Harmonizing Tagsets for Multilingual Corpora via Concept Lattice

Alexandr Rosen

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

Dialogue 2010, Bekasovo
29 May 2010

The problem

- In a multilingual parallel corpus, each language comes with a potentially incompatible morphosyntactic tagset.
- The variety of tagsets is a problem for human users and for applications, too.

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))**

Reasons for the variety of tagsets

- Different labels for the same concept
- Typological differences among languages
- Different choice of a theory or its aspect

A solution

- Map each tag in every language to a common abstract tagset, a kind of tagset interlingua.
- Use *Formal Concept Analysis* to build a hierarchy of categories, linked to language-specific tags.

Linguistically motivated choices can be reflected in the common tagset

- In some tagsets, adjectives, ordinal numerals and possessive pronouns are all in a single class.
- In other tagsets, they are different lexical classes.
- Anyway they share the property of being a syntactic or inflectional adjective.

$\longrightarrow$ We propose a 3-way cross-classification of word classes.

Cross-classification

	lexical	inflectional	syntactic
ordinal numeral *pátý*	num	adj	adj
cardinal numeral *pět*	num	noun	noun
personal prn *ty*	prn	noun	noun
possessive prn *tvůj*	prn	adj	adj
relative prn *který*	prn	adj	noun
interrogative prn *který*	prn	adj	noun/adj
adverbial pple *volajíce*	verb	pple	advb

Further support for the cross-classification: sets of morphological categories appropriate to cross-classified POS

- Czech possessive pronouns show pronoun–antecedent agreement as lexical pronouns
- and attribute–noun agreement as syntactic adjectives

(1) Jana přišla, ale **jejího** syna jsem neviděl.
Jana$_{\text{FEM,3,NOM}}$ came but her$^{\text{FEM,3}}_{\text{MASC,ACC}}$ son$_{\text{MASC,ACC}}$ I haven't seen
'Jana has come, but I haven't seen her son.'

Multiple inheritance hierarchy

- Sets of categories as partially ordered types in multiple inheritance ISA hierarchy.
- Each type denotes a set of language-specific tags.
- 'Translational' ambiguities due to mismatching tags are not resolved but properly represented.

Formal Concept Analysis (FCA)

- A logical formalism and tool to build the hierarchy and to do reasoning about categories.
- FCA classifies objects (tags) according to their properties (categories).

Formal context for adjectives and cardinal/ordinal numerals

- A table to identify objects and their attributes.

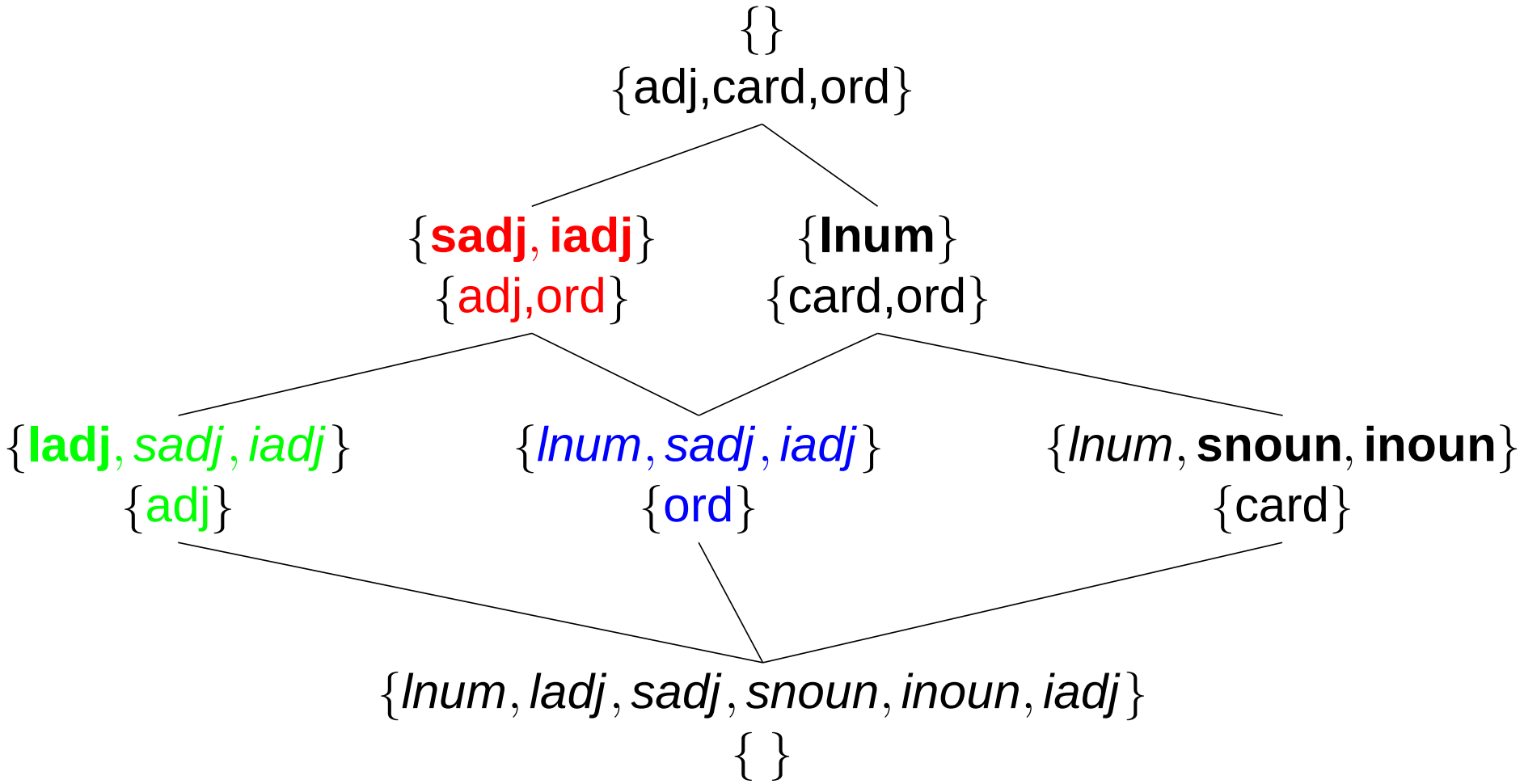
	ladj	*lnum*	*iadj*	*inoun*	*sadj*	*snoun*
adj	✓		✓		✓	
ord		✓	✓		✓	
card		✓		✓		✓

Formal concepts derived from the formal context table

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*}⟩

- **Concept**: a set of objects and a set of attributes.
- Objects of a concept are also part of its superconcept.
- Concepts are partially ordered by specificity (roughly: the more attributes, the more specific).

Concept lattice for adjectives and ordinal numerals



Reasoning about attributes

- Possible implications: *ladj* $\Rightarrow$ *sadj* or *snoun* $\Rightarrow$ *lnum*
- ... useful for general queries ("show me all adjectives")
- ... or to match incompatible language-specific tags

Problems

- Attributes in FCA formal contexts interpreted in conjunction, for disjunctive attributes we have to introduce a more general attribute, perhaps in a pre-processing step.
- The main task: manually specifying formal context for all tags.