ГАРМОНИЗАЦИЯ СИСТЕМ ПОМЕТ ДЛЯ МНОГОЯЗЫЧНЫХ КОРПУСОВ ПОСРЕДСТВОМ РЕШЕТКИ ПОНЯТИЙ

HARMONIZING TAGSETS FOR MULTILINGUAL CORPORA VIA CONCEPT LATTICE*

*Rosen A. (*alexandr.rosen@ff.cuni.cz*)*

Institute of Theoretical and Computational Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic

Сравнение систем морфосинтаксических помет обнаруживает различные предположения, заслоняющие сходства и различия между языками. Чтобы преодолеть формальные и концептуальные несоответствия, мы строим абстрактную межъязыковую систему помет как иерархию категорий, используя анализ формальных понятий.

1 Introduction

Multilingual corpora can be annotated with morphosyntactic tags by monolingual tools. However, each of the tools is typically bundled with a specific tagset. This variety of tagging schemes may be a problem for the user: *InterCorp*, a parallel corpus, currently offers on-line concordances in 22 languages, 11 of them tagged with 11 different tagsets.[1] Fig. 1 illustrates the tagset variety using comparable examples of prepositional phrases in all of the 11 presently tagged languages.[2]

We are aiming at a solution that would delegate the task of dealing with multiple tagsets to the system, allowing the user to interact with an abstract interlingual hierarchy of linguistic categories. In order to reflect the differences between various tagsets, the common "tagset" takes three different perspectives 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.

Tags in all tagsets can be described as objects with properties and the methods of Formal Concept Analysis [2] 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 employed by Janssen [3] for dealing with lexical gaps in a multilingual lexical database.

This is certainly not the first attempt to design an interlingual representation of linguistic categories in the context of multilingual corpora. We wish to mention at least *MULTEXT-East* [4], whose tagging scheme became a *de facto* standard for inflectional languages, and *Interset*,

* Work on this project was supported by grant no. MSM0021620823 of the Czech Ministry of Education, Youth and Sports.

[1] For more details about the project see [1] or the project site at http://korpus.cz/intercorp/. The corpus can be queried at korpus.cz/Park after registration at http://ucnk.ff.cuni.cz/english/dohody.php.

[2] For details about the tagging tools and tagsets see http://korpus.cz/english/intercorp-info.php. Here and below, Czech positional tags are truncated: **RR-6** stands for **RR-6**---------- (tag for a preposition selecting local case).

a truly interlingual tagset [5], designed primarily for translating tags from one tagset into another. However, neither quite satisfies our requirements: they miss some categorial correspondences between languages and do not support the idea of arbitrary levels of specificity.

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
	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

2 Word classes in three flavours

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 decline independently in typical nominal positions, referring to entities; attributive or predicative adjectives, representing properties, 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 abandoning the traditional list in favour of a cross-classification along the three dimensions seems attractive. Distinctions between the three aspects are borne out also by the tagsets. Our tagset for Czech has a preference for lexically-based classification, the Polish tagset for inflectional word classes, the German tagset distinguishes pronouns by their syntactic function.

Fig. 2 shows a simple case – nouns and adjectives are nouns and adjectives, respectively, on all three criteria.[3] The topmost node *wcl* stands for both nouns and the adjectives. Its daughters are labelled by the three aspects: *lexical* (for 'semantic'), *inflectional* (for 'morphological') and *syntactic*.[4] The boxes around the labels suggest that the sets of objects denoted by the nodes have a non-empty intersection. In fact, all four sets involved are

[3] All hierarchies shown here are partial: they cover only a fraction of morphological categories and languages.
[4] We use *lexical* rather than *semantic* – *lexical* word classes have their properties specified in the lexicon.

identical, which is a feature of cross-classification. The other nodes stand for 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.

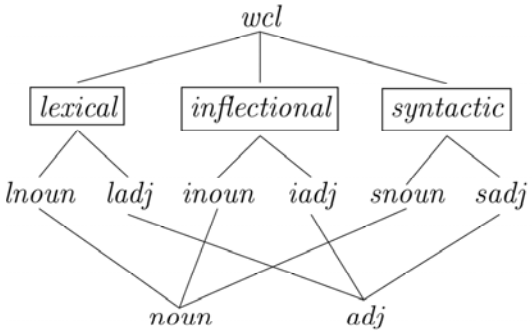

Figure 2: Nouns and adjectives are nouns and adjectives from all three aspects

The hierarchy of categories or *types* is partially ordered by their specificity. Each type denotes a set of objects – language-specific tags, identified by their name and specific tagset. The topmost type denotes all tags in all tagsets. 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 its supertypes.

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. 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 – 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.'

To express this triple membership, the Czech tag **P4** for relative pronouns[5] is a subtype of the cross-classifying word classes, each representing a different dimension – see fig. 3.

[5] We ignore all but the first two positions in the tag.

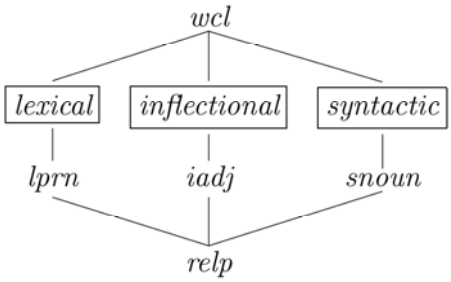

Figure 3: A hierarchy fragment for the Czech relative pronoun který *'which'*

The fragment can be extended by other objects as in fig. 4: cardinal and ordinal numerals, personal, possessive and interrogative 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 *který* in its relative use 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*.

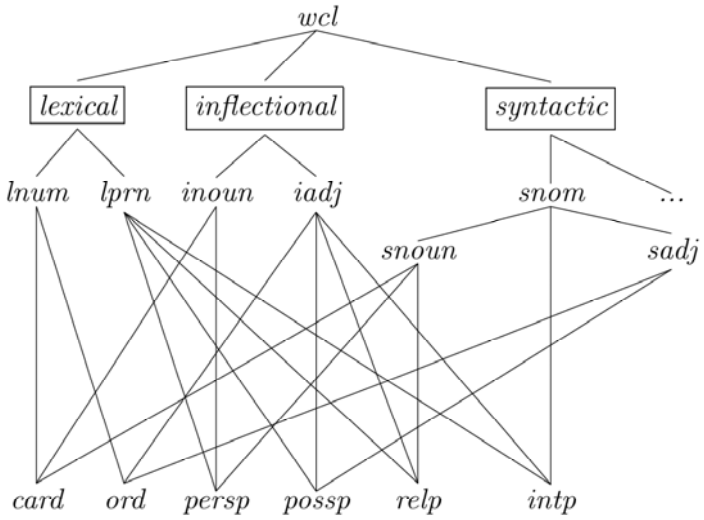

Figure 4: Distinguishing types of numerals and pronouns in a hierarchy

However, there is a single Czech tag covering both the relative and the interrogative use of *který* (**P4**), which should be 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. 5 captures this ambiguity. The Czech tag **P4** corresponds to a node labelled *lprn* ∧ *iadj* ∧ *snom*.

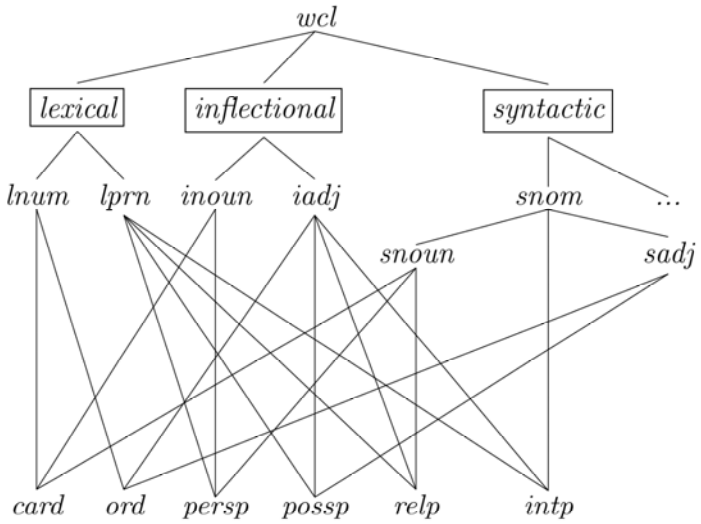

Figure 5: A single node for interrogative and relative pronouns

The three views of word class allow for proper mapping between language-specific tagsets. The tag for adjective in the English, German, French, Italian and Polish tagsets covers also ordinal numerals. If all these tags are mapped 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. 6 shows a fragment of the hierarchy with a node representing both ordinal numerals and adjectives, labelled $(lord \vee ladj) \wedge iad\,j \wedge sadj$ and corresponding to the German tag **ADJA**.

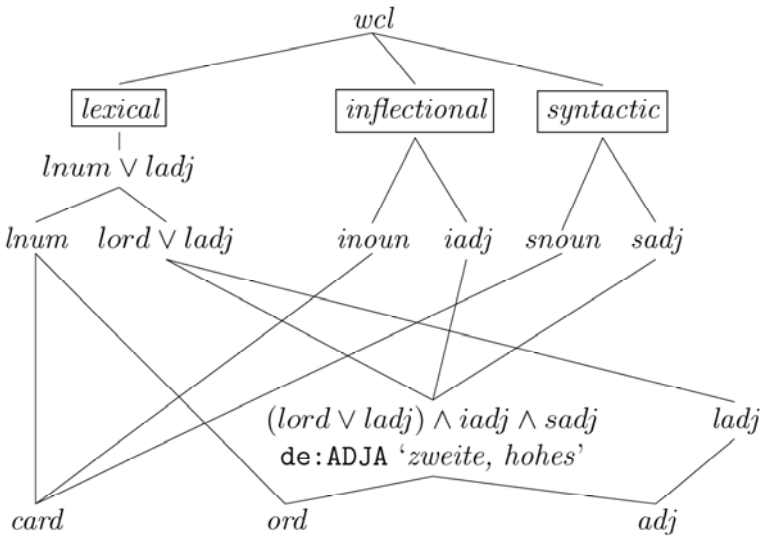

Figure 6: A single node for ordinal numerals and adjectives

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* ∨ *lord*).

Partial hierarchies can be merged. The result of merging the above two hierarchies (figures 5 and 6) is shown in fig. 7.

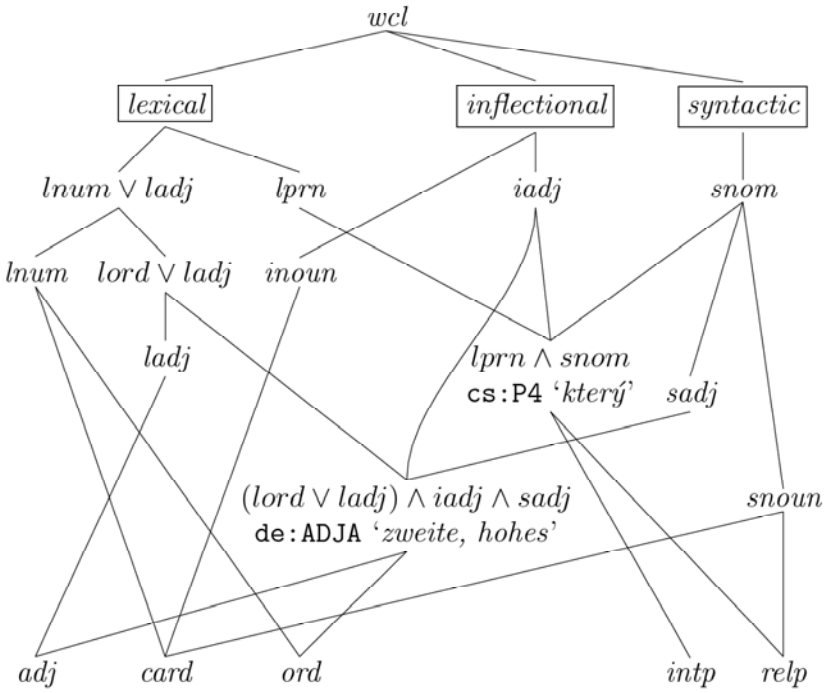

Figure 7: Hierarchies in figures 5 and 6 merged

We have barely 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.

3 Morphological categories

Tags often encode more information than just word class. Word class of any flavour may be required to co-occur with a set of other 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 possessive pronoun such as *jejího* is *lexically* 3rd person, singular and feminine, while *inflectionally* it is masculine or neuter, singular, genitive or accusative (2).[6]

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

(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.'

The set of categories appropriate to a word class may be defined as types in the hierarchy, which further cross-classify types corresponding to language-specific tags. Then the user can refer to all plural items by specifying them merely as **pl**.

The tag for the Czech possessive pronoun *jejího* in fig. 8 is a subtype of lexical pronoun (*lprn*) and inflectional adjective (*iadj*).[7] As a possessive pronoun, it is required by the specification of the hierarchy[8] to be a subtype of *lexical* gender (*lgend*), number (*lnum*) and person (*lpers*), more precisely of their intermediate subtypes, specifying the morphological categories. As an inflectional adjective, it is required to be a subtype of *inflectional* gender *igend*, case (*icase*) and 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*.

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

[8] More general co-occurrence restrictions could be specified at a meta-level to ease the initial manual task of mapping tags to categories.

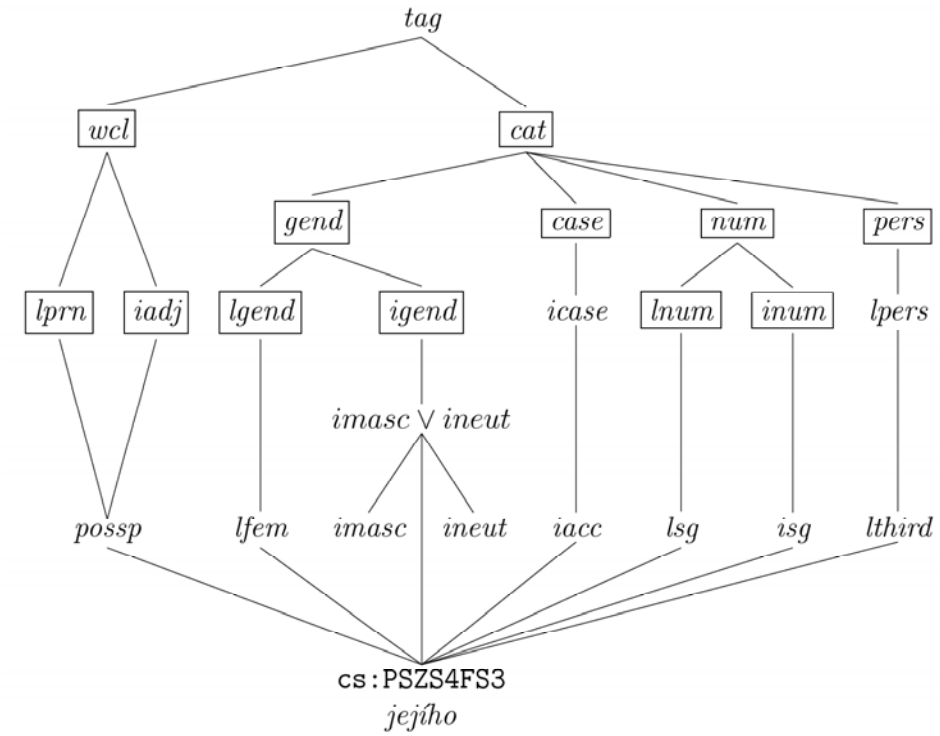

Figure 8: Morphological categories used to tag a Czech possessive pronoun jejího*, a category-based view*

The hierarchy in fig. 8 leaves the lexical/inflectional distinction implicit. In fig. 9 this distictions is shown at the top level, as in all previous hierarchies. For clarity, general category labels (*gend*, *case*, etc.) are omitted.

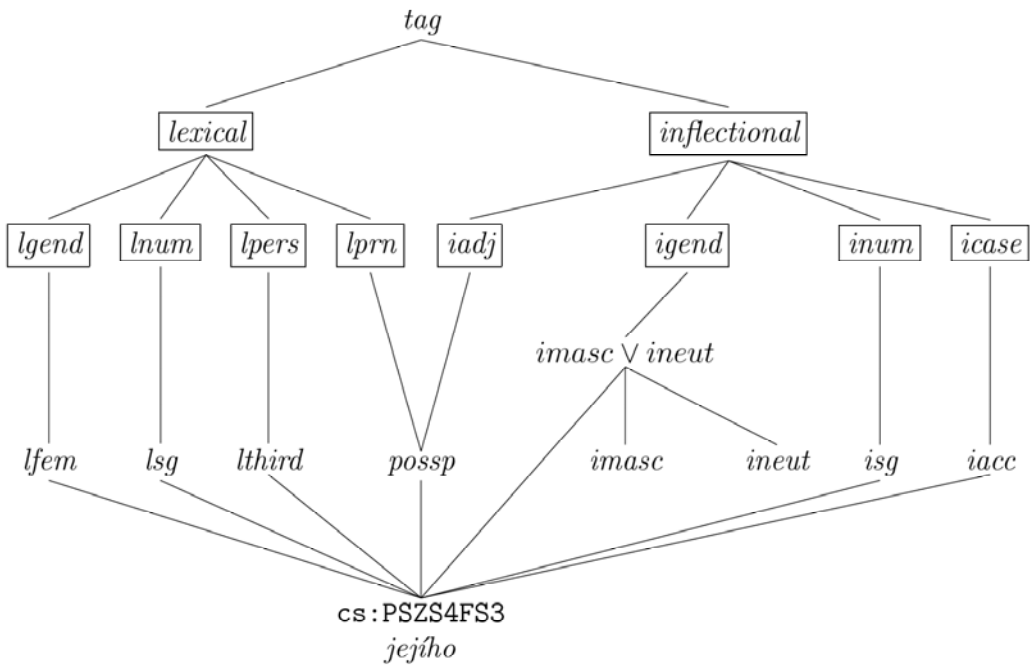

Figure 9: Morphological categories used to tag a Czech possessive pronoun jejího*, a lexical/inflectional view*

4 Building and using the common tagset

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 [2,6]. 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.

The first step of the analysis is to identify the objects and their 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 numerals (fig. 6). Attributes corresponding to the boxed labels in fig. 6 are omitted: they would be specified for all objects 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, and a set of attributes. 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 11

Finally, the concept lattice can be drawn (fig. 10). Its geometry is significantly simpler than the hierarchy constructed intuitively (as in fig. 6), while the concept ambiguous between adjectives and cardinal numerals is still present. The last two steps can be done automatically.[9]

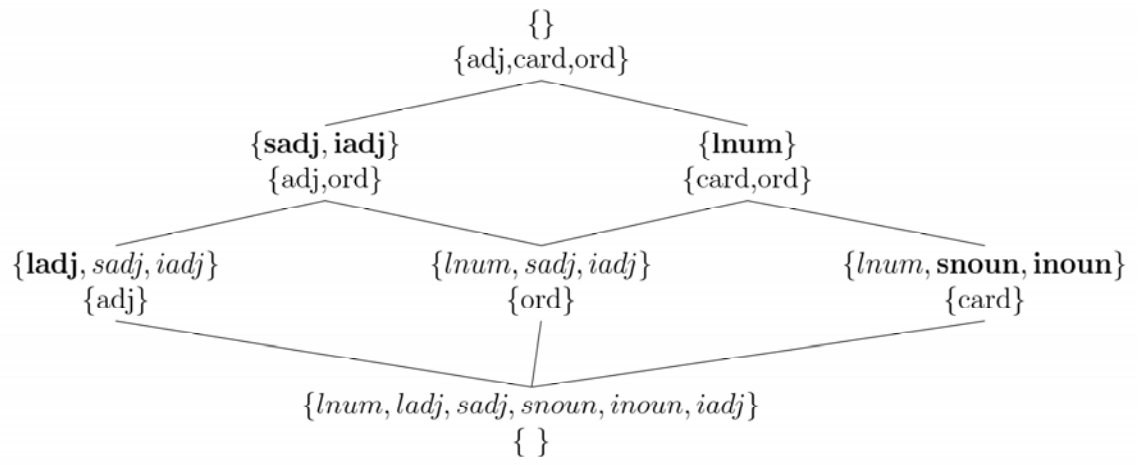

Figure 10: Concept lattice for adjectives and ordinal numerals

The concept lattice can be used for reasoning about attributes, as in the following implications: *ladj* ⇒ *sadj* or *snoun* ⇒ *lnum*. Such statements can be used 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. To look up its Czech equivalent we have to find a Czech tag corresponding to the {adj,ord} concept. In the absence of such a tag,

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

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 **cs:Nr** is similar to the scenario when the user asks for "ord" in a German text. It's easy in a Czech text, 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} 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 conjuction. Thus, specifying both *snoun* and *sadj* as attributes of an interrogative pronoun (intp) would mean that it is simultaneously 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. 11.

	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

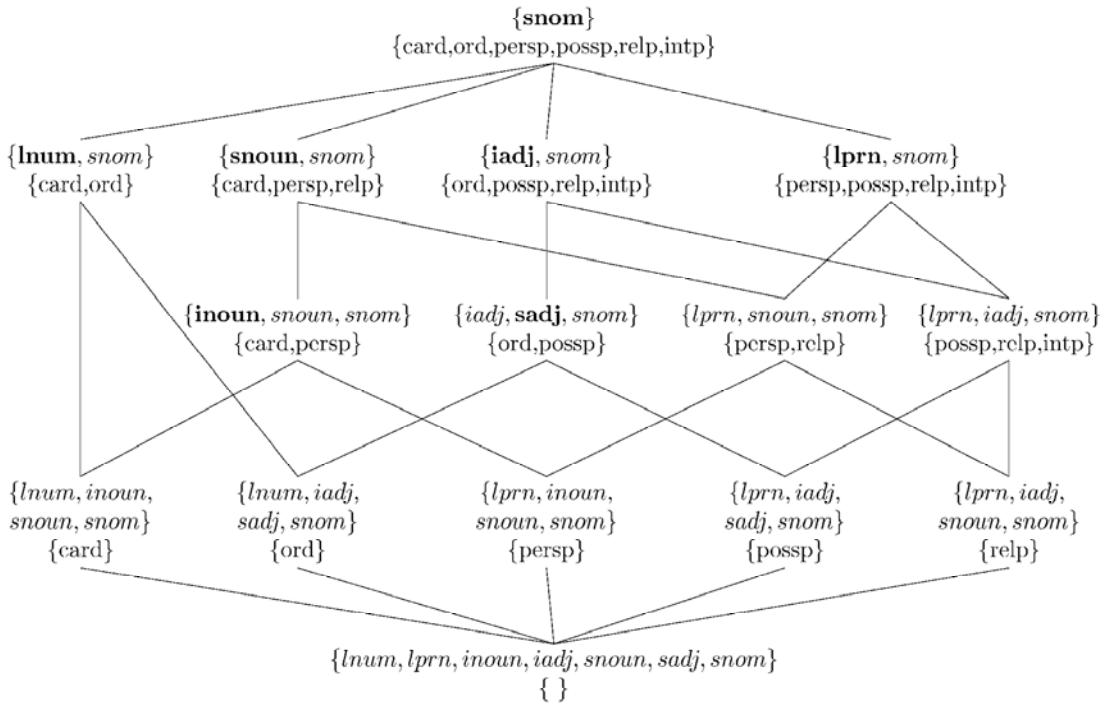

Figure 11: Concept lattice for numerals and pronouns

This is not the first application of FCA in the field of linguistics, not even in a multilingual setting. Priss [7] gives an overview of linguistic applications of FCA and Janssen [3] is concerned with multilingual lexical databases. His lattice, a structured lexical interlingua connecting words from different languages, is 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.

5 Conclusion

A solution to the problem of tagset variety in a multilingual corpus can be an abstract, hierarchically structured interlingual tagset, based on a three-way distinction in the system of word classes, allowing for intuitive and underspecified queries and principled mappings between different language-specific tagsets. 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.

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 customising 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.

References

1. Vavřín M., Rosen A. InterCorp: A Multilingual Parallel Corpus Project // Proceedings of the International Conference Corpus Linguistics – 2008. St. Petersburg State University, 2008. P. 97–104
2. Ganter B., Wille R. Formal Concept Analysis. Mathematical Foundations // Berlin/Heidelberg: Springer, 1999.
3. Janssen M. Multilingual Lexical Databases, Lexical Gaps, and SIMuLLDA // International Journal of Lexicography, 2004. 17 № 2.
4. Erjavec T. Harmonised Morphosyntactic Tagging for Seven Languages and Orwell's 1984 // Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, NLPRS'01, 2001. P. 481–492.
5. Zeman D. Reusable tagset conversion using tagset drivers // Proceedings of the Language Resources and Evaluation Conference, LREC 2008, Marrakech, Marocco, 2008.
6. Wille R. Formal concept analysis as mathematical theory of concepts and concept hierarchies // Ganter B., editor, Formal Concept Analysis. Foundations and Applications, volume 3626 of Lecture Notes in Artificial Intelligence, Berlin/Heidelberg: Springer, 2005. P. 1–33.
7. Priss U. Linguistic applications of formal concept analysis // Ganter B., editor, Formal Concept Analysis. Foundations and Applications, volume 3626 of Lecture Notes in Artificial Intelligence, Berlin/Heidelberg: Springer, 2005. P. 149–160.