Slavonic Information Extraction and Partial Parsing

Adam Przepiórkowski
Insitute of Computer Science
Polish Academy of Sciences
Ordona 21, 01-237 Warsaw, Poland
adamp@ipipan.waw.pl

Abstract

Information Extraction (IE) often involves some amount of partial syntactic processing. This is clear in cases of interesting high-level IE tasks, such as finding information about who did what to whom (when, where, how and why), but it is also true in case of simpler IE tasks, such as finding company names in texts. The aim of this paper is to give an overview of Slavonic phenomena which pose particular problems for IE and partial parsing, and some phenomena which seem easier to treat in Slavonic than in Germanic or Romance; I also mention various tools which have been used for the partial processing of Slavonic.

1 Introduction

The aim of this paper is to give a general but rather biased overview of the problems of Information Extraction (IE) in Slavonic. In particular, I discuss linguistic phenomena which make IE in Slavonic harder than in Germanic or Romance, §2, but also those which seem to make it easier, §3. We will also look at various general tools which have been used in IE tasks in the context of Slavonic languages, especially at tools for partial (or shallow) parsing, §4.

I deal mainly with Polish, as a good representative of the Slavonic family: although Polish is a relatively large language, with about 44 million native speakers world-wide (over 38 million in Poland), the availability of linguistic resources and tools for this language does not reflect this fact: it compares unfavourably with Czech, and probably favourably with, say, Ukrainian.

2 Slavonic is Hard

There are various characteristics of Slavonic languages[1] that make them more difficult for automatic processing, whether shallow or deep, than Germanic and Romance languages.[2] The two of them which are most conspicuous, and identified as most problematic, e.g., in (Collins et al., 1999), are rich nominal inflection (§2.1) and free word order (§2.6). Others, causing problems to varying extents, include: idiosyncratic inflection of Slavonic proper names (§2.2); unstable inflection of some foreign names (§2.3); high degree of trans- and, especially, intra-paradigmatic syncretisms (§2.4); and, on the more syntactic level, the infamous quirkiness of numeral phrases (NumPs; §2.5).

2.1 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 in-

[1] Many of the typological features discussed below distinguish between, on the one hand, 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), and, on the other hand, the Eastern subgroup of South Slavonic (Bulgarian and Macedonian). In this and the next section I concentrate on the former group of Slavonic languages.

[2] By shallow or, equivalently, partial processing, I mean the task of finding *some* syntactic structure *without* using lexical resources such as valence dictionaries; by contrast, deep processing involves finding the *complete* sentence structures *with* the use of such lexical resources.

Balto-Slavonic Natural Language Processing 2007, June 29, 2007, pages 1–10,
Prague, June 2007.

flecting 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).

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).

2.2 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.

2.3 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.

On the other hand, the inflection of proper names

[3] A new version of Morfeusz, containing a large dictionary of proper names, is being prepared, but it is currently not clear if it is going to be freely available for non-commercial research purposes (M. Woliński, p.c.).

[4] See `http://so.pwn.pl/`.

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.

2.4 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.

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).

2.5 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]

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]

[5] Traditionally, 3 genders were assumed for Polish, as for many other European languages, but (Mańczak, 1956) conclusively shows that at least 5 gender values must be adopted in Polish: *virile* (called also *m1*, *personal masculine* and *human masculine*), *animate masculine* (*m2*), *inanimate masculine* (*m3*), *neuter* and *feminine*. Although this repertoire of genders was only recently adopted in general dictionaries (Bańko, 2000), it is still rather conservative; e.g., (Saloni, 1976) proposes 9 genders.

[6] One of the largest formal grammars of Polish, (Świdziński, 1992), implemented as a wide coverage deep parser in (Woliński, 2004), does not deal with NumPs at all. Later modifications of the parser in (Ogrodniczuk, 2006) include some limited treatment of numerals.

[7] This property turned out to be problematic for adapting the GF Parallel Resource Grammar to Russian (Khegai, 2006).

[8] Another exception is JEDEN '1', which is actually an adjective, rather than a numeral (Przepiórkowski, 2006a). Also, the description above holds for non-virile genders, but is even more complicated for virile.

[9] The latter may occur in contexts like: ... *według paragrafu 155 dolary nie są środkiem płatniczym w Polsce* '... according to paragraph 155 dollars are not a valid currency in Poland'.

[10] I argue elsewhere (Przepiórkowski, 1996; Przepiórkowski, 2004b) that such NumPs in subject position actually bear the accusative case; hence, the lack of agreement.

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.

2.6 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.

3 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*.

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

Coordination is infamous both in theoretical linguistics and in Natural Language Processing (NLP); in fact, while recent years witnessed an increase of theoretical linguistic works on various aspects of coordination, it seems that NLP lags behind in addressing this phenomenon head on. One of the exceptions is (Dale and Mazur, 2007), which deals with the problem of identifying the number of Named Entities (NEs) in expressions of the form "X and Y", where X and Y are sequences of capitalised words, e.g.: "Victorian Casino and Gaming Authority" (single entity) or "American Express and Visa International" (two entities). (Dale and Mazur, 2007) note that the problem is statistically non-negligible, as around 5.7% of sequences of capitalised words with an optional conjunction (i.e., candidates for NEs) actually contain a conjunction. Similarly, (Rus et al., 2007, p. 229) discuss the bracketing problem in phrases such as "[soccer and tennis] player" and "navy and [marine corps]", noting that "[p]arsing base Noun Phrases ... is not handled by current state-of-the-art syntactic parsers". Another kind of coordination ambiguity is considered in (Steiner, 2006), namely, the "NP and NP" sequence as either an NP-coordination, or a part of sentential coordination (where the first NP is an object of the preceding verb and the second NP is the subject of the following verb).

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 ac-

[11] Of course, the term *free word order* as applied to Slavonic means that the word order is conditioned largely by information structure (i.e., not really free); modelling the constraints of information structure on word order is particularly important in text generation (Kruijff-Korbayová and Kruijff, 1999).

tually 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.

4 Slavonic is Processable

After discussing ways in which Slavonic languages seem to be hard or easy for Information Extraction, let us look at practical attempts at Slavonic IE, especially those involving partial parsing.

It seems that there have been relatively few attempts at applying shallow (or partial; cf. fn. 2) grammars to particular practical tasks. In some of these attempts no particular dedicated language processing system was used to implement shallow grammars: apparently they were coded directly in the host programming language.

One example is (Sharoff, 2004), where shallow parsing is used for the identification of prepositional Multi Word Expressions in Russian, with the following explanation of reasons for performing some language-dependent processing: "Given that the word order in Russian (and other Slavonic languages) is relatively free and a typical word (i.e. lemma) has many forms (typically from 9 for nouns to 50 for verbs), the sequences of exact N-grams are much less frequent than in English, thus rendering purely statistical approaches useless."

For Polish, simple shallow grammars were implemented for the tasks of question answering (Piechociński and Mykowiecka, 2005) and automatic valence acquisition (Fast and Przepiórkowski, 2005; Przepiórkowski and Fast, 2005); in the latter case a grammar was implemented as a cascade of Perl regular expressions. Similarly, (Zeman, 2001) describes a Perl regular expression implementation of a shallow preprocessor for a deep statistical parser. Much earlier, (Nenadić and Vitas, 1998; Nenadić, 2000) developed shallow grammars of Serbo-Croatian for the recognition of noun phrases (NPs) and certain kinds of coordinate structures. See also (Bekavac and Tadić, 2007) on the recognition of Croatian NEs with regular grammars.

Moreover, for Bulgarian a more general integrated system was developed, called LINGUA (Tanev and Mitkov, 2002), which — apart from modules for tokenisation, morphosyntactic analysis and disambiguation, and anaphora resolution — includes an NP extractor and a bottom-up grammar of Bulgarian. This system, together with a set of shallow patterns for identifying definition patterns, has been employed in a Question Answering prototype system (Tanev, 2004). Bulgarian pattern-matching grammars are also employed in (Koeva, 2007).

Apart from these language-specific implementations, there exist tools and toolboxes which facilitate various IE tasks, including shallow parsing. Probably the best known such a general system is GATE (Cunningham et al., 1995; Cunningham et al., 2002), which contains some NE resources for Bulgarian and Russian (Humphreys et al., 2002; Popov et al., 2004) and allows to write shallow (regular) grammars in the JAPE subsystem (Cunningham et al., 2000).

A system similar in scope is SProUT (Becker et al., 2002), whose shallow parsing language allows to write regular grammars over HPSG-style (Pollard and Sag, 1994) typed feature structures and which includes the operation of unification. Preliminary work on adapting SProUT to the processing of Baltic and Slavonic languages is presented in (Drożdżyński et al., 2003), with much subsequent work devoted to the processing of Polish, especially, in the area of Information Extraction from medical texts (Piskorski et al., 2004; Piskorski, 2004a; Piskorski, 2004b; Marciniak et al., 2005; Mykowiecka et al., 2005a; Mykowiecka et al., 2005b; Marciniak and Mykowiecka, 2007).

Although GATE and SProUT may be adapted to the processing of XML documents, they are perhaps not the most natural choice for the further processing of morphosyntactically annotated documents in, for example, the XCES (XML Corpus Encoding Standard; (Ide et al., 2000)) format, as assumed, e.g., in the IPI PAN Corpus of Polish (Przepiórkowski, 2004a), in the Slovak National Corpus (Garabík and Gianitsová-Ološtiaková, 2005), or in the LT4eL project (`http://www.lt4el.eu/`). Specialised XML-aware tools exist for such tasks.

One of the earliest collections of XML processing tools is the LT XML library (Brew et al., 2000), whose second edition, LT-XML2 is currently under preparation. One of the tools in that new edition, `lxtransduce` (Tobin, 2005), is an efficient pro-

[12] Again, this test may fail due to case syncretisms; cf. §2.4.

gram to add mark-up to XML files via regular grammars over XML elements; this tool is currently used for implementing definition-extraction grammars for Bulgarian, Czech and Polish (Przepiórkowski et al., 2007).

A system well-known in Slavonic NLP is CLaRK (Simov et al., 2001; Simov et al., 2002); it implements various XML mechanism and proposes a language for developing shallow grammars over XML documents; such grammars have been implemented for Bulgarian, as reported in (Simov et al., 2004; Simov and Osenova, 2004).

Finally, a new system, SPADE (*Shallow Parsing and Disambiguation Engine*), abbreviated to "♠" (Unicode character 0x2660), has recently been developed at the Institute of Computer Science, Polish Academy of Sciences (Przepiórkowski, 2007b; Buczyński, 2007). This tool, unlike many other shallow parsing tools,[13] accepts a possibly morphosyntactically ambiguous (XCES-encoded) input and performs simultaneous morphosyntactic disambiguation and shallow parsing. For example, the rule below, called `P + co/kto`, will match a possible preposition followed by a possible form of one of the pronouns CO 'what' or KTO 'who',[14] it will try to unify the selected case of the preposition with the case of the pronoun and, if that succeeds, it will mark any non-unified interpretations as rejected and it will mark the two words as a prepositional group with the preposition (cf. `1` below) as its syntactic head and the pronoun (cf. `2`) as its semantic head.[15] Moreover, any non-prepositional interpretations of the first segment of the match and any non-nominal interpretations of the second segment will be marked as incorrect. The language for specification of segments is based on the query syntax of the Poliqarp corpus search engine (Przepiórkowski et al., 2004; Janus and Przepiórkowski, 2007), in turn based on CQP (Christ, 1994).

```
RULE P + co/kto

Match: [pos~"prep"][base~"co|kto"]
Cond:  unify(case,1,2)
Synt:  group(PrepNG,1,2)
Morph: leave(pos~~"prep",1)
Morph: leave(pos~~"subst",2)
```

SPADE is currently employed for the shallow processing of the IPI PAN Corpus of Polish.

5 Conclusion

While the relatively free word order of Slavonic languages makes the processing of Slavonic unambiguously harder, I claim that the effects of the rich nominal inflection are mixed: rich inflection dramatically increases the complexity of low-level IE tasks such as NER, but it is beneficial for high-level IE tasks which involve filling scenario templates, as it facilitates identifying grammatical roles, parsing coordination, etc. Moreover, as becomes clear on the basis of the overview of practical work on Slavonic IE in the last decade, recent years have witnessed substantially increased interest and activity in the area. I am convinced that the Balto-Slavonic Natural Language Processing workshop at ACL 2007 will further catalyse the development of this field.

Acknowledgments

I am grateful to Jakub Piskorski, Bruno Pouliquen, Ralf Steinberger and Hristo Tanev for their generous invitation to the BSNLP workshop, and to the European Commission's Joint Research Centre in Ispra, Italy, for financial support. This paper was written during my stay at the University of Regensburg, Germany, within the DAAD-MNiSW cooperation programme *Automatic methods of extracting linguistic knowledge from corpora* (2006–2007). Many thanks to Qba Piskorski and Roland Meyer for comments on the (as it turned out, pre-) final versions of the paper.

References

Mirosław Bańko, editor. 2000. *Inny słownik języka polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Markus Becker, Witold Drożdżyński, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich Schäfer, and Feiyu Xu. 2002. SProUT — shallow processing with typed feature structures and unification. In *Proceedings of the International Conference on NLP (ICON 2002)*, Mumbai, India.

[13] But the shallow grammars for Serbo-Croatian described in (Nenadić and Vitas, 1998; Nenadić, 2000) were developed with similar goals in mind.

[14] Left and right context of a match may be specified; here they are empty.

[15] A rationale for distinguishing these two kinds of heads is given in (Przepiórkowski, 2007a).

Božo Bekavac and Marko Tadić. 2007. Implementation of Croatian NERC system. In Piskorski et al. (Piskorski et al., 2007a).

Leonard Bolc, Zbigniew Michalewicz, and Toyoaki Nishida, editors. 2005. *Intelligent Media Technology for Communicative Intelligence, Second International Workshop, IMTCI 2004, Warsaw, Poland, September 13-14, 2004, Revised Selected Papers*, volume 3490 of *Lecture Notes in Computer Science*. Springer-Verlag.

Chris Brew, David McKelvie, Richard Tobin, Henry Thompson, and Andrei Mikheev, 2000. *The XML Library LT XML version 1.2: User documentation and reference guide*. Language Technology Group, University of Edinburgh. `http://www.ltg.ed.ac.uk/software/xml/xmldoc/xmldoc.html`.

Aleksander Buczyński. 2007. An implementation of combined partial parser and morphosyntactic disambiguator. In *Proceedings of ACL 2007 Student Research Workshop*.

Oli Christ. 1994. A modular and flexible architecture for an integrated corpus query system. In *COMPLEX'94*, Budapest.

Michael Collins, Jan Hajič, Lance Ramshaw, and Christoph Tillmann. 1999. A statistical parser for Czech. In *Proceedings of ACL 1999*, pages 505–518, University of Maryland.

Greville G. Corbett. 1978. Numerous squishes and squishy numerals in Slavonic. In Bernard Comrie, editor, *Classification of Grammatical Categories*, pages 43–73. Linguistic Research, Inc., Edmonton.

Hamish Cunningham, Robert Gaizauskas, and Yorick Wilks. 1995. A general architecture for text engineering (GATE) — a new approach to language engineering R&D. Technical report, Department of Computer Science, University of Sheffield. `http://xxx.lanl.gov/abs/cs.CL/9601009`.

Hamish Cunningham, Diana Maynard, and Valentin Tablan. 2000. JAPE: a Java Annotation Patterns Engine (second edition). Technical Report CS–00–10, Department of Computer Science, University of Sheffield.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

Robert Dale and Paweł Mazur. 2007. Handling conjunctions in named entities. In Gelbukh (Gelbukh, 2007), pages 131–142.

Witold Drożdżyński, Petr Homola, Jakub Piskorski, and Vytautas Zinkevičius. 2003. Adapting SProUT to processing Baltic and Slavonic languages. In Hamish Cunningham, E. Paskaleva, Kalina Bontcheva, and G. Angelova, editors, *Information Extraction for Slavonic and Other Central and Eastern European Languages*, pages 18–25, Borovets, Bulgaria.

ELRA. 2004. *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*, Lisbon.

Jakub Fast and Adam Przepiórkowski. 2005. Automatic extraction of Polish verb subcategorization: An evaluation of common statistics. In Vetulani (Vetulani, 2005), pages 191–195.

Steven Franks. 1995. *Parameters of Slavic Morphosyntax*. Oxford University Press, New York.

Radovan Garabík, editor. 2005. *Computer Treatment of Slavic and East European Languages: Proceedings of the Third International Seminar, Bratislava, Slovakia, 10–12 November 2005*, Bratislava. VEDA: Vydavatel'stvo Slovenskej akadéme vied.

Radovan Garabík and Lucia Gianitsová-Ološtiaková. 2005. Manual morphological annotation of Slovak translation of Orwell's novel 1984 — methods and findings. In Garabík (Garabík, 2005), pages 59–66.

Alexander Gelbukh, editor. 2007. *Computational Linguistics and Intelligent Text Processing (CICLing 2007)*, Lecture Notes in Computer Science, Berlin. Springer-Verlag.

Włodzimierz Gruszczyński. 1989. *Fleksja Rzeczowników Pospolitych we Współczesnej Polszczyźnie Pisanej (na materiale „Słownika języka polskiego" PAN pod redakcją W. Doroszewskiego)*, volume 122 of *Prace Językoznawcze*. Ossolineum, Wrocław.

Jan Hajič and Barbara Hladká. 1997. Probabilistic and rule-based tagger of an inflective language - a comparison. In *Proceedings of the ANLP'97*, pages 111–118, Washington, DC.

K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. 2002. Slavonic named entities in GATE. Research Memorandum CS-02-01, University of Sheffield.

Nancy Ide, Patrice Bonhomme, and Laurent Romary. 2000. XCES: An XML-based standard for linguistic corpora. In *Proceedings of the Linguistic Resources and Evaluation Conference*, pages 825–830, Athens, Greece.

Roman O. Jakobson. 1958. Morfologičeskie nabljudenija nad slavjanskim skloneniem. In *Selected Writings II*, pages 154–183. Mouton, The Hague.

Daniel Janus and Adam Przepiórkowski. 2007. Poliqarp: An open source corpus indexer and search engine with syntactic extensions. In *Proceedings of ACL 2007 Demo Session*.

Janna Khegai. 2006. GF parallel resource grammars and Russian. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 475–482, Sydney, Australia. Association for Computational Linguistics.

Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors. 2005. *Intelligent Information Processing and Web Mining*. Advances in Soft Computing. Springer-Verlag, Berlin.

Svetla Koeva. 2007. Multi-word term extraction for Bulgarian. In Piskorski et al. (Piskorski et al., 2007a).

Ivana Kruijff-Korbayová and Geert-Jan M. Kruijff. 1999. Handling word order in a multilingual system for generation of instructions. In Václav Matousek, Pavel Mautner, Jana Ocelíková, and Petr Sojka, editors, *Text, Speech and Dialogue - Second International Workshop, TSD'99, Plzen, Czech Republic, September 1999*, pages 83–88, Berlin. Springer-Verlag.

Witold Mańczak. 1956. Ile jest rodzajów w polskim? *Język Polski*, XXXVI(2):116–121.

Małgorzata Marciniak and Agnieszka Mykowiecka. 2007. Automatic processing of diabetic patients' hospital documentation. In Piskorski et al. (Piskorski et al., 2007a).

Małgorzata Marciniak, Agnieszka Mykowiecka, Anna Kupść, and Jakub Piskorski. 2005. Intelligent content extraction from Polish medical texts. In Bolc et al. (Bolc et al., 2005), pages 68–78.

Jiří Mirovský, Roman Ondruška, and Daniel Průša. 2002. Searching through Prague Dependency Treebank: Conception and architecture. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT2002)*, pages 114–122, Sozopol, Bulgaria.

Agnieszka Mykowiecka, Anna Kupść, and Małgorzata Marciniak. 2005a. Rule-based medical content extraction and classification. In Kłopotek et al. (Kłopotek et al., 2005), pages 237–246.

Agnieszka Mykowiecka, Małgorzata Marciniak, and Anna Kupść. 2005b. Making shallow look deeper: Anaphora and comparisons in medical information extraction. In Vetulani (Vetulani, 2005).

Goran Nenadić. 2000. Local grammars and parsing coordination of nouns in Serbo-Croatian. In *Proceedings of Text, Dialogue and Speech (TSD) 2000*, pages 57–62. Springer-Verlag.

Goran Nenadić and Duško Vitas. 1998. Using local grammars for agreement modeling in highly inflective languages. In *Proceedings of Text, Dialogue and Speech (TSD) 1998*, pages 91–96.

Maciej Ogrodniczuk. 2006. *Weryfikacja korpusu wypowiedników polskich (z wykorzystaniem gramatyki formalnej Świdzińskiego)*. Ph. D. dissertation, Warsaw University, Warsaw.

Karel Pala, Pavel Rychlý, and Pavel Smrž. 1998. Corpus annotation in inflectional languages: Czech. In A Min Tjoa and Roland R. Wagner, editors, *Ninth International Workshop on Database and Expert Systems Applications*, pages 149–153, Los Alamitos, California.

Dariusz Piechociński and Agnieszka Mykowiecka. 2005. Question answering in Polish using shallow parsing. In Garabík (Garabík, 2005), pages 167–173.

Jakub Piskorski. 2004a. Extraction of Polish named-entities. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004* (ELR, 2004), pages 313–316.

Jakub Piskorski. 2004b. Rule-based named-entity recognition for Polish. In *Proceedings of the Workshop on Named-Entity Recognition for NLP Applications held in conjunction with the 1st International Joint Conference on NLP, March 2004*, Sanya, Hainan Island, China.

Jakub Piskorski. 2005. Named-entity recognition for Polish with SProUT. In Bolc et al. (Bolc et al., 2005).

Jakub Piskorski, Peter Homola, Małgorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiórkowski, and Marcin Woliński. 2004. Information extraction for Polish using the SProUT platform. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 227–236. Springer-Verlag, Berlin.

Jakub Piskorski, Bruno Pouliquen, Ralf Steinberger, and Hristo Tanev, editors. 2007a. *Proceedings of the BSNLP workshop at ACL 2007*, Prague.

Jakub Piskorski, Marcin Sydow, and Anna Kupść. 2007b. Lemmatization of Polish person names. In Piskorski et al. (Piskorski et al., 2007a).

Carl Pollard and Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. Chicago University Press / CSLI Publications, Chicago, IL.

Borislav Popov, Angel Kirilov, Diana Maynard, and Dimitar Manov. 2004. Creation of reusable components and language resources for Named Entity Recognition in Russian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004* (ELR, 2004), pages 309–312.

Adam Przepiórkowski. 1996. Case assignment in Polish: Towards an HPSG analysis. In Claire Grover and Enric Vallduví, editors, *Studies in HPSG*, volume 12 of *Edinburgh Working Papers in Cognitive Science*, pages 191–228. Centre for Cognitive Science, University of Edinburgh.

Adam Przepiórkowski. 2004a. *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Adam Przepiórkowski. 2004b. O wartości przypadka podmiotów liczebnikowych. *Biuletyn Polskiego Towarzystwa Językoznawczego*, LX:133–143.

Adam Przepiórkowski. 2006a. O dystrybutywnym PO i liczebnikach jedynkowych. *Polonica*, XXVI–XXVII:171–178.

Adam Przepiórkowski. 2006b. The potential of the IPI PAN Corpus. *Poznań Studies in Contemporary Linguistics*, 41:31–48.

Adam Przepiórkowski. 2007a. On heads and coordination in valence acquisition. In Gelbukh (Gelbukh, 2007), pages 50–61.

Adam Przepiórkowski. 2007b. A preliminary formalism for simultaneous rule-based tagging and partial parsing. In Georg Rehm, Andreas Witt, and Lothar Lemnitzer, editors, *Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*, pages 81–90. Gunter Narr Verlag, Tübingen.

Adam Przepiórkowski and Jakub Fast. 2005. Baseline experiments in the extraction of Polish valence frames. In Kłopotek et al. (Kłopotek et al., 2005), pages 511–520.

Adam Przepiórkowski and Marcin Woliński. 2003a. A flexemic tagset for Polish. In *Proceedings of* Morphological Processing of Slavic Languages*, EACL 2003*, pages 33–40, Budapest.

Adam Przepiórkowski and Marcin Woliński. 2003b. The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. In *Proceedings of the* 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, EACL 2003*, pages 109–116.

Adam Przepiórkowski, Zygmunt Krynicki, Łukasz Dębowski, Marcin Woliński, Daniel Janus, and Piotr Bański. 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* (ELR, 2004), pages 1235–1238.

Adam Przepiórkowski, Łukasz Degórski, Miroslav Spousta, Kiril Simov, Petya Osenova, Lothar Lemnitzer, Vladislav Kuboň, and Beata Wójtowicz. 2007. Towards the automatic extraction of definitions in Slavic. In Piskorski et al. (Piskorski et al., 2007a).

Vasile Rus, Sireesha Ravi, Mihai C. Lintean, and Philip M. McCarthy. 2007. Unsupervised method for parsing coordinated base noun phrases. In Gelbukh (Gelbukh, 2007), pages 229–240.

Zygmunt Saloni. 1976. Kategoria rodzaju we współczesnym języku polskim. In Roman Laskowski, editor, *Kategorie gramatyczne grup imiennych we współczesnym języku polskim*, volume 14 of *Prace Instytutu Języka Polskiego*, pages 43–78. Ossolineum, Wrocław.

Serge Sharoff. 2004. What is at stake: a case study of Russian expressions starting with a preposition. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 17–23, Barcelona, Spain. Association for Computational Linguistics.

Kiril Simov and Petya Osenova. 2004. A hybrid strategy for regular grammar parsing. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004* (ELR, 2004), pages 431–434.

Kiril Simov, Z. Peev, Milen Kouylekov, Alexander Simov, M. Dimitrov, and A. Kiryakov. 2001. CLaRK — an XML-based system for corpora development. In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie, and Shereen Khoja, editors, *Proceedings of the Corpus Linguistics 2001 Conference*, pages 558–560, Lancaster.

Kiril Simov, Petya Osenova, Milena Slavcheva, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krassimira Ivanova, Alexander Simov, and Milen Kouylekov. 2002. Building a linguistically interpreted corpus of Bulgarian. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002*, pages 1729–1736, Las Palmas, Canary Islands, Spain. ELRA.

Kiril Simov, Petya Osenova, Sia Kolkovska, Elisaveta Balabanova, and Dimitar Doikoff. 2004. A language resources infrastructure for Bulgarian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004* (ELR, 2004), pages 1685–1688.

Ralf Steinberger and Bruno Pouliquen. 2007. Cross-lingual Named Entity Recognition. *Linguisticae Investigationes*. Special Issue on Named Entity Recognition and Categorisation, Satoshi Sekine and Elisabete Ranchhod (eds.), Forthcoming.

Ilona Steiner. 2006. Coordinate structures: On the relationship between parsing preferences and corpus frequencies. In *Pre-Proceedings of the International Conference on Linguistic Evidence: Empirical, Theoretical and Computational Perspectives, Tübingen, 2–4 February 2006*, pages 88–92, Tübingen. SFB 441 “Linguistic Data Structures”, University of Tübingen, Germany.

Marek Świdziński. 1992. *Gramatyka formalna języka polskiego*, volume 349 of *Rozprawy Uniwersytetu Warszawskiego*. Wydawnictwa Uniwersytetu Warszawskiego, Warsaw.

Hristo Tanev. 2004. Socrates: A question answering prototype for Bulgarian. In *Recent Advances in Natural Language Processing III, Selected Papers from RANLP 2003*, pages 377–386. John Benjamins.

Hristo Tanev and Ruslan Mitkov. 2002. Shallow language processing architecture for Bulgarian. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei.

Richard Tobin, 2005. *Lxtransduce, a replacement for fsgmatch.* University of Edinburgh. `http://www.cogsci.ed.ac.uk/~richard/ltxml2/lxtransduce-manual.html`.

Dominika Urbańska and Agnieszka Mykowiecka. 2005. Multi-words Named Entity Recognition in Polish texts. In Garabík (Garabík, 2005), pages 208–215.

Zygmunt Vetulani, editor. 2005. *Proceedings of the* 2nd Language & Technology Conference, Poznań, Poland.

Marcin Woliński. 2004. *Komputerowa weryfikacja gramatyki Świdzińskiego*. Ph. D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Marcin Woliński. 2006. Morfeusz — a practical tool for the morphological analysis of Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining. Proceedings of the International IIS: IIPWM’06 Conference held in Ustron, Poland, June 19-22, 2006*, Advances in Soft Computing. Springer-Verlag, Berlin.

Daniel Zeman. 2001. How much will a RE-based preprocessor help a statistical parser? In *Proceedings of the Seventh International Workshop on Parsing Technologies (IWPT-2001), 17-19 October 2001, Beijing, China*. Tsinghua University Press.