A Search Tool for Corpora with Positional Tagsets and Ambiguities

Adam Przepiórkowski*, Zygmunt Krynicki†, Łukasz Dębowski*, Marcin Woliński*, Daniel Janus‡, Piotr Bański§

*Polish Academy of Sciences, Institute of Computer Science
ul. Ordona 21, 01-237 Warsaw, Poland
{adamp, ldebowsk, wolinski}@ipipan.waw.pl

†Polish-Japanese Institute of Information Technology
ul. Koszykowa 86, 02-008 Warsaw, Poland
zygmunt.krynicki@pjwstk.edu.pl

‡University of Warsaw, Institute of Computer Science
ul. Banacha 2, 02-097 Warsaw, Poland
nathell@bach.ipipan.waw.pl

§University of Warsaw, Institute of English
ul. Nowy Świat 4, 00-497 Warsaw, Poland
bansp@ipipan.waw.pl

Abstract

This article describes POLIQARP, a corpus indexing and query tool, which understands positional tagsets and which does not assume that word forms are annotated with unique morphosyntactic tags. POLIQARP is designed to be applicable to a variety of languages and tagsets: it works with XML-encoded texts, uses the UTF-8 character set, and allows for an external specification of the tagset. Currently, POLIQARP is used for indexing and searching a morphosyntactically annotated corpus of Polish.

1. Introduction

The aim of this article is to present POLIQARP,[1] a corpus management tool developed at the Institute of Computer Science of the Polish Academy of Sciences within a corpus project financed by the State Committee for Scientific Research (KBN; grant number 7 T11C 043 20). The aim of this project is to build a large morphosyntactically-annotated publicly-available corpus of Polish (Przepiórkowski et al., 2003), as well as to create tools for its markup, linguistic annotation, searching, and concordancing. The functionality of POLIQARP is to some extent based on that of CQP / IMS Corpus Workbench system, but — in addition — POLIQARP provides a number of novel features interesting to the broader corpus community. Of these, we concentrate on POLIQARP's ability to handle complex positional tagsets and ambiguities.

2. Motivation

For morphologically rich languages, such as Slavic languages, it makes sense to use part-of-speech (POS) tagsets more structured than those assumed for English (e.g., CLAWS 7 used in the British National Corpus), where POS and morphosyntactic (number, gender, case, etc.) information is clumped into atomic symbols. Such more structured or 'positional' tagsets are used, e.g., within the Czech National Corpus (Hajič and Hladká, 1998) and were proposed for a number of languages within the Multext-East project (Erjavec, 2001). However, currently used corpus search tools, such as CQP, are unaware of this internal structure of tags, which results in at best cumbersome access to the values of individual morphosyntactic positions. For example, in order to search for all masculine or plural nouns in Czech National Corpus, a rather cryptic query like the following must be formulated:

(1) `[tag="N.M.*"]|[tag="N..P.*"]`

A much more important deficiency of existing corpus search tools is their assumption that each corpus position is associated with a single grammatical tag, i.e., that POS annotations are fully disambiguated. It has been argued (Oliva, 2001) that it is wrong to expect a fully disambiguated annotation and that, especially in case of morphosyntactically rich languages, there are various cases of morphosyntactic ambiguities which cannot be disambiguated in a non-arbitrary way. One type of relevant examples from Polish involves both i) a verb subcategorising optionally either for a genitive or an accusative noun phrase, without any change in the meaning, and ii) a noun syncretic between the genitive and the accusative, as in (3) below, with (2) illustrating the fact that the object of the verb *pożądać* 'to covet, to desire' may occur either in the accusative or in the genitive.

[1] POLish Indexing Query and Retrieval Processor, apocryphally known as POLy-interpretation Indexing Query and Retrieval Processor.

(2) a. Pożądał ją.
desired.MASC her.ACC
'He desired her.'
b. Pożądał jej.
desired.MASC her.GEN

(3) Pożądała go.
desired.FEM him.ACC/GEN
'She desired him.'

A similar example, given below, involves predicative adjectives, which, in Polish, may either agree in case with the noun phrase they modify, or occur in the instrumental, cf. (4). However, feminine forms of such adjectives are syncretic between the accusative case and the instrumental case, so, when they predicate of an accusative noun phrase, it cannot be decided whether they bear the accusative, or the instrumental, cf. (5).

(4) a. Pamiętam go pijanego.
remember.1ST him.ACC drunk.ACC
'I remember him drunk.'
b. Pamiętam go pijanym.
remember.1ST him.ACC drunk.INS

(5) Pamiętam ją pijaną.
remember.1ST her.ACC drunk.ACC/INS
'I remember her drunk.'

Again, just as in the previous example, there is no change in meaning between the agreeing form and the instrumental form, so any attempt at 'disambiguation' would amount to an arbitrary and unjustified decision. The only sensible course of action in such cases seems to be to mark such forms as ambiguous (even after disambiguation).

Given that well-annotated corpora are bound to contain ambiguities, the user of such corpora should be allowed to search, e.g., either for nouns which are unambiguously genitive, or for nouns with the genitive case as one of the possible case values after disambiguation. Moreover, given that there are no 100% correct disambiguation methods, stochastic methods used in the current project being no exception (Dębowski, 2003; Dębowski, 2004), the user should be allowed to also search for words with certain morphosyntactic characteristics *as assigned by the morphological analyser*, regardless of any later disambiguation decisions, e.g., for all possibly genitive forms, regardless of whether they have been disambiguated as actually genitive in the given context.

3. POLIQARP — indexer and query tool

POLIQARP is a corpus management tool which makes the above tasks possible and easy. On the basis of the XML encoding of the corpus, as well as a tagset configuration file, which specifies the repertoire of POSs, morphosyntactic categories and their possible values, POLIQARP creates an internal representation of the text (a so-called index) and uses it for efficient query processing. Technical issues concerning corpus indexing and query processing are described in more detail in §4. The following subsections concentrate on the user interaction with POLIQARP and pertain to the text version of the tool; the graphical version is currently under development.

3.1. Basic query syntax

The powerful query language made available by POLIQARP is based on that of CQP[2] (Christ, 1994). In particular, regular expressions may be formulated over corpus positions, as in (6), where any non-empty sequence of adjectives is sought, or within values of attributes, as in (7), concerning forms tagged with POSs whose names start with an `a`, e.g., `adj` and `adv`:

(6) `[pos="adj"]+`
(7) `[pos="a.*"]`

When values of attributes do not contain special regular expression characters (such as `*`, `+` or `|`), quotes can be omitted, as in, e.g., (8) or (10)–(11) below.

Moreover, again as in CQP, conditions on corpus positions may be combined, e.g., for finding the sequence of at least two 5-letter nominal or adjectival forms beginning with an *a*:

(8) `[(pos=subst|pos=adj)&orth="a...."]{2,}`
(9) `[pos="(subst|adj)"&orth="a.{4}"]{2,}`

3.2. Positional tagset

POLIQARP understands positional tagsets. Each tag is a list whose first element is a POS and other elements are values of morphosyntactic categories specific for this POS. The repertoire of POSs (e.g., nouns), their morphosyntactic categories (e.g., case and gender), and possible values of those categories (e.g., for case, nominative, accusative, genitive, etc.) are defined in an external configuration file, which is consulted by POLIQARP when the corpus is indexed. In particular, it is possible to define no morphosyntactic categories, in which case the tagset reduces to an atomic tagset (such as CLAWS 7 for English).

Parts of speech and morphosyntactic categories may be queried separately, e.g., query (10) could be posed when searching for various masculine forms, regardless of the POS or other categories, while query (11) can be used to find nominal forms which are masculine or plural.

(10) `[gend=masc]`
(11) `[pos=subst&(gend=masc|numb=pl)]`

It is also possible to specify the POS and various morphosyntactic categories more compactly (and obtusely), using the `tag` attribute. For example, given the positional tagset for Polish proposed in (Przepiórkowski and Woliński, 2003a; Przepiórkowski and Woliński, 2003b), the following two queries, finding sequences of at least two nouns in the nominative or the accusative, are equivalent:

(12) `[pos=subst&case="nom|acc"]{2,}`
(13) `[tag="subst:.*:[na].*:.*"]{2,}`

For the purpose of such queries, full tags are supposed to have the form `pos:cat`$_1$`:cat`$_2$`:...:cat`$_n$, where `cat`$_i$s are morphosyntactic categories relevant for `pos` (e.g., case, gender and number, but not aspect, for nouns). The order of these categories is specified in the tagset configuration file.

[2] `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`.

3.3. Ambiguities

Assuming a corpus with some ambiguities possibly left after POS disambiguation, a query like `[case=acc]` will find all forms which, after disambiguation, are considered to be possibly accusative, e.g., it will find a syncretic accusative/genitive complement of a verb taking accusative or genitive complement; cf. *go* 'him' in (3). In order to find forms which are disambiguated to the accusative only, a different equality sign should be used: `[case==acc]`; this query will not find the form *go* in (3). Any corpus position matched by `[X==Y]` will also be matched by `[X=Y]`; in short, `[X==Y]` → `[X=Y]`. Note that queries involving '==' cannot be used to find uniquely disambiguated forms, but rather forms for which all tags after disambiguation have the accusative as the case value (but may differ in, say, gender or even POS).

Similarly, in order to find forms which may be in the accusative case according to the morphological analyser, i.e., before disambiguation, the following query should be posed: `[case~acc]`. Finally, the query `[case~~acc]` finds those forms whose all tags assigned by the morphological analyser involve the accusative case. Note that such queries can be useful for finding syncretisms; e.g., the query below can be used to find all accusative/genitive syncretic forms in the corpus:

(14) `[case~acc & case~gen]`

In summary, the following implications hold:

(15) a. `[X=Y]` → `[X~Y]`,
b. `[X==Y]` → `[X=Y]`,
c. hence, also: `[X==Y]` → `[X~Y]`,
d. `[X~~Y]` → `[X==Y]`,
e. hence, also: `[X~~Y]` → `[X=Y]`,
f. and also: `[X~~Y]` → `[X~Y]`.

4. Technical issues

Technically, POLIQARP is implemented in C and it is composed of two separate parts: an indexer that builds an efficient corpus representation, cf. §4.1. and §4.2., and a shell-like tool for end-user interaction, as well as for batch processing, cf. §4.3. and §4.4.

4.1. XML input

Each text is represented in the corpus by three XML files, (roughly) compliant with the Corpus Encoding Standard (X)CES[3] (Ide et al., 2000), cf. also (Bański, 2001; Bański, 2003): `text.xml`, satisfying (a slightly modified version of) xcesDoc.dtd, containing the text and some structural markup; `header.xml`, logically part of `text.xml`, which contains meta-data; and `morph.xml`, satisfying (a modified version of) xcesAna.dtd and containing the morphosyntactic annotation.

More precisely, each `morph.xml` contains elements representing paragraphs and sentences, and — within them — `<tok>` elements representing tokens (segments, corpus positions), all their morphosyntactic interpretations and the information about which of these interpretations are valid in the given context. In the example below, the ambiguous verbal/nominal form *myślą* is disambiguated to its verbal (finite plural 3rd person imperfective) interpretation.

```
<tok>
 <orth>myślą</orth>
 <lex disamb="1">
  <base>myśleć</base>
  <ctag>fin:pl:ter:imperf</ctag>
 </lex>
 <lex>
  <base>myśl</base>
  <ctag>subst:sg:inst:f</ctag>
 </lex>
</tok>
```

The indexer, described in more detail in the following subsection, reads a prespecified set of `morph.xml` files and creates a search-efficient representation of the information contained there.

4.2. Indexing

Source XML documents are transformed, by means of the Expat library, into several dictionaries of forms and tags, and an internal representation of the corpus which contains indexes to these dictionaries. Since high efficiency is one of the main design factors of POLIQARP, the process of parsing source XML documents is streamlined, so that only one pass is required.

Within dictionaries, all items are sorted by frequency so that most used items remain close to each other; this helps conserve memory. Also for efficiency reasons, dictionaries contain indexes that specify certain alphabetic orderings which are used for sorting the results: this includes both the usual *a fronte* alphabetic order, and the *a tergo* (or reversed-word alphabetic) ordering. Since dictionaries contain variable-length items (words and tags of different lengths), offsets are used to enable constant-time dictionary lookup.

The internal representation of the corpus consists of constant-length (16 byte) records, which enables constant-time access to any corpus position. Each such record contains, *inter alia*, an index to the orthographic word (token) occupying the given position, an index to the sequence of morphosyntactic interpretations assigned to that word, and an index to the disambiguation information for that word and that sequence. Note that, again in order to conserve memory, the dictionary of sequences of interpretations that may be assigned to forms, e.g., the sequence of adjectival plural dative interpretations differing in gender values (in Polish dative plural adjectival forms are fully syncretic for gender), is separated from the information on which of these interpretations are correct in the given context. Assuming the repertoire of 9 genders in Polish proposed by (Saloni, 1976), the conflation of the morphosyntactic information with the disambiguation information would result in $2^9 = 512$ theoretically possible sequences of adjectival plural dative interpretations in context. Finally, for each morphosyntactic interpretation, there is an index to the lemma of the orthographic word. This slightly complex indexing

[3] `http://www.cs.vassar.edu/XCES/`.

architecture allows for efficient matching of queries to corpus positions.

4.3. Query processing

Generally, queries have a 3-tiered structure. The most basic expression is a match expression using one of the four match operators: *match any interpretation*, '~', *match all interpretations*, '~~', *match any disambiguated interpretation*, '=', and *match all disambiguated interpretations*, '=='. The left hand side of such an operator is an attribute, e.g., `pos` or a category name (as defined in the tagset configuration file), and the right hand side is a regular expression defining the value of this attribute. At the second tier, match expressions are combined, with the use of the three logical operators of *conjunction* '`&`', *disjunction* '`|`' and *negation* '`!`' into position expressions, which define conditions on a single corpus position. Finally, those position expressions are treated as letters of the alphabet of regular expressions which constitute the full queries.

The syntax of queries is defined externally and it is converted (with the use of the standard tools *bison* and *flex*) into a query parser. This parser is then used for parsing queries into attribute expressions (represented as simple evaluation trees) and regular expressions (represented as nondeterministic finite state automata with ϵ-symbols, subsequently converted into deterministic finite state automata, optimised for efficiency).

4.4. Concordancing

Queries may contain a special focus marker, '^', which defines the alignment position. For example, for the query '`[pos=adj]+^[pos=subst]+`', the results will be aligned with respect to the first substantive word (i.e., the first noun). More precisely, each match is split into four parts: left-context (a few positions before the first matching position), left-match (match up to the focus point), right-match, and right-context. Such list of results may be sorted using any combination of ascending or descending *a fronte* or *a tergo* orders, as applied to each of these four parts.

POLIQARP features a special LISP-like language that describes how the results should be displayed. The formatting specifications are compiled into a program executable on a simple stack-based virtual machine specialised for list and text processing. In particular, it is possible to define the HTML format of the results and to render it with a WWW browser such as Mozilla.

5. Conclusion

The main emphasis in corpus linguistics has long shifted from morphosyntactic annotation to higher levels of linguistics representation (mainly syntax and, more recently, semantics). However, the vast majority of work on morphosyntactically annotated corpora pertains to English and other Germanic languages, whose morphology, in comparison with Slavic, is severely impoverished. As a result, various standards, best practices and tools are not optimal for languages with richer morphosyntactic structure.

Our earlier work towards filling this gap, reported in (Przepiórkowski and Woliński, 2003a; Przepiórkowski and Woliński, 2003b), is concerned with the principles of designing a tagset for morphologically rich languages, with a detailed application of these principles to Polish. The present article, describing a corpus management tool which understands such positional tagsets and makes sense of morphosyntactic ambiguities, omnipresent in morphologically rich languages, marks another step in the same direction.

6. References

Bański, Piotr, 2001. The proposed annotation scheme for the IPI PAN corpus. IPI PAN Research Report 936, Institute of Computer Science, Polish Academy of Sciences.

Bański, Piotr, 2003. Anotacja zewnętrzna: wpływ architektury korpusu IPI PAN na efektywność jego tworzenia oraz wykorzystania [Eng.: Stand-off annotation: the impact of the architecture of the IPI PAN corpus on the effectiveness of its creation and use]. *Polonica*, XXII–XXIII:77–91.

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

Dębowski, Łukasz, 2003. A reconfigurable stochastic tagger for languages with complex tag structure. In *Proceedings of* Morphological Processing of Slavic Languages*, EACL 2003*.

Dębowski, Łukasz, 2004. Trigram morphosyntactic tagger for Polish. In *Proceedings of IIS:IIPWM 2004*.

Erjavec, Tomaž (ed.), 2001. *Specifications and Notation for MULTEXT-East Lexicon Encoding*. Ljubljana.

Hajič, Jan and Barbara Hladká, 1998. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proceedings of COLING-ACL'98*. Montréal.

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

Oliva, Karel, 2001. On retaining ambiguity in disambiguated corpora. *TAL (Traitement Automatique des Langues)*, 42(2).

Przepiórkowski, Adam, Piotr Bański, Łukasz Dębowski, Elżbieta Hajnicz, and Marcin Woliński, 2003. Konstrukcja korpusu IPI PAN [Eng.: The construction of the IPI PAN corpus]. *Polonica*, XXII–XXIII:33–38.

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

Przepiórkowski, Adam 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*.

Saloni, Zygmunt, 1976. Kategoria rodzaju we współczesnym języku polskim [Eng.: The category of gender in contemporary Polish]. In *Kategorie gramatyczne grup imiennych we współczesnym języku polskim [Eng.: The grammatical categories of nominal groups in contemporary Polish]*. Wrocław: Ossolineum, pages 41–75.