Alexandr Rosen
Charles University, Faculty of Arts, Prague

ON THE ART OF TAMING AND EXPLOITING PARALLEL TAGS IN A MULTILINGUAL CORPUS[1]

Abstract

Multilingual parallel corpora can be annotated with monolingual tools, such as morphosyntactic taggers. However, even taggers for typologically similar languages use incompatible tagsets, which results in a conceptual and formal variety of tags. Retraining taggers on data annotated with a common tagset is not a realistic option. However, differences between tagsets are often rooted in different linguistic perspectives rather than in real distinctions between the languages, which means a common ground could be found. Moreover, a different perspective may provide additional information missing in one tagset but present in another. Our first goal is to delegate the task of dealing with multiple tagsets to an abstract interlingual representation of linguistic categories. Ideally, each tag in every language-specific tagset used in the corpus is linked to a position in a tangled hierarchy of concepts. To accommodate the different perspectives, the hierarchy takes three views of word class: inflectional, syntactic, and semantic (lexical), each with its appropriate morphological characteristics. Mismatches between tags are properly represented, which allows for a principled mapping strategy between languages-specific tagsets, and for intuitive and underspecified queries. The hierarchy can be built and the mismatches partially resolved using Formal Concept Analysis. Our second goal is to refine existing morphosyntactic annotation by projecting distinctions in one tagset onto a conceptually different tagset. The hierarchy and automatic word-to-word alignment is used to learn from tags in another language. We show results of an experiment aimed at discovering how feasible this approach is for Czech and two other languages (English and Polish).

---

1 This is a revised and extended version of a paper entitled "Mediating between incompatible tagsets," published in *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora*, co-located with the Ninth International Workshop on Treebanks and Linguistic Theories in Tartu, Estonia, 2010. The work was supported by grant no. MSM0021620823 of the Czech Ministry of Education, Youth and Sports.

1 Introduction

Multilingual corpora are often annotated with a variety of language-specific morphosyntactic tagsets. To use tags in a query or to understand its results may require 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.

There seem to be two possible solutions: (i) to transform each language-specific tagset into a tagset notationally and conceptually compatible with the other tagsets according to a common standard, or (ii) to provide mapping between the language-specific tagsets, either directly or via an interlingual representation. If available, the single consistent standardised annotation scheme (the former option) in the spirit of *MULTEXT-East* (Erjavec, 2009) is preferable. However, to build a multilingual corpus using such a scheme is not realistic, especially when more than a handful of languages are involved.[2] Available taggers are trained on different tagsets, and consistently annotated training data are seldom available even for typologically close languages.

Confronted with texts already tagged in different ways, the user may still believe that a tagset can be translated into a common standard or into another existing tagset (the latter option). But a given tag may be too specific or too general to be expressed by a tag from a different tagset. Table 1 illustrates the tagset variety using comparable examples of prepositional phrases in 11 languages, tagged by available tools.[3] While some corresponding tags used in the examples are indeed notational equivalents, other tags

---

2 In its release 3, dated February 2011, the parallel corpus *InterCorp* offers on-line concordances in 23 languages, 14 of them tagged with different morphosyntactic tagsets. The texts (114 million words – the total for all languages) 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 Čermák (this volume) or http://korpus.cz/intercorp/?lang=en.

3 Bulgarian, Dutch, English, French, German, Italian, Russian and Spanish are tagged by TreeTagger (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/), Czech and Slovak by Morče (http://ufal.mff.cuni.cz/morce/) using appropriate morphological analyzers (by Jan Hajič and Radovan Garabík, respectively), Polish by TaKIPI and Morfeusz (http://nlp.ipipan.waw.pl/TaKIPI/), Hungarian by HunPOS (http://code.google.com/p/hunpos/). Kudos for Lithuanian and Norwegian is due to http://donelaitis.vdu.lt/~vidas/ and http://maximos.aksis.uib.no/Aksis-wiki/Oslo-Bergen_Tagger. The tags used here and below are often truncated for brevity.

are not related 1:1. The English tag IN is used both for prepositions and subordinating conjunctions. The German tag ADJA covers attributive adjectives, 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.

Thus "translating" a tag often means to deal with a situation common in real translation. The source tag may be more general than the target tag (the English tag IN above), more specific (a tag for preposition in any language except English with English as the target), or both at the same time (the German tag ADJA is too general when compared with the English tag JJS because it does not distinguish degree, but too specific because it is only used for attributive adjectives). It seems that the conversion inevitably involves cases where the target tag conveys too much and/or too little information.

For closed-class items (pronouns, function words), lexeme-specific information corresponding to the source tag can be used to derive a more fine-grained target tag (see Kotsyba *et al.*, 2009). Tags for open word classes can be translated into an intermediate representation, as in *Interset* (Zeman, 2010), or even into a standardized ontology (such as that developed by www.isocat.org). Such a "tagset interlingua" (possibly restricted to the tagsets of immediate interest for practical reasons) should properly capture the meaning of any language-specific tag. Of course, it is impossible for an arbitrary abstract concept to be encoded precisely as a tag in any target tagset, but at least the missed or redundant information can be identified. In the context of many different languages and tagsets, an interlingual representation 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.[4]

4 Chiarcos (2008) and Chiarcos *et al.* (2008) present a technically mature framework for integrating conceptually incompatible linguistic annotation schemes, very much in line with the present proposal.

bg	на	това	приятелско	движение
	R	**Pde-os-n**	**Ansi**	**Ncnsi**
cs	v	těch	nejodlehlejších	zástavbách
	RR--6	**PDXP6**	**AAFP6----3A**	**NNFP6-----A**
de	in	den	abgelegensten	Außenbezirken
	APPR	**ART**	**ADJA**	**NN**
en	in	the	remotest	exurbs
	IN	**DT**	**JJS**	**NNS**
fr	dans	les	plus lointaines	banlieues
	PRP	**DET:ART**	**ADV ADJ**	**NOM**
hu	a	szép	katalán	lányba
	ART	**ADJ**	**ADJ**	**NOUN(CAS(ILL))**
it	da	queste	lingue	babeliche
	PRE	**PRO:demo**	**NOM**	**ADJ**
lt	už	tos	juokingos	nosies
	prln	**įvrd neįvardž**	**bdvr teig nelygin.l**	**dktv mot.gim vnsk K**
no	med	den	latterlige	nesen
	prep	**det dem mask ent**	**adj be ent pos**	**subst apell mask be ent**
nl	in	dit	schitterende	appartement
	600	**370**	**103**	**000**
pl	w	tym	wspaniałym	apartamencie
	prep:loc:nwok	**adj:sg:loc:m3:pos**	**adj:sg:loc:m3:pos**	**subst:sg:loc:m3**
ru	в	самых	отдалённых	районах
	Sp-l	**P---pl**	**Afp-plf**	**Ncmpln**
sk	v	tejto	dejinnej	chvíli
	Eu6	**PFfs6**	**AAfs6x**	**SSfs6**
sp	en	las	zonas	más remotas
	PREP	**ART**	**NC**	**ADV ADJ**

Table 1: Different tags for words in similar prepositional phrases

Our goal is to delegate the task of dealing with multiple tagsets 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 Komárek (1999) 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-spe-

cific tagsets and stay open to future extensions. The hierarchy can be built and mismatches between tagsets partially resolved using Formal Concept Analysis (Ganter & Wille, 1999). In a word-aligned corpus, morphosyntactic annotation can be refined by adding information from corresponding tags in other languages.

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 (Hajič, 2004), the Polish tagset (Przepiórkowski & Woliński, 2003) for inflectional word classes, the German tagset distinguishes pronouns by their syntactic function.[5]

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^). In either case *Interset* retains the Polish tagset specifications in the intermediate representation as POS=adj and POS=part. They cover a much larger set of categories than their namesakes in Czech and most other languages.

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. 1, 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 'morpholo-

5 Díaz-Negrillo *et al.* (2010) propose a very similar tripartite analysis of word classes to describe errors in learner language. Their classification, based on lexical stem, distribution, and morphology, is an independent piece of evidence that the approach is well-founded and useful.

gical') and *syntactic*.[6] 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.

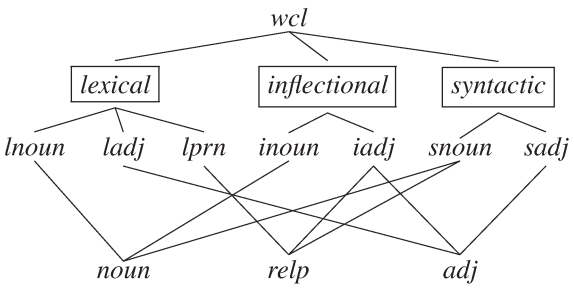

Figure 1: A hierarchy for nouns, adjectives and relative pronouns

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.

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 – see example (1) below – 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. The three dimensions are exemplified in Table 2 using examples of few Czech word classes.[7]

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

6 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.

7 The specific assignment of lexical, inflectional and syntactic classes is open to refinement and modification, e.g. the category of *noun* as the inflectional class for personal pronouns may be too general. In fact, the inflectional dimension could be seen as a hierarchically structured specification of morphological paradigms.

		example	gloss	*lexical*	*inflectional*	*syntactic*
numerals	**ordinal**	*pátý*	fifth	numeral	adjective	adjective
	cardinal	*pět*	five	numeral	noun	noun
pronouns	**personal**	*ty*	you	pronoun	noun	noun
	possessive	*tvůj*	your	pronoun	adjective	adjective
	relative	*který*	which	pronoun	adjective	noun
	interrogative	*který*	which	pronoun	adjective	noun ∨ adjective
adverbial participle		*volajíce*	calling	verb	participle	adverbial

Table 2: Examples of cross-classification in Czech

The hierarchy in Fig. 2 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*.

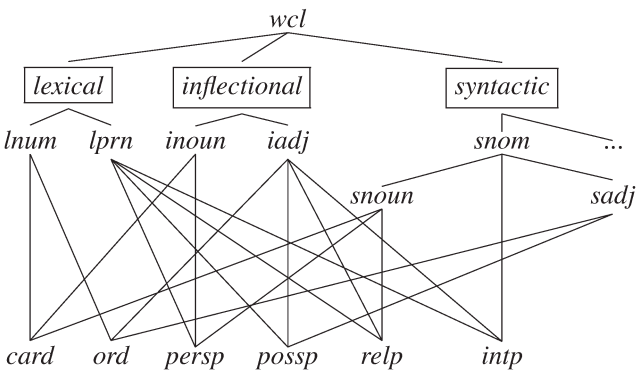

Figure 2: Distinguishing types of numerals and pronouns in a hierarchy

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. 3 captures this ambiguity. The Czech tag P4 corresponds to a node labelled $lprn \wedge iadj \wedge snom$.

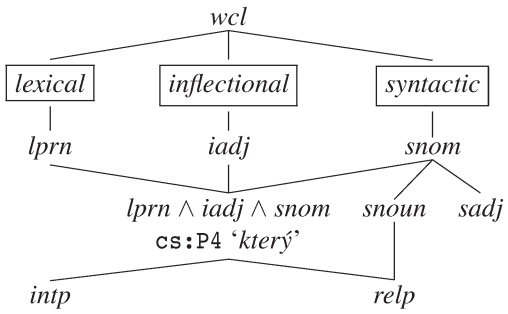

Figure 3: A single node for interrogative and relative pronouns

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. 4 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.

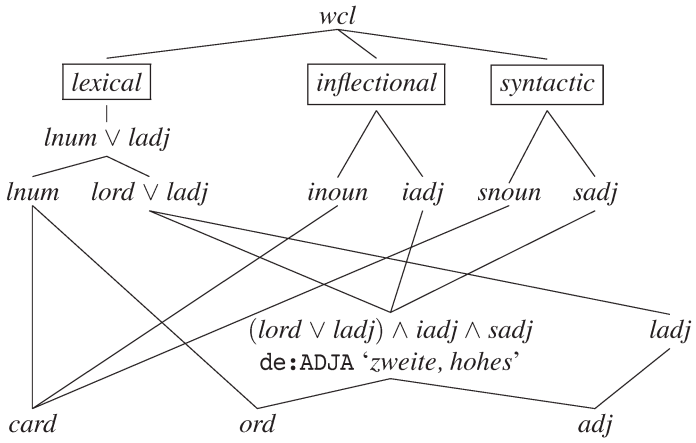

Figure 4: A single node for ordinal numerals and adjectives

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 per-

son, singular and feminine, while *inflectionally* it is masculine or neuter, singular, genitive or accusative.[8] 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).[9] 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 3 is the formal context for our previous example of adjectives and numerals (Fig. 4). Attributes corresponding to the boxed labels in Fig. 4 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 (Table 4). 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. 5). Its geometry is significantly simpler than the hierarchy constructed intuitively (as in Fig. 4), but the concept ambiguous between adjectives and cardinal numerals is still there. The latter two steps can be done automatically.[10]

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

Table 3: 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*}>

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

9 For an overview of linguistic applications of FCA see Priss (2005). Janssen (2004) is concerned with a lexical interlingua, similar to our hierarchy of linguistic categories.

10 See http://www.fcahome.org.uk/fca.html.

3	<{card},	{*lnum,inoun,snoun*}>
4	<{},	{*ladj,lnum,iadj,inoun,sadj,snoun*}>

Table 4: Formal concepts derived from Table 1

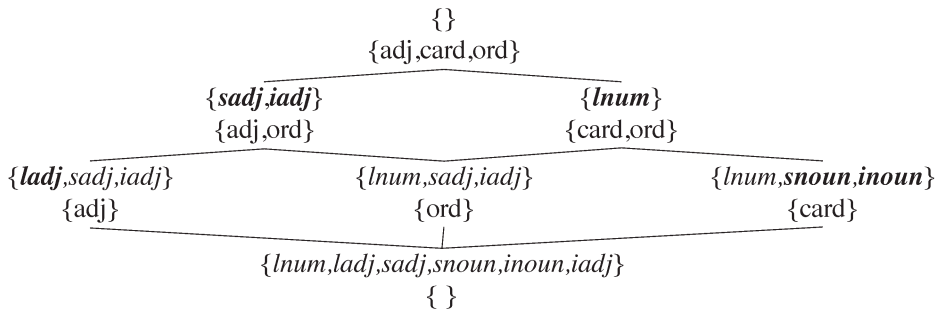

Figure 5: Concept lattice for adjectives and ordinal numerals

Attributes specified for an object in a formal context are interpreted in conjunction. 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 5 and the corresponding lattice in Fig. 6.

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

Table 5: Formal context for numerals and pronouns

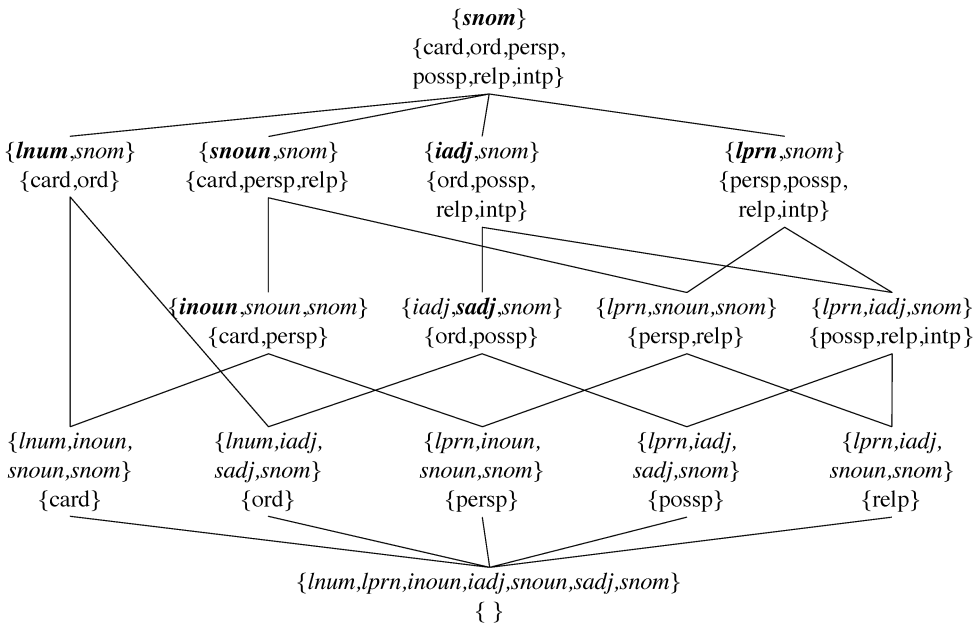

Figure 6: Concept lattice for numerals and pronouns

Lattices can be used for reasoning about attributes, as in the implications *ladj*⇒*sadj* or *snoun*⇒*lnum*, refering to Fig. 5. 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".

4 Projecting annotation via word-to-word alignment

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 coun-

terparts, 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. Tables 6 and 7 show a list of most frequent pairs of tags in word-aligned subsets of the English-Czech and Polish-Czech sections of *InterCorp*, comparing the frequency of the pair with the frequency of either tag in the sample, counted independently. Rows including the relevant pairs are greyed, with the more general tag that could be made more specific by the projection set in bold. In some cases the profit can be mutual even within a single pair, as in the Czech Polish pairs PD-adj and PD-subst.

In 1.5 million English-Czech word-to-word alignments, 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 conjunction 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.

tags cs	freq cs	cs-en/cs	tags en	freq en	cs-en/en	freq cs-en
NN	356,512	55.07	NN	249,501	78.69	196,324
RR	94,871	75.17	**IN**	108,205	65.90	71,310
NN	356,512	18.84	NNS	75,906	88.48	67,163
AA	109,297	61.34	**JJ**	101,445	66.08	67,039
J^	81,564	79.95	CC	69,870	93.33	65,207
NN	356,512	16.23	NP	88,209	65.59	57,855
Vp	144,515	38.92	VVD	66,984	83.96	56,243
Db	71,090	53.53	RB	81,163	46.89	38,057

Vp	144,52	12.76	VVN	32,332	57.03	18,438
Vf	30,098	55.82	VV	43,022	39.05	16,801
PP	19,425	84.92	PP	68,377	24.12	16,495
Dg	30,428	53.55	RB	81,163	20.08	16,295
AA	109,297	13.89	NN	249,501	6.08	15,177
J,	81,564	18.33	**IN**	108,205	13.82	14,953
PS	16,491	90.10	PP$	26,163	56.79	14,859
X@	20,401	71.80	NP	88,209	16.60	14,647
PD	24,123	55.90	DT	37,932	35.55	13,484

Table 6: Most frequent tags in Czech-English word-to-word alignments

tags cs	freq cs	cs-pl/cs	tags pl	freq pl	cs-pl/pl	freq cs-pl
NN	420,685	89.13	subst	473,524	79.18	374,945
Vp	166,877	84.57	praet	156,372	90.25	141,130
RR	140,320	96.73	prep	155,074	87.53	135,738
AA	132,121	79.70	**adj**	194,285	54.20	105,298
J^	96,430	85.21	**conj**	150,451	54.62	82,169
VB	115,189	59.95	fin	81,460	84.77	69,051
Db	104,472	55.97	**qub**	123,525	47.34	58,474
J,	44,420	91.70	**conj**	150,451	27.07	40,732
P7	31,265	83.94	**qub**	123,525	21.25	26,244
Vf	34,297	74.55	inf	34,408	74.31	25,569
Dg	40,115	63.10	adv	42,190	60.00	25,313
VB	115,189	19.26	aglt	26,315	84.31	22,186
PD	42,982	44.47	**adj**	194,285	9.84	19,116
X@	29,700	59.18	subst	473,524	3.71	17,576
NN	420,685	4.17	ger	22,788	77.07	17,563
Db	104,472	14.75	conj	150,451	10.24	15,412
PD	42,982	30.35	**subst**	473,524	2.75	13,045
PP	26,260	45.59	ppron12	24,410	49.05	11,973

Table 7: Most frequent tags in Czech-Polish word-to-word alignments

5 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 *Interset* and *PAULA/ANNIS/OLiA* projects (Zeman, 2010; Chiarcos, 2008; Chiarcos *et al.* 2008). 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

Chiarcos Ch. 2008: An Ontology of Linguistic Annotations, in: "LDV Forum" (GLDV-Journal for Computational Linguistics and Language Technology), 23/2008, pp. 1–16.

Chiarcos Ch., Dipper, S., Götze M., Leser U., Lüdeling A., Ritz J., Stede, M. 2008: A Flexible Framework for Integrating Annotations from Different Tools and Tagsets, in: "TAL" (Traitement automatique des langues), 2/2008, pp. 217–246.

Díaz-Negrillo A., Meurers D., Valera S., Wunsch H. 2010: Towards interlanguage POS annotation for effective learner corpora in SLA and FLT, in: "Language Forum", 1–2/2010, pp. 139–154.

Erjavec T. 2009: MULTEXT-East Morphosyntactic Specifications: Towards Version 4, in: *Metalanguage and Encoding Scheme Design for Digital Lexicography,* (ed. R. Garabík), Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, pp. 59–70.

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

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

Janssen M. 2004: Multilingual Lexical Databases, Lexical Gaps, and SIMuLLDA, in: "International Journal of Lexicography", 2/2004, pp. 137–154.

Komárek M. 1999: Autosemantic Parts of Speech in Czech, in: *Travaux du Cercle linguistique de Prague,* vol. 3*,* (eds. E. Hajičová, T. Hoskovec, O. Leška, P. Sgall, Z. Skoumalová),

pp. 195–210.

Kotsyba N., Radziszewski A., Derzhanski I. 2009: 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*, Ljubjana, Slovenia, pp. 37–55.

Priss U. 2005: Linguistic Applications of Formal Concept Analysis, in: *Formal Concept Analysis. Foundations and Applications*, vol. 3626 of *Lecture Notes in Artificial Intelligence*, (ed. B. Ganter), Springer, Berlin/Heidelberg, pp. 149–160.

Przepiórkowski A., Woliński M. 2003: A flexemic tagset for Polish, in: *Proceedings of Morphological Processing of Slavic Languages, EACL 2003*, pp. 33–40.

Zeman, D. 2010: Hard Problems of Tagset Conversion, in: *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, (eds. A. Fang, N. Ide, J. Webster), Hong Kong, pp. 181–185.