Компютърна компаративистика

Езикови технологии за славистични изследвания

SlavicNLP@online.de

ст.н.с. II ст. д-р **Елена Паскалева**, ИПОИ БАН

PD Dr. phil. habil. **Таня Августинова**, DFKI & Saarland University

Discussion based on [Przepiórkowski 2007]
http://langtech.jrc.it/BSNLP2007/m/BSNLP-2007-proceedings.pdf

- Information Extraction (IE) often involves partial syntactic processing
  - highlevel IE → about who did what to whom (when, where, how and why)
  - simpler IE → finding company names in texts
- Overview of Slavonic phenomena which
  - (i) pose particular problems for IE and partial parsing
  - (ii) seem easier to treat in Slavonic than in Germanic or Romance

*N.B. many of the typological features distinguish:

1. East Slavonic (Russian, Ukrainian, Belorussian, Rusyn), West Slavonic (Czech, Slovak, Upper and Lower Sorbian, Polish, Kashubian) and the Western subgroup of South Slavonic (Croatian, Bosnian, Serbian, Slovenian)
2. the Eastern subgroup of South Slavonic (Bulgarian and Macedonian)

"Slavonic is hard"

Rich Nominal Inflection

The rich nominal inflection of Slavonic makes already the most basic IE task, namely Named Entity Recognition (NER), more difficult than in Germanic or Romance. Slavonic nouns, apart from inflecting for number (singular and plural; in Slovenian and Sorbian also dual), famously inflect for about 6 (e.g., Russian, Slovenian) or 7 (e.g., Czech, Croatian, Polish, Ukrainian) cases: the exact number of cases cited in the literature for any particular language often depends on the granularity of description, so Belorussian and Slovak have either 6 or 7 cases, depending on the inclusion in the description the rare vocative forms, among the 7 Serbian cases, dative and locative are sometimes conflated because they "only" differ in accent, the Polish case system may be extended to 8 cases by postulating the distributive case (Gruszczyński, 1989, p. 89), while the number of Russian cases may also be reasonably increased to 8 by adding a second genitive and a second locative case (Jakobson, 1958).

"Slavonic is hard"

Rich Nominal Inflection

While for many European languages a dictionary of lemmata of proper names is sufficient for the task of NER, (Steinberger and Pouliquen, 2007, §3.3) note that "a minimum of morphological treatment" is required for languages with rich nominal inflection, such as Balto-Slavonic or Finno-Ugric languages. Unfortunately, for the majority of Slavonic languages, there are no (freely) publicly available resources that could provide such "minimum morphological treatment" of proper names. For example, the only large free (but not open source) morphological analyser for Polish, Morfeusz (Woliński, 2006), contains very few proper names.[3] Moreover, the NE content of commercial analysers is often rather low, so that simple resource-light heuristics sometimes give better results (Urbańska and Mykowiecka, 2005, p. 214). Such heuristics usually involve the creation of inflected forms by adding typical suffixes (Popov et al., 2004; Urbańska and Mykowiecka, 2005; Steinberger and Pouliquen, 2007), where the suffix addition/substitution rules are either hand-generated (Urbańska and Mykowiecka, 2005) or automatically acquired (Steinberger and Pouliquen, 2007).

"Slavonic is hard"

Different Inflection of Homonymous Common and Proper Nouns

As mentioned in (Piskorski, 2005) and discussed at length in (Piskorski et al., 2007b), many Polish surnames have the same base forms as common names, for example, GRZYB (lit. 'mushroom'), GOŁĄB (lit. 'pigeon') or KOWALSKI (lit. an adjective from 'smith'). This is a problem in itself in recognising proper names, but it is further exacerbated by the fact that such proper nouns may have different gender values, and different inflectional paradigms, than the corresponding common nouns. For example, while the common nouns GRZYB and GOŁĄB are, respectively, inanimate masculine and animate masculine (cf. fn. 5), the corresponding surnames are virile or feminine, depending on the denotation; in case of singular feminine names, they would not overtly inflect at all, while in case of singular masculine or plural uses, the forms are often different than corresponding common forms, e.g., the accusative singular and plural forms of GOŁĄB would be *gołębia* and *gołębie*, when used as a common noun, and *Gołąba* and *Gołąbów*, when used as a surname, etc. Obviously, once properly described, such inflectional differences may actually help in NER.

"Slavonic is hard"

Difficult Inflection of Foreign Names

A problem relatively minor in comparison to other problems discussed here is the inflection of foreign names: although it is governed by strict prescriptive rules, native speakers are often unaware of them and different variants of the same form may be encountered in text; for example, while in Polish the correct spelling of the singular instrumental form of LINUX is *Linuksem*, the variant *Linuxem* is at least as common, and the starkly wrong *Linux'em* and *Linux-em* are also quite frequent. Similarly, probably few Poles realise that the correct locative forms of BRANDT and PEIRCE are *Brandcie* and *Peirsie*, and not, say, *Brandtcie* and *Peirce'ie*, and that although the locative of REMARQUE is *Remarque'u*, the instrumental is *Remarkiem*.[4] A comprehensive NER should be able to deal with various incorrect forms of foreign NE occurring in Slavonic texts.

"Slavonic is hard"

Difficult Inflection of Foreign Names

On the other hand, the inflection of proper names depends on their pronunciation, i.e., on their origin. For example, the genitive of CHARLES is either *Charlesa* or *Charles'a*, depending on whether it is an English name or a French name. Another example, from (Piskorski et al., 2007b), is WILDE, whose genitive form is either *Wilde'a* (English) or *Wildego* (German). This feature, when properly encoded, may actually help distinguish between entities in NER.

"Slavonic is hard"

Tagset Size and Syncretisms

A rich inflection system also implies that the size of the tagset is very large. For example, given that a Polish nominal form may have one of 2 numbers, one of 7 cases and one of 5 genders,[5] there are 70 possible nominal tags, not counting gerundial and pronominal forms. In fact, there are 4179 possible tags in the IPI PAN Tagset of Polish (Przepiórkowski and Woliński, 2003a; Przepiórkowski and Woliński, 2003b), of which around 1150 occur in nature (Przepiórkowski, 2006b). Similarly, sizes of Czech tagsets range from 1171 (Hajič and Hladká, 1997), through 1631 (Pala et al., 1998), to theoretically 4257, but "only" about 1100 actually used (Mirovský et al., 2002). Such detailed tagsets make it difficult to reach high accuracy, which — on the assumption that syntactic parsing is preceded by full morphosyntactic disambiguation — has negative influence on syntactic processing.

"Slavonic is hard"

Tagset Size and Syncretisms

Another problem connected to the rich inflection system of Slavonic languages is the large number of syncretisms. For example, a typical Polish adjective may have 11 textually different forms (e.g., for BIAŁY 'white': *biali*, *biała*, *białą*, *białe*, *białego*, *białej*, *białemu*, *biały*, *białych*, *białym*, *białymi*), but as many as 70 different tags (2 numbers × 7 cases × 5 genders). There are also various systematic nominal syncretisms which to some extent annul the advantages that rich case system presents for the identification of grammatical roles. For example, in plural, Polish non-virile (non-human-masculine) nouns have the same form in the nominative and in the accusative, while in the singular, inanimate masculine and neuter forms do. Similarly, virile and animate masculine nouns have the same singular accusative and singular genitive forms. So, for example, in the rather artificial sentence *Samochody dwie minuty wyprzedzają autobusy* '(The) cars (for) two minutes are overtaking (the) buses', each of *samochody*, *dwie minuty* and *autobusy* may be interpreted as either nominative or accusative, i.e., as the subject (nominative), the object (accusative) or a temporal adjunct (accusative).

"Slavonic is hard"

Numeral Phrases

An area of Slavonic syntax very well-known in theoretical linguistics is the syntactic behaviour of NumPs (Corbett, 1978; Franks, 1995); numerals also turn out to be awkward for automatic processing in various ways.[6]

First, the case of the noun (phrase) within an NumP depends on the numeral[7] and on the position of the whole NumP in the sentence. For example, for NumPs in the subject position, the noun is in the nominative case, roughly, if the numeral is or ends in 2, 3 or 4 (with the exception of 12, 13 and 14), and it is genitive otherwise.[8] This means that the shallow processor should recognise as a possible currency quantity the sequence *152 dolary* and *155 dolarów*, but not **152 dolarów* or **155 dolary*.[9]

"Slavonic is hard"

Numeral Phrases

Second, in case of "typical" numerals (not ending in 2, 3 or 4), the Polish NumP in subject position does *not* agree with the verb; instead, the verb occurs in the default 3rd person singular neuter form,[10] which may make discovering the subject-verb relation more difficult.

Finally, and rather marginally, "typical" NumPs in copular constructions trigger very atypical agreement with the predicative adjective, e.g.: *40 głosów było nieważnych/nieważne* '40 votes be-3RD.SG.NEUT invalid-PL.GEN/ACC'. It is easy to overlook such constructions when developing a shallow grammar, and — since they are rare — it is difficult to learn them automatically from corpora.

"Slavonic is hard"

Free Word Order

Last but certainly not least, the relatively free word order[11] makes the discovering of who did what to whom (when, where, how and why) much more difficult than finding the relative order of NPs and PPs in the sentence. It may seem that the rich case system may help here, as — with active forms of verbs — subjects are usually nominative and objects are often accusative, but matters are much more complicated because of the widespread syncretisms mentioned in §2.4, esp. the systematic nominative-accusative and accusative-genitive syncretisms, and because both complements and adjuncts may be expressed by the same cases (e.g., accusative temporal adjuncts may look like objects of transitive verbs).

While the relatively free word order is seriously felt in deep parsers and leads to the multiplication of analyses, to the best of our knowledge most IE work in Slavonic to date has concentrated on lower-level tasks such as NER and, hence, has not yet tried to systematically deal with this problem.

"Slavonic is easy"

On a more positive note, the rich Slavonic inflectional system may help at the higher levels of processing. There are various linguistic phenomena where overt case, gender and number agreement allows to differentiate between interpretations and, hence, to extract the information about who did what to whom. To give two trivial constructed examples: the English sentence *I saw him drunk* is ambiguous in ways that are necessarily disambiguated by the two Polish translations of that sentence: *Widziałem go pijany* '(I) saw him drunk-NOM' and *Widziałem go pijanego* '(I) saw him drunk-ACC'. Perhaps more interestingly, the lexical aspect of Slavonic verbs may make conspicuous the meanings which are only implicit in other languages, as in the Polish *Skoczył na stół* '(He) jumped-PERF on (the) table-ACC' versus *Skakał na stole* '(He) jumped-IMPERF on (the) table-LOC', both translated into the English *He jumped on the table*.

"Slavonic is easy"

One phenomenon important for high level IE where the rich inflectional system plays a positive role, however, is coordination.

Slavonic rich inflection makes the processing of such potentially coordinate structures easier. For example, case disagreement between two apparently coordinated NPs is a strong clue that they in fact belong to separate coordinated clauses, while agreement is a (perhaps weaker) clue that they form an actually coordinated NP.[12] Similarly, (dis)agreement in case, number and gender may help decide whether two apparently coordinated adjectival forms actually form a coordinate structure.