these formats to be used interchangeably, be it in queries, in rendering query results, or in the corpus data. The common abstract tagset can be used as a source to derive tagsets for various languages, formatted in an arbitrary way.

In any case, the original tags should be preserved in the data. The common tags can either be added to every word token in the corpus or translated back and forth on the fly after specifying a query and before presenting results. Then the common tagset can be a truly abstract structure, mediating between a set of intuitive categories, available in the search interface, and the language-specific tagsets.

4.2. Three flavours of word class

Ever since a grammar of Greek attributed to Dionysius Thrax was written in the 2nd century BCE, a mix of morphological, syntactic and semantic criteria has been used to define the traditional list of seven word classes (later extended to eight). As Komárek (2006) points out, for some word classes (nouns, adjectives, prepositions, conjunctions, finite verbs – if treated as a distinct word class) the three criteria coincide; all point to the same word class. Nouns decline independently in typical nominal positions as subjects, objects, nominal predicates, non-agreeing attributes and adverbials, typically referring to entities; adjectives decline in agreement with a noun as attributes or predicates, representing properties. On the other hand, numerals and pronouns offer a completely different picture; their class membership is justified solely by semantic criteria, while syntactic and morphological behaviour of ordinals and

[23] Although it might be possible to derive a more adequate tag by comparing results of more than one tagger using different tagsets.

cardinals, personal and possessive pronouns cannot be described as that of pronouns or numerals, but rather as that of nouns or adjectives.

Komárek (2006, p. 14–15) suggests the option of abandoning the traditional list in favour of a cross-classification along the three dimensions, but rejects it as superficial and destructive, in our view without persuasive arguments. Without speculating about its explanatory merits, we assume that the three-angled view of word classes correctly describes their behaviour, is very useful for defining their properties and allows for relating word classes defined by various criteria in language-specific tagsets to more or less specific categories, potentially cross-classified along up to three dimensions.

Distinctions between the three aspects are borne out also by the tagsets. The Prague tagset for Czech has a preference for lexically-based classification, the Polish tagset for inflectional word classes, the German tagset distinguishes substitutive (nominal), attributive and – sometimes – adverbial use of interrogative, relative, demonstrative, indefinite and possessive pronouns.

The hierarchical structure in fig. 4 shows a simple case – nouns and adjectives are nouns and adjectives, respectively, on all three criteria.[24] The topmost node *wcl* stands for both objects, i.e. for the noun and the adjective. Its daughters are labelled by the three aspects: *lexical* (for 'semantic'), *inflectional* (for 'morphological') and *syntactic*.[25] The boxes around the labels are supposed to suggest that the sets of objects denoted by the nodes have a non-empty intersection, i.e. that they do not partition the set of objects denoted by the mother node. In fact, all the four sets involved are identical. This is precisely what cross-classification requires. The daughters of the nodes labelled by aspects are the word classes in the three respective flavours, distinguished in their labels by the initial letter. The six types of word classes share only two daughters, the objects to be classified. Each of the two objects inherits the property of being a word class according to the three criteria.

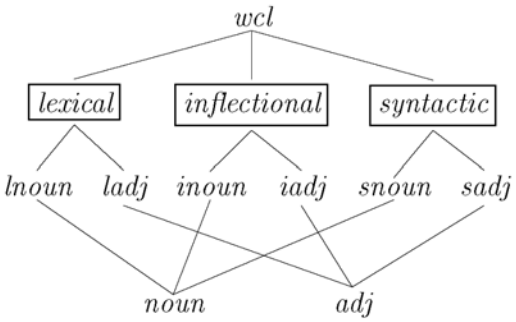

Figure 4: Nouns and adjectives are nouns and adjectives from all three aspects

The common tagset is specified as a hierarchy of concepts or *types*, partially ordered by their specificity. Each type denotes a set of objects – language-specific tags, identified by their name and membership in a language-specific tagset. The topmost type is the most general one,

[24] All hierarchies shown here merely illustrate how a common abstract tagset could be designed. They are partial in two ways: (i) they cover only a fraction of morphological categories, (ii) they make no attempt to cover more than very few languages, often just one or two. However, each of the sample hierarchies can be extended by inserting more types and links.

[25] Our use of the term *lexical* rather than *semantic* word class may be misleading. The rationale behind the preference derives from the fact that *lexical* word classes have their properties specified in the lexicon, rather than by rules of morphology or syntax.

denoting all tags in all tagsets.[26] Immediate subtypes of a supertype denote subsets of that supertype. A tag in the denotation of the supertype must be in the denotation of at least one of the subtypes. A subtype can have more than one supertype. In this case, the subtype denotes a subset of the intersection of the sets denoted by the supertypes. This means that a tag in the denotation of a type can be referred to by any of its supertypes; the higher the supertype, the less focused reference.

Unlike regular nouns and adjectives, a Czech *wh-* form *který* 'which' in its use as a relative (rather than interrogative) pronoun belongs to three different word classes at the same time, according to the three aspects; it is in fact a pronoun from the inherently *lexical* perspective, an adjective from the *inflectional* perspective, and a noun from the *syntactic* perspective. In (1), *který* is at the same time a syntactic noun as the subject of the relative clause, a lexical pronoun with "dog" as its antecedent, and – due to its adjectival declension paradigm – an inflectional adjective.

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

Now how should we express this triple membership? The Czech relative pronoun *který*, tagged as **P4** in the Czech tagset,[27] is a subtype of lexical pronoun (*lprn*), inflectional adjective (*iadj*) and syntactic noun (*snoun*), each of the word classes a subtype of a type representing a different dimension. The type corresponding to the Czech tag **P4** inherits from all three of its word-class supertypes, can be labelled by their conjunction and referred to by any of them. The corresponding fragment of the hierarchy is shown in fig. 5.

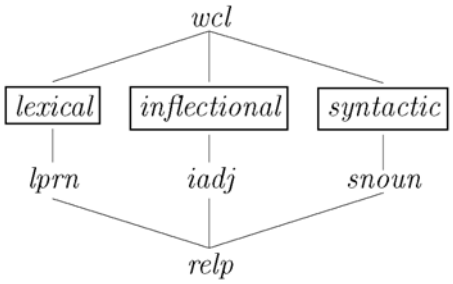

Figure 5: A hierarchy fragment for the Czech relative pronoun *který* 'which'

We can extend our fragment by other objects as in fig. 6: cardinal and ordinal numerals, personal, possessive and interrogative pronouns. Ordinal numerals such as *pátý* 'fifth' are treated as *lexical* numeral and adjective – both *inflectional* and *syntactic*. Possessive pronouns behave in a similar way, except in that they are *lexical* pronouns. Personal pronouns are inflectional and syntactic nouns, similarly as are cardinal numerals. Interrogative pronouns are worth more attention. Unlike *který* in its relative use, its homonymous interrogative counterpart can be used in the syntactic position of adjective or noun. While *intp* inherits from a new node *snom*, representing all objects with the property of being *either* a syntactic noun *or* a syntactic adjective, *relp* has an additional ancestor, namely *snoun*, which excludes *který* as a relative pronoun from the class of syntactic adjectives.

[26] Except for highly parochial tags, excluded from the common tagset.
[27] We ignore all but the first two positions in the tag.

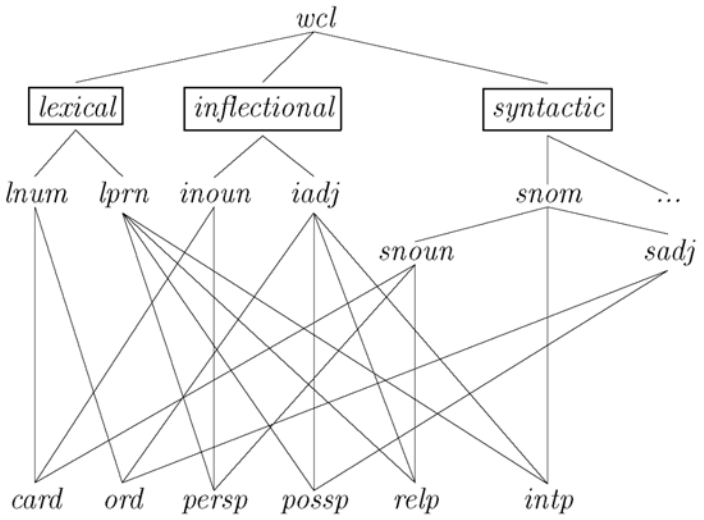

Figure 6: Distinguishing types of numerals and pronouns in a hierarchy

However, the Prague tagset for Czech does not have a tag for *který* as a relative pronoun. The tag **P4** covers both the relative and the interrogative use of *který*. The tag **P4** should be properly represented as 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. 7 correctly captures this ambiguity. The Czech tag **P4** corresponds to a node labelled *lprn* $\wedge$ *iadj* $\wedge$ *snom*, whose two daughters stand for interrogative and relative pronouns.

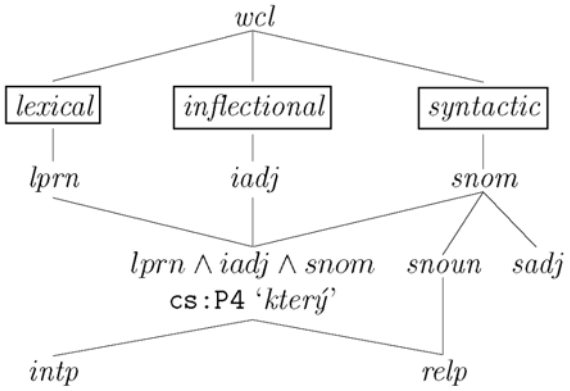

Figure 7: A single node for interrogative and relative pronouns

The three views of word class allow for proper mapping between language-specific tagsets and the common tagset. The tags for adjective in the English, German, French, Italian and Polish tagsets cover also ordinal numerals. If all these tags are translated as common tags for *syntactic* rather than *lexical* adjectives, they end up correctly in the same class as Czech, Spanish, Russian or Bulgarian adjectives, ordinal numerals and possessive pronouns. Their *inflectional* word class will be specified – most likely – also as adjective, but what about their *lexical* word class? It is not the case that it can be arbitrary; a German word tagged as an adjective is unlikely to be a lexical preposition or a finite verb. When a specific word class is unknown, we need a means of specifying that the word belongs to a more general word class.

Fig. 8 shows a fragment of the hierarchy with a node representing both ordinal numerals and adjectives, labelled $(lord \vee ladj) \wedge iadj \wedge sadj$ and corresponding to the German tag **ADJA**.

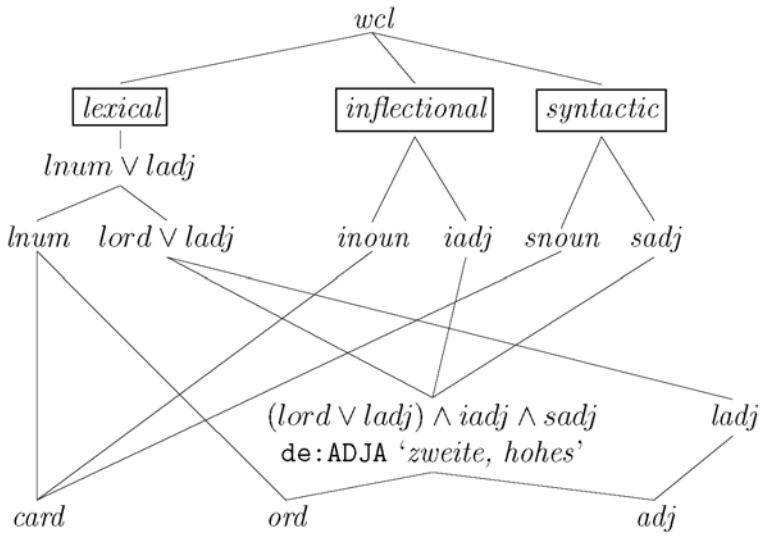

Figure 8: A single node for ordinal numerals and adjectives

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 ($ladj \vee lord$).

Partial hierarchies can be merged. The result of merging the above two hierarchies (figures 7 and 8) is shown in fig. 9.

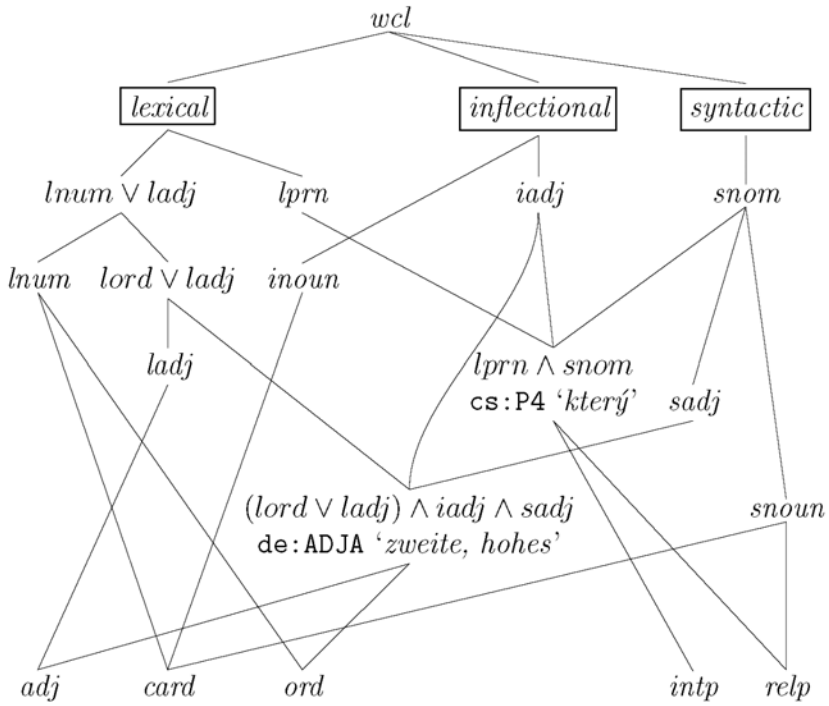

Figure 9: Hierarchies in figures 7 and 8 merged

We are aware of the fact that we have just scratched the surface of the topic of cross-classifying word classes. Obvious candidates for this treatment could be derived words. However, the possibility of multiple derivation and the constraints of the language-specific tagsets may present a prohibitive obstacle to any significant extension of the approach.

4.3. Morphological categories

Tags often encode more information than just word class, especially in highly inflected languages. The relevance of specific morphological categories for a given class is not random; it is dependent on the word class. More precisely, it depends on the aspect of the word class. Word class of any flavour may be required to co-occur with a set of other categories as its properties – 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 possessive pronoun such as *jejího* is *lexically* 3rd person, singular and feminine, while *inflectionally* it is masculine or neuter, singular, genitive or accusative (2).[28]

(2) *Martina* *je* *moje* *sousedka.*
Martina is my neighbour$_{\text{FEM,SG,NOM}}$.
Jejího *syna* *často potkávám* *v tramvaji.*
her$_{\textit{lex}:\ \text{3RD,FEM,SG};\ \textit{infl}:\ \text{MASC,SG,ACC}}$ son$_{\text{MASC,SG,ACC}}$ often meet$_{\text{1ST,SG}}$ in tram.
'Martina is my neighbour. I often meet her son on the tram.'

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

The set of categories or properties appropriate to a word class may be defined as attribute-value pairs, or as types in the hierarchy, which further cross-classify types corresponding to language-specific tags. The latter approach allows the user to use only types rather than types and attributes, perhaps referring to all plural items by specifying them merely as **pl**. However, the two formats are interchangeable, which may be useful for working with existing tagsets, such as *Interset*.

The tag for the Czech possessive pronoun *jejího* in fig. 10 is a subtype of lexical pronoun (*lprn*) and inflectional adjective (*iadj*).[29] As a possessive pronoun, it is required by the specification of the hierarchy and more general co-occurrence restrictions[30] to be a subtype of lexical gender (*lgend*), lexical number (*lnum*) and lexical person (*lpers*), more precisely of their intermediate subtypes, specifying the morphological categories.[31] As an inflectional adjective, it is required to be a subtype of inflectional gender *igend*, inflectional case (*icase*) and inflectional number (*inum*). In isolation, the form *jejího* is ambiguous between (inflectional) genitive and accusative and inflectional masculine and neuter genders. As the tag suggests, the former ambiguity is assumed to be resolved (the digit "4" at the 5th position stands for accusative), unlike the latter ambiguity, which is retained (the character "Z" at the third position stands for all genders, except feminine). Therefore, the tag is a subtype of *imasc* $\vee$ *ineut*, covering both *imasc* and *ineut*.

[29] It is also a subtype of syntactic adjective. Types less relevant for the current discussion are omitted for brevity.

[30] See 4.5 below for more details.

[31] Again, irrelevant types are omitted, including the animate vs. inanimate distinction in the masculine gender. We also leave aside the issue of the proper lemma for *jejího*, tagged as a (lexical) 3rd person singular feminine form; whether the lemma is in fact *její*, the base form of the (lexical) 3rd person feminine singular possessive pronoun (or even *ona*, the base form of the 3rd person feminine singular personal pronoun), or a corresponding representative of possessive (personal) pronouns of both numbers and all persons and genders. For lemmas, we make no attempt at this stage to introduce a uniform policy and rely on the output of the language-specific taggers. The tagger currently used in *InterCorp* for Czech suggests *její* as the lemma for *jejího*.

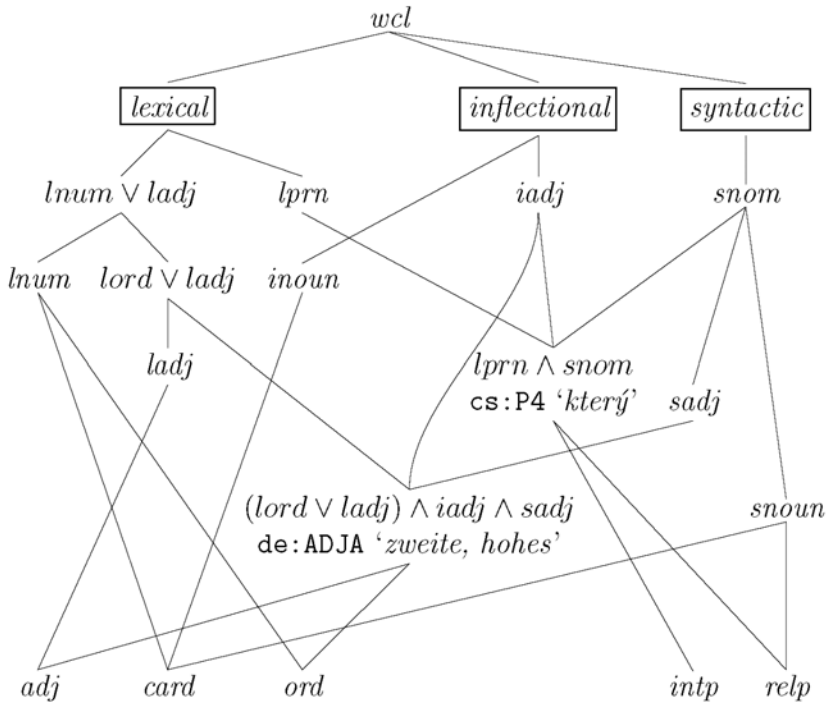

Figure 10: Morphological categories used to tag a Czech possessive pronoun *jejího*, a category-based view

The hierarchy in fig. 10 leaves the lexical/inflectional distinction implicit. In fig. 11 this distinction is shown at the top level, as in all previous hierarchies. For clarity, general category labels (*gend*, *case*, etc.) are omitted.

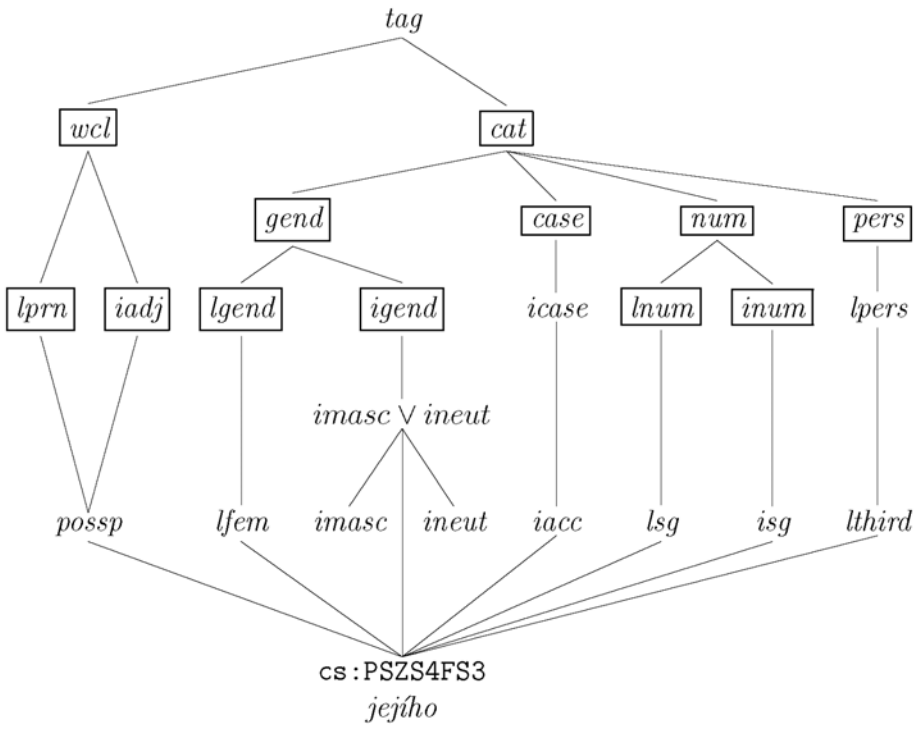

Figure 11: Morphological categories used to tag a Czech possessive pronoun *jejího*, a lexical/inflectional view

4.4. Building and using the common tagset

We still have to show that the task to build the complex hierarchy is realistic and that the hierarchy can be used as suggested.

The type hierarchies presented so far are equivalent to concept lattices of Formal Concept Analysis (FCA), a logical formalism equipped with methods of constructing and using the lattices (Ganter & Wille, 1996, 1999), (Wille, 2005). The task of FCA is to classify objects according to their properties (attributes). The classification is based on the notion of *concept*, consisting of a set of objects as its extension and a set of attributes as its intension. Objects sharing a common set of attributes are extensions of the same concept.

The first step of the analysis is to identify the objects and their (definitional) attributes. This is done in a tabular data structure called *formal context*. Table 1 is an example of a formal context for our previous example of adjectives and cardinal and ordinal numerals (as in fig. 8). Note that attributes corresponding to the boxed labels in fig. 8 are not included. They would be specified for all objects in the formal context and would not make the resulting lattice more informative.

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

Table 1: Formal context for adjectives and ordinal numerals

Next, a set of formal concepts is built, each of the concepts consisting of a pair of the set of objects (its extension), and a set of attributes (its intension). 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).

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

Finally, the concept lattice can be drawn (fig. 12). Note that its geometry is significantly simpler than the hierarchy constructed intuitively (as in fig. 8), while the concept corresponding to the tag covering both adjectives and cardinal numerals is still present.

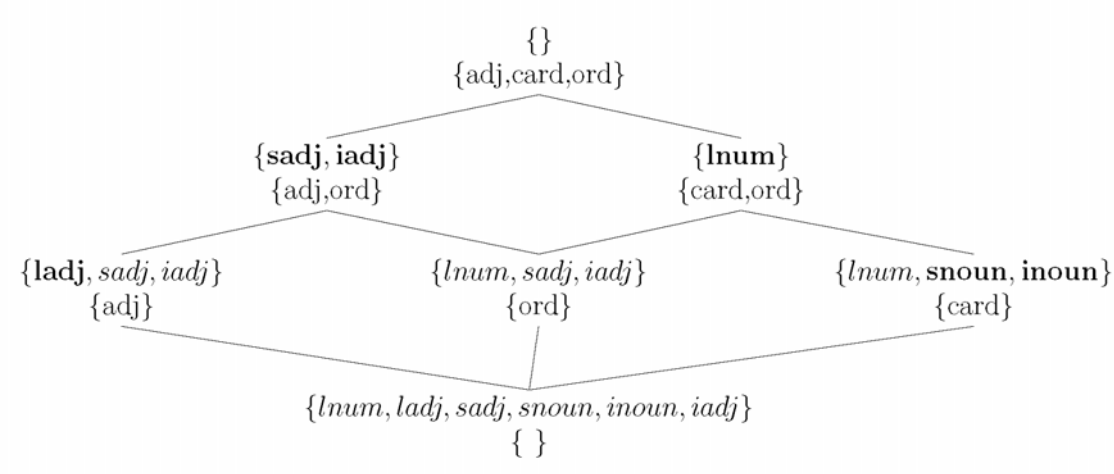

Figure 12: Concept lattice for adjectives and ordinal numerals

The last two steps can be done automatically. Some software is freely available, even as an online application.[32]

The concept lattice can be used for reasoning about attributes. For example, we can make implications such as *ladj* ⇒ *sadj* or *snoun* ⇒ *lnum*. They are valid only within the model, but we can use these and similar statements to assist the user in making queries including language-independent category labels (such as "adj"), or to match incompatible language-specific tags.

The concept with the extension {ord} corresponds to **cs:Nr**, the Czech tag for ordinal numerals, while the concept with the extension {adj,ord} corresponds to **de:ADJA**, the German tag covering adjectives and ordinal numerals. Looking up a Czech equivalent of

[32] Online Java Lattice Building Application: http://maarten.janssenweb.net/jalaba/JaLaBA.pl, Galois Lattice Interactive Constructor: http://www.iro.umontreal.ca/~galicia/, RubyFCA – A Formal Concept Analysis Tool: http://www.kotonoba.net/rubyfca/doc/about/

de:ADJA involves searching for a Czech tag corresponding to the {adj,ord} concept. Because there is no such tag, the more specific concepts are traversed and the disjunction of the Czech tags corresponding to {adj} and {ord} concepts is offered as the equivalent. Looking up a German equivalent of **cs:Nr** is much more complex and is similar to the corpus query scenario described below, in which the user searches for "ord" in a German text.

When the user searches for "ord" in a Czech text, the search is easy, because the appropriate tag **cs:Nr** is available. For German, there is no tag corresponding to "ord". There are also no concepts more specific than {ord} available in the hierarchy that would correspond to German tags. The only option is to resort to a more general concept {adj,ord}, with a 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".

Attributes specified for an object in a formal context are interpreted in conjunction. Thus, specifying both *snoun* and *sadj* as attributes of an interrogative pronoun (intp) would mean that it is simultaneously a syntactic noun and a syntactic adjective. To model disjunction of attributes we have to introduce a more general attribute covering the two options. The formal context and concepts for numerals and pronouns are shown below in tables 3 and 4 and the corresponding lattice in fig. 13.

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

Table 3: Formal context for numerals and pronouns

1	⟨{card,ord,persp,possp,relp,intp},	{snom}⟩
2	⟨{card,ord},	{lnum,snom}⟩
2	⟨{card,persp,relp},	{snoun,snom}⟩
2	⟨{ord,possp,relp,intp},	{iadj,snom}⟩
2	⟨{persp,possp,relp,intp},	{lprn,snom}⟩
3	⟨{card,persp},	{inoun,snoun,snom}⟩
3	⟨{ord,possp},	{iadj,sadj,snom}⟩
3	⟨{persp,relp},	{lprn,snoun,snom}⟩
3	⟨{possp,relp,intp},	{lprn,iadj,snom}⟩
4	⟨{card},	{lnum,inoun,snoun,snom}⟩
4	⟨{ord},	{lnum,iadj,sadj,snom}⟩
4	⟨{persp},	{lprn,inoun,snoun,snom}⟩
4	⟨{possp},	{lprn,iadj,sadj,snom}⟩
4	⟨{relp},	{lprn,iadj,snoun,snom}⟩
5	⟨{},	{lnum,lprn,inoun,iadj,snoun,sadj,snom}⟩

Table 4: Formal concepts derived from table 3

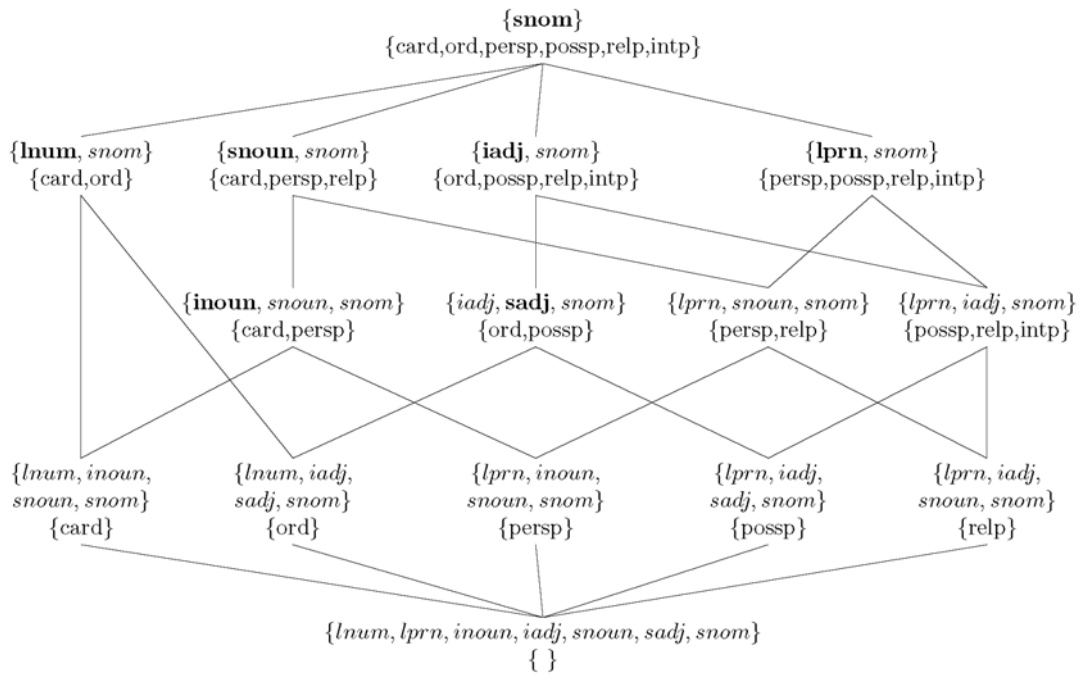

Figure 13: Concept lattice for numerals and pronouns

This is not the first application of FCA in the field of linguistics, not even in a multilingual setting. Priss (2005) gives an overview of linguistic applications of FCA and Janssen (2002a,b, 2004) is concerned with multilingual lexical databases. His lattice is in fact a structured lexical interlingua, connecting words from different languages. The attributes are not language-specific, giving rise to a hierarchy of interlingual concepts. Every such concept has a potentially empty set of words lexicalizing it in every language, and every word of every language has a set of interlingual concepts it expresses. Such a lattice allows for finding a nearest equivalent in another language even when there is a lexical gap. The solution is to find a translational hyperonym, possibly supplying the attribute(s) discriminating the more specific source word, a "definitional surplus". In the concept lattice of horse terms, where no French equivalent for the English term *colt* can be found, the system can suggest the equivalent of its hyperonym *foal – poulain* in French, with the additional specification that it is in fact a male foal. This can be useful even within one language for generating definitions of terms based on hyperonyms.

It is easy to see that Janssen's multilingual lexical concept lattice is very similar to the common abstract tagset. Given that the world of morphosyntactic tags is simpler than the world of words, this is a reassuring finding.

4.5. Modularity and formal rigour

The common tagset should be exhaustively specified, including its language-specific subsets. An explicit formal specification is not an end in itself – it serves to guide the user and supports software tools, including the corpus manager and conversion modules. In addition to error checking it allows for underspecified tag queries and simplifies maintenance and extensions of the tagset.

Linguistic categories and their values correspond to types in a hierarchy, ordered by their specificity from the most general type covering everything else down to the most specific type.

Immediate subtypes of a type are required to fully cover the domain of the supertype – each member of a category denoted by the supertype must be included in at least one category denoted by a subtype. A type can have more than one supertype. A pair of immediate supertypes is interpreted in conjunction, the denotation of the type being the intersection of the denotations of the supertypes.

Types with multiple supertypes typically correspond to language-specific tags and their position in the tangled hierarchy (the choice and relationships of their supersets) is restricted by the categorial values encoded in the tags. However, they should be required to follow some more general patterns in any case (see 4.3 above). In a feature-based format a set of attribute-value pairs may be "strongly typed", including all the appropriate attributes and their values. Our system is equivalent to a hierarchy of strongly typed flat (non-embedding) feature structures. Every intersecting type, including the most specific types, has its known set of supertypes, therefore a known set of appropriate categories. However, in practice, especially when the types and mappings between the hierarchy and the tagsets are defined, it may be unwieldy to rely solely on this inherent property of the formalism, due to the large numbers of tags and relevant categories. As an auxiliary mechanism, the specification of intersecting/unioning types may be governed by a formally weaker notion of general and language-specific co-occurrence restrictions. Two possible restrictions for Czech are shown in (3).

(3) $lprn \Rightarrow lgend \wedge lnum \wedge lpers$
$iadj \Rightarrow igend \wedge inum \wedge icase$

The hierarchy of types can (and should) be specified once for all languages, with all language-specific tags corresponding to some type. This would allow the use of an arbitrary tagset for tag queries and tag display, including underspecified queries and underspecified display (unavoidable in the case of a missing tag equivalent), and also for tag conversions.

The common abstract tagset (**CTS**), specified as a hierarchy of types, is the core of the system, a knowledge base used by all other components. The other components form several layers of patchwork shells (patchwork being a metaphor for their multilingual variety). Types in the hierarchy can be linked to tags in the "external" language-specific tagsets ($\mathbf{ETS}_L$). Some parts of the hierarchy may be language-universal, other parts specific to a group of languages or even to a single language; typically only a subset of the common tagset is relevant to a language. Abstract language-specific subsets of the common tagset ($\mathbf{ATS}_L$) are defined as functions of **CTS**. A function ${F^{extr}}_L$, extracting an $\mathbf{ATS}_L$, traverses the hierarchy, selecting paths with at least one type linked to a tag in that language-specific tagset. The functions are only allowed to eliminate types and hierarchical links from the common tagset definition; they cannot add any new types or links.

Any $\mathbf{ATS}_L$ may be rendered in a format (positional, attributive or other: $\mathbf{FTS}_{L,F}$) according to the user's or the task's preference and used in queries, displays, or even in corpus data. The format may also depend on factors such as the choice of a corpus manager or its user interface, and may even be ready to support a menu-driven specification of tag queries.

Tagset types

CTS	common abstract tagset
$\mathbf{ATS}_L$	abstract tagset for language L, derived from **CTS**
$\mathbf{ETS}_L$	external tagset for language L
$\mathbf{FTS}_{L,F}$	tagset for language L in format F, derived from ATS_L

Mappings between tagsets

CTS → $\mathbf{ATS}_L$ (common tagset → abstract L-specific tagset)
$\mathbf{ATS}_L$ ↔$\mathbf{FTS}_{L,F}$ (abstract L-specific tagset ↔ formatted tagset)
$\mathbf{ATS}_L$ ↔$\mathbf{ETS}_L$ (abstract L-specific tagset ↔ external tagset)

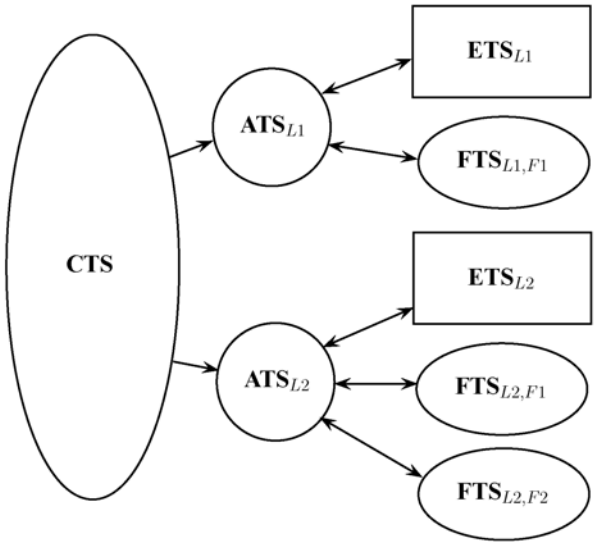

Figure 14: Mappings between tagsets

5. Conclusion

Users of a multilingual parallel corpus deserve some solution to the problem of tagset variety. We have shown that one of the solutions can be an abstract common tagset, designed in a formally sound, highly modular and linguistically informed fashion (based on a three-way distinction in the system of word classes), allowing for intuitive and underspecified queries, supporting the option to define various tag formats for available abstract tagsets and to map tags between different tagsets. Importantly, the system is expected to make use of existing work (such as results of the *Interset* project).

The complex multiple inheritance system allows the user to make queries underspecified to an arbitrary degree and along multiple dimensions. It is also well suited to the role of a common tagset. If corpus data include only original, language-specific tags, the system can be easily modified and extended without touching the corpus data.

The cost is higher complexity, both conceptual and formal/implementational; a module to resolve queries using the type hierarchy specification is needed. And some users may even prefer a menu-driven specification of tag-based queries, an approach that does not necessarily require cross-classification of linguistic categories. However, we believe that the price is well justified and that the modular framework of our proposal allows for customizing the setup of the system according to specific preferences. Formal Concept Analysis seems to be the answer to concerns about the costs of designing the hierarchy.

Nevertheless, it is still difficult to foresee the real costs and benefits of the proposed solution and compare it with alternatives. Therefore, the next step must be a more detailed investigation of all aspects of the solution.

Together with the effort to extract knowledge from monolingual texts, tags can be compared and perhaps made more precise across languages by using word-to-word alignment.

In this way, tags other than those belonging to the language-specific tagset may be used physically in the text.

References

Baskaran, S., Bali, K., Bhattachrya, T., Bhattachrya, P., & Jha, G. N., 2008, Designing a Common POS-Tagset Framework for Indian Languages. In *The 6th Workshop on Asian Language Resources*, pages 89–92, Hyderabad, India. Indian School of Business.

Chiarcos, C., 2008, An ontology of linguistic annotations. In U. Mönnich and K.-U. Kühnberger, editors, *Foundations of Ontologies in Text Technology – Applications*, volume 2 of *LDV Forum*, pages 1–16. Gesellschaft für linguistische Datenverarbeitung.

Derzhanski, I. A. & Kotsyba, N., 2008, The Category of Predicatives in the Light of the Consistent Morphosyntactic Tagging of Slavic Languages. In O. Shemanayeva, editor, *Lexicographic Tools and Techniques*, pages 68–79, Moscow. IITP RAS.

Derzhanski, I. A. & Kotsyba, N., 2009, Towards a Consistent Morphological Tagset for Slavic Languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian. In Garabík, pages 9–26.

Dimitrova, L., Kosecka-Toszewa, V., Derzhanski, I., & Roszko, R., 2009a, Annotation of Parallel Corpora (on the Example of the Bulgarian–Polish Parallel Corpus). In Shyrokov & Dimitrova, pages 47–54.

Dimitrova, L., Garabík, R., & Majchráková, D., 2009b, Comparing Bulgarian and Slovak Multext-East morphology tagset. In Shyrokov & Dimitrova, pages 38–46.

Erjavec T., 2001, Harmonised Morphosyntactic Tagging for Seven Languages and Orwell's 1984. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, NLPRS'01*, pages 481–492.

Erjavec, T., 2009, MULTEXT-East Morphosyntactic Specifications: Towards Version 4. In Garabík, pages 59–70.

Erjavec, T., Krstev, C., Petkevič, V., Simov, K., Tadić, M., & Vitas, D., 2003, The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. In *Proceedings of the EACL 2003 Workshop on Morphological Processing of Slavic Languages*, pages 25–32. ACL.

Farrar, S. & Langendoen, T., 2003, A linguistic ontology for the semantic web. *GLOT International*, **7**(3).

Feldman, A. & Hana, J., 2010, *A resource-light approach to morpho-syntactic tagging*. Rodopi, Amsterdam/New York, NY.

Ganter, B. & Wille, R., 1996, *Formale Begriffsanalyse. Mathematische Grundlagen*. Springer, Berlin/Heidelberg.

Ganter, B. & Wille, R., 1999, *Formal Concept Analysis. Mathematical Foundations*. Springer, Berlin/Heidelberg.

Garabík, R., editor, 2009, *Metalanguage and Encoding Scheme Design for Digital Lexicography*, Bratislava. Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences.

Hajič, J., 2004, *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Charles University Press, Prague.

Hughes, J., Souter, C., & Atwell, E., 1995, Automatic extraction of tag set mappings from parallel annotated corpora. In *From Text to Tags: Issues in Multilingual Language Analysis, Proc. ACL-SIGDAT Workshop*, pages 10–17.

Ide, N. & Véronis, J., 1996, Multext (multilingual tools and corpora). In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 90–96, Kyoto. ACL.

Janssen, M., 2002a, Differentiae Specificae in EuroWordNet and SIMuLLDA. In *Proceedings of ONTOLEX 2002*. Las Palmas, Gran Canaria.

Janssen, M., 2002b, *SIMuLLDA: a Multilingual Lexical Database Application using a Structured Interlingua*. Ph.D. thesis, Utrecht University.

Janssen, M., 2004, Multilingual Lexical Databases, Lexical Gaps, and SIMuLLDA. *International Journal of Lexicography*, **17**(2).

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

Komárek, M., 2006, *Příspěvky k české morfologii*. Periplum, Olomouc.

Kotsyba, N., Shypnivska, O., & Turska, M., 2008, Principles of organising a common morphological tagset for PolUKR (Polish-Ukrainian Parallel Corpus). In M. A. Kłopotek, A. Przepiórkowski, S. T. Wierzchoń, and K. Trojanowski, editors, *Intelligent Information Systems*, pages 475–484, Warsaw. Akademicka Oficyna Wydawnicza EXIT.

König, E., Bakker, D., Dahl, E., Haspelmath, M., Koptjevskaja-Tamm, M., Lehmann, C., & Siewierska, A., 1993, Eurotyp guidelines. Technical report, European Science Foundation Programme in Language Typology.

Leech, G. & Wilson, A., 1996, Eagles recommendations for the morphosyntactic annotation of corpora. Technical report, AGLES, Istituto di Linguistica Computazionale, Pisa. AGLES-Guidelines EAG–TCWG–MAC/R.

Osolsobě, K., Pala, K., & Sedláček, R., 2006, Brněnský atributivní tagset. NLP FI MU Brno, http://nlp.fi.muni.cz/projekty/ajka/tags.pdf.

Pořízka, P. & Schäfer, M., 2009, MorphCon – A Software for Conversion of Czech Morphological Tagsets. In J. Levická and R. Garabík, editors, *NLP, Corpus Linguistics, Corpus Based Grammar Research. Proceedings of the Fifth International Conference Slovko 2009*, pages 292–301, Bratislava. Slovak National Corpus, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences. http://korpus.juls.savba.sk/~slovko.

Priss, U., 2005, Linguistic applications of formal concept analysis. In B. Ganter, editor, *Formal Concept Analysis. Foundations and Applications*, volume 3626 of *Lecture Notes in Artificial Intelligence*, pages 149–160. Springer, Berlin/Heidelberg.

Przepiórkowski, A., 2009, A comparison of two morphosyntactic tagsets of Polish. In V. Koseska-Toszewa, L. Dimitrova, and R. Roszko, editors, *Representing Semantics in Digital Lexicography: Proceedings of MONDILEX Fourth Open Workshop*, pages 138–144, Warsaw.

Przepiórkowski, A. & Woliński, M., 2003, A flexemic tagset for Polish. In *Proceedings of* Morphological Processing of Slavic Languages*, EACL 2003*.

Przepiórkowski, A., Krynicki, Z., Dębowski, Ł., Woliński, M., Janus, D., & Bański, P., 2004, A search tool for corpora with positional tagsets and ambiguities. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*, pages 1235–1238.

Shyrokov, V. & Dimitrova, L., editors, 2009, *Organization and Development of Digital Lexical Resources*, Kyiv. National Academy of Sciences of Ukraine, Ukrainian Lingua-Information Fund.

Wille, R., 2005, Formal concept analysis as mathematical theory of concepts and concept hierarchies. In B. Ganter, editor, *Formal Concept Analysis. Foundations and Applications*, volume 3626 of *Lecture Notes in Artificial Intelligence*, pages 1–33. Springer, Berlin/Heidelberg.

Zeman, D., 2008, Reusable tagset conversion using tagset drivers. In *Proceedings of the Language Resources and Evaluation Conference, LREC 2008*, Marrakech, Marocco.