CMST 24(1) 21–27 (2018) DOI:10.12921/cmst.2018.0000005

Korpusomat – a Tool for Creating Searchable Morphosyntactically Tagged Corpora

Witold Kieraś, Łukasz Kobyliński, Maciej Ogrodniczuk

Institute of Computer Science, Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warszawa

E-mail: w.kieras@ipipan.waw.pl, l.kobylinski@ipipan.waw.pl, m.ogrodniczuk@ipipan.waw.pl

Received: 21 March 2017; revised: 30 November 2017; accepted: 16 January 2018; published online: 31 March 2018

Abstract: The paper presents Korpusomat, a web application aimed at building annotated corpora for the purpose of corpus linguistic studies. Korpusomat combines existing tools, such as morphological analyser, tagger and corpus search engine, and provides an easy-to-use environment for building corpora technically compatible with the National Corpus of Polish from almost any text, including texts in binary formats. In the paper we present the current state of the project, its features and functionalities, as well as some future plans and developments tasks. A usage example is also presented.

Key words: corpus linguistics, corpus management, language corpora

I. INTRODUCTION

Korpusomat[1] is a simple web application enabling researchers to create morphosyntactically annotated text corpora without much technical knowledge about the underlying computational linguistic components. The resulting corpora can be then queried using standard search tools such as Poliqarp, the corpus search engine used by the National Corpus of Polish (Pol. Narodowy Korpus Języka Polskiego, NKJP) [1]. The application builds on existing resources and tools created by the Linguistic Engineering Group, Polish Academy of Sciences, and can be used in various types of studies in digital humanities.

Development of the application was motivated by discrepancy between availability of electronic data ready for use in many research domains and usability of tools capable of performing its linguistic analysis. One of our authors was frequently contacted by researchers who have successfully gathered large amounts of texts representing their narrow domain (and as such unavailable in balanced corpora of Polish) but were unable to install and configure tools for further processing and querying this data in order to perform some basic quantitative analyses. Korpusomat intends to fill this gap by offering configuration-free environment facilitating such tasks. While research on general Polish can be effectively completed by accessing the services around the National Corpus of Polish, Korpusomat is intended to be used for analysing domain data, specialist texts or thematic collections.

II. PREVIOUS WORK

Historically, the work concerning tools in the area of English-language corpus linguistics concentrated either around query engines developed for specific corpora (e.g. the user interface to the British National Corpus, the BYU-BNC[2]), or focused on simple concordancers, which allowed to perform text queries on a set of documents (e.g. AntConc concordancer[3]).

The most notable exception is the Sketch Engine [2], a general corpus manager which allows for creating and analyzing text documents provided by the user. The Sketch Engine is a feature-rich and multi-language system, but it is provided with a commercial license and thus is not easily accessible for individual linguists, or researchers from smaller institutions.

[1] http://www.korpusomat.pl
[2] http://corpus.byu.edu/bnc/
[3] http://www.laurenceanthony.net/software/antconc/

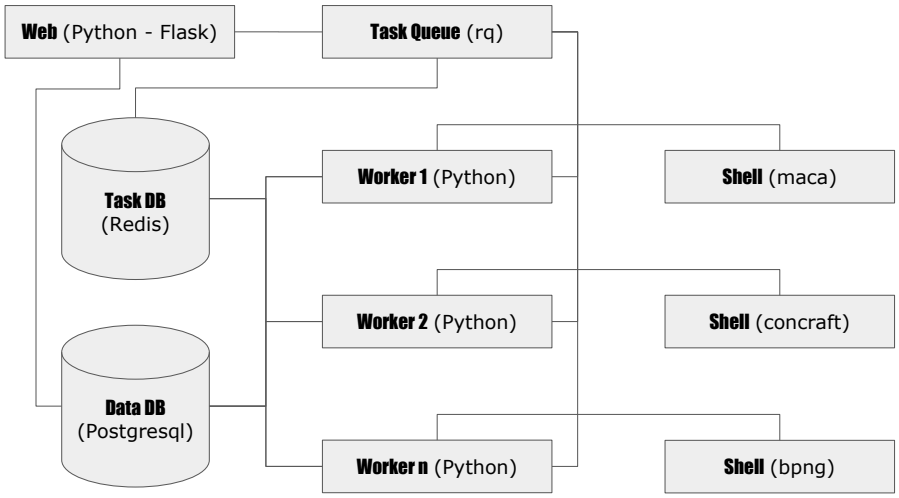

Fig. 1. Simplified system architecture

In the case of Polish language the situation is similar: considerable work has been done in the area of developing corpus query engines, as well as automated text processing, such as morphosyntactic tagging but these tools are difficult to use for non-technical users and up to date have been provided with more approachable user interfaces only in combination with existing corpora (e.g. the National Corpus of Polish[4]).

For this reason, we have decided to propose Korpusomat, a web-based service which combines several components to provide an easy-to-use user interface for creating searchable text corpora in Polish. Korpusomat comprises:

- morphological analyser Morfeusz 2 [3, 4] based on the Grammatical Dictionary of Polish (Polish: Słownik gramatyczny języka polskiego, SGJP) [5]
- disambiguating tagger Concraft-pl [6]
- corpus search tool Poliqarp [7, 8]
- various third-party extraction and conversion components
- web application interface and backend, allowing for task-based processing of user requests.

III. SYSTEM ARCHITECTURE AND COMPONENTS

Korpusomat is a web-based application intended to be used by multiple users simultaneously. As such, besides including such typical components as a user authentication mechanism, its server-side logic is based on a task queue paradigm, which allows to process user requests in the order in which they are created and dynamically limit the server resources used for that purpose (see Fig. 1).

The processing of the input data provided by the user starts with a file format conversion procedure. As the user is free to provide both text and binary files, a heuristic procedure is used to check if binary conversion is needed. In such a case a conversion module from the Calibre suite of e-book processing tools is used to convert files in EPUB and MOBI formats. In the case of all other binary formats, the Apache Tika library is used. Korpusomat uses the metadata extraction capabilities of these tools to include this information in further steps of processing and make the metadata available in the final corpus. The resulting text file is finally converted to the universal UTF-8 character encoding and is ready for further processing at this point.

Automated morphosyntactic tagging of the provided text is a two-step process. First, morphological analysis is performed, which results in context-independent assignment of (potentially many) morphosyntactic interpretations[5] to segments – sequences of orthographic characters no longer than orthographic words ("from space to space"), representing the internal structure of Polish words[6] Morfeusz 2, the most frequently used morphological analyser for Polish is used in this part of the process. Secondly, these interpretations are disambiguated in a context-aware manner using statistical methods with Concraft-pl tagger, one of the best performing and actively developed taggers for Polish (see e.g. [11] for tagger performance comparison).

For the purpose of making direct comparisons with NKJP possible, Korpusomat offers also a possibility to use an older version of Morfeusz, which was used to annotate the NKJP corpus, but is now obsolete and no longer updated. However it is highly recommended to choose Morfeusz 2 which contains

[4] `http://www.nkjp.pl`

[5] See e.g. [9] for details of the tagset used.

[6] [10] defines segments as "those sequences of orthographic characters to which tags are assigned." See chapter 2.1 of this article for an extensive list of examples.

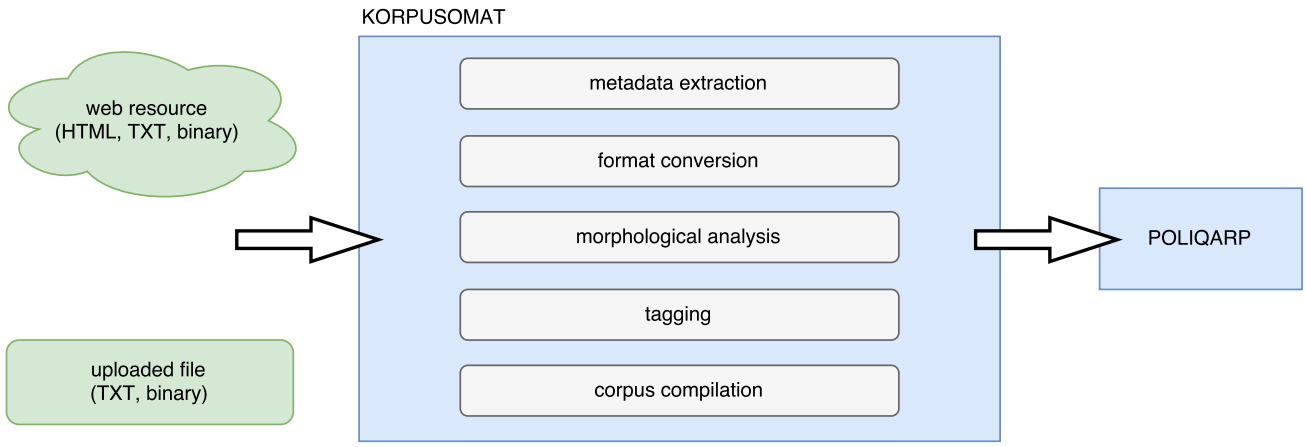

Fig. 2. System overview

an up-to-date inflectional dictionary of SGJP.

Maca [12], a supporting library providing mechanisms for interfacing the application code with Morfeusz and Morfeusz 2, as well as for corpus format conversion, is used alongside these tools.

The final step of text processing is compiling the corpus – converting it to a binary format, which is suitable for efficient querying and browsing. This step is performed by the Poliqarp corpus compiler and the resulting set of binary files is compressed into a single zip archive and provided for the user to download.

Morfeusz and Concraft-pl are actively developed and their new versions will be consequently adopted in our application. On the contrary, Poliqarp is currently no longer updated but, at the same time it is still under use in several projects [13-15].

IV. USAGE

The general overview of the system, summarizing the processing stages described in the previous section is presented in Fig. 2.

The input data can be provided by the user both in the form of online resources (web pages, as well as text and binary documents) and in the form of local files uploaded to the server. Processing such data in Korpusomat results in a compiled corpus provided in a binary format readable by the Poliqarp corpus query engine.

Corpora are created in private user space thus registration in the service is necessary. After logging in the user can input a name for their new corpus, upload a set of texts, and, after they are processed (in background), trigger compilation of the corpus. When it is completed, its binary version can be downloaded to be queried offline on the user's computer.

IV. 1. Creating a corpus

Korpusomat can convert various document formats and character encodings, starting with plain text files, through popular word processing and e-book formats, up to searchable PDFs.[7] Both locally accessible texts as well as URL-based texts can be added[8] multiple texts at a time. After the text has been retrieved, user can fill in its basic metadata: author, title, publication date and genre. For EPUB and HTML files this information is retrieved from the file and can be later corrected and updated. Custom user metadata fields can also be defined. Metadata labels are useful not only in managing corpus content but can also be used directly in corpus queries to restrict the search domain.

Morphological analysis and tagging are the most time-consuming stages of text processing; for an average book consisting of 80–100k words the processing time should amount to approx. 2–3 minutes but it obviously depends on the current server load. The processing is time-restricted; presently the timeout is set to 10 minutes. When processing of all the texts included in the corpus is completed, the final binary version may be created. The resulting file can be downloaded to a local computer in the form of a zip archive and opened in Poliqarp search engine. Texts can be added and removed from existing corpus, but the changes are reflected in the binary version only after a new compilation is triggered.

IV. 2. Querying a compiled corpus

The compiled corpus can be queried using a standalone Poliqarp desktop application,[9] which should be installed on the user's computer. After unzipping the binary corpus file to a folder of choice, Poliqarp can be started and the corpus can be loaded from the folder. The corpus can be then searched using the query language familiar to users of the National

[7] The list of formats depends on the external converter; see `http://tika.apache.org/1.13/formats.html`.
[8] Uploaded texts will not be used in any way other than automatic processing for the purpose of corpus compilation.
[9] Current version: 1.3.13, see `http://clip.ipipan.waw.pl/Poliqarp`. Poliqarp installation requires Java environment.

Fig. 3. Sienkiewicz's *Trilogy* corpus view in Korpusomat

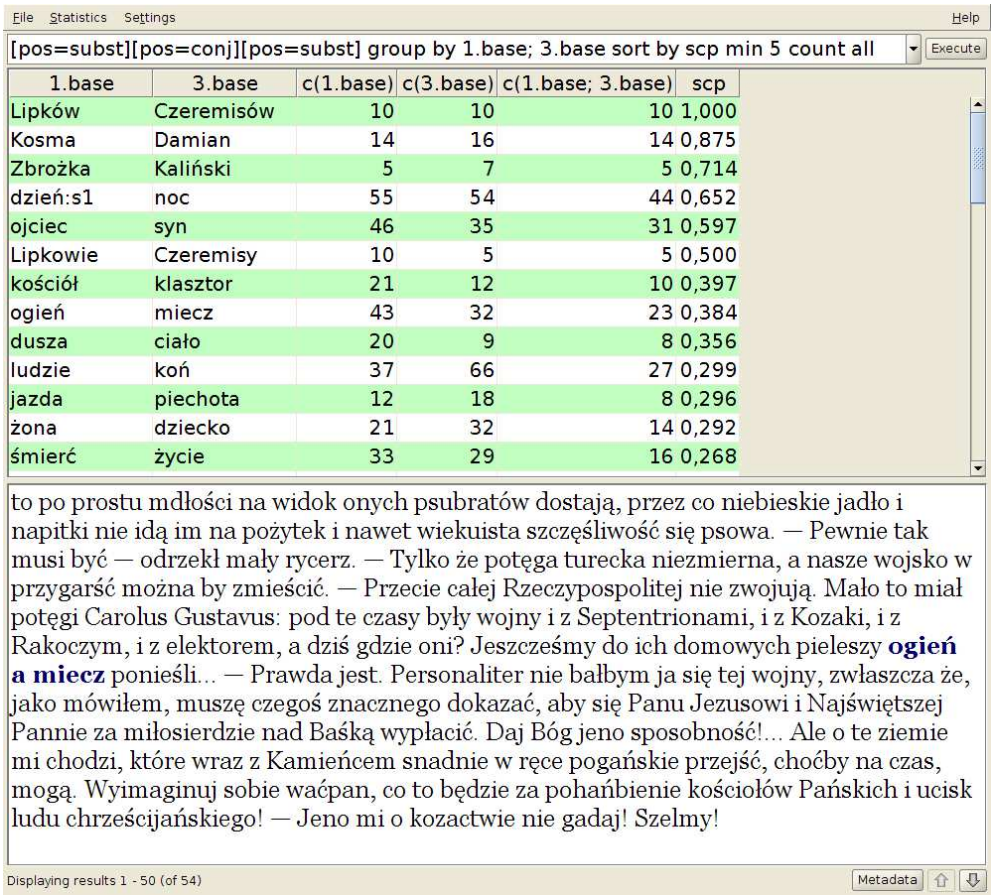

Fig. 4. Poliqarp search engine executing a statistical query

Corpus of Polish. Apart from simple content-based queries, also the metadata and morphosyntactic descriptions can be used, e.g. to find all feminine nouns.[10] Moreover, statistical queries are available, e.g. to create the frequency list of all nouns in the corpus.[11]

IV. 3. Usage examples

Let us consider a very short usage example of the service. A wide variety of public domain literary texts are available for free in services such as *Wolnelektury.pl* and *Wikisource*. We have chosen Henryk Sienkiewicz's trilogy novels from the former as our working example. After uploading text files to Korpusomat, the corpus is ready to download and open in Poliqarp within minutes. Let us assume that we want to check how many occurrences of *ogniem i mieczem* ('with fire and sword'), a phrase which is the title of the first novel, appear in the whole trilogy. For a query performed by simply typing the phrase, the search engine returns 15 hits. By narrowing the query down to one of the three novels (e.g. `ogniem i mieczem meta Tytuł=Potop`) it is easy to notice that the phrase appears mostly in *The Deluge* (12 hits), that is the second part of the trilogy.

Subsequently we can check for occurrences of the more generalized version of the phrase, namely all occurrences of nouns *ogień* 'fire' and *miecz* 'sword' conjoined with any conjunction, by performing the following query:

```
[base=ogień][pos=conj][base=miecz]
```

It returns 23 hits, most of them in the instrumental case and with *i* conjunction (as in the novel title). With Poliqarp, it is possible to check how frequent such conjunction of nouns is in a given corpus:

```
[base=subst][base=conj][base=subst]
group by 1.base;3.base
sort by freq count all
```

By executing this query the user obtains a frequency list of phrases consisting of a conjunction placed between two nouns. It can be noticed that coordination of the nouns *ogień* and *miecz* is among the most frequent in the trilogy.

It is also possible to compute some basic probabilities of word sequences. Fig. 4 shows an example of such a query: a symmetrical probability of two nouns appearing in the context of conjunction between them narrowed down to those with at least five occurrences in the corpus. As can be observed, the sequence is once again among the most probable ones. On the other hand, it is also worth noting that no automatic annotation is perfect. Proper names unrecognized by the morphological analyser were not lemmatized correctly and thus one of the word sequences appears on the list twice (once as *Lipków* and *Czeremisów* and once as *Lipkowie* and *Czeremisy*).

V. PERSPECTIVES

The basic goal for Korpusomat is to keep all of its components up to date, particularly the morphological analyser and tagger which are in constant development.

One of the main functional enhancements that we are planning to include in the service in the near future is the integration of the system with a web-based search engine client. Although the process of installing the desktop Poliqarp search engine is relatively easy, it limits the flexibility of Korpusomat. Integrating a web-based search engine directly into Korpusomat will allow the user to access his/her personal corpora collection from virtually any computer with any web browser without the need to install any additional software locally.

Another major enhancement planned is the development of a graphical corpus queries creator, which would simplify the process of querying the corpus by inexperienced users. The query language of Poliqarp, although very expressive, is considered very difficult to master by many current users. Thus this simplification would be valuable and further improve the accessibility of the system.

The process of compiling a corpus is currently straightforward and involves almost no customization. In the future, we are planning to make it possible for the user to choose between at least two different morphological dictionaries by including the Polimorf dictionary alongside the SGJP. It is possible that some other variants of morphological dictionaries will be provided by Korpusomat, e.g. these suited to use with historical texts or non-standard variants of Polish (slang, dialects, etc). This however depends on actual outcomes of third party research projects concerning development of morphological resources of that kind.

Annotated corpora produced by Korpusomat are now available in a specific binary format of Poliqarp as well as XML source files. Other formats are possible in the future, but it depends heavily on the users' request. Other minor enhancements of the service also depend on the users' input which is most welcome.

Acknowledgement

Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education

References

[1] A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk, eds, *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*, Wydawnictwo Naukowe PWN, Warsaw, 2012.

[10] `[pos=subst & gender=f]`

[11] See `[pos=subst] group by base sort by freq count all`.

[2] A. Kilgarriff et al., *The Sketch Engine: ten years on*, Lexicography, pages 1–30, (2014).

[3] M. Woliński, *Morfeusz Reloaded*, [In:] N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, eds, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1106–1111, Reykjavík, Iceland, 2014 European Language Resources Association.

[4] W. Kieraś, *Co jest zgodne z duchem kraftu? Próba korpusowego badania słownictwa związanego z piwem*, Język Polski, (2017), in print.

[5] M. Woliński, W. Kieraś, *The On-Line Version of Grammatical Dictionary of Polish*, [In:]N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, editors, *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2589–2594, Portorož, Slovenia, 2016 European Language Resources Association.

[6] J. Waszczuk, *Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language*, [In:] *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 2789–2804, Mumbai, India, 2012.

[7] D. Janus, A. Przepiórkowski, POLIQARP 1.0: Some technical aspects of a linguistic search engine for large corpora, [In:] J. Waliński, K. Kredens, S. Goźdź-Roszkowski, editors, *Proceedings of Practical Applications in Language and Computers Conference (PALC 2005)*, Frankfurt am Main, 2006 Peter Lang.

[8] A. Przepiórkowski, Z. Krynicki, Ł. Dębowski, M. Woliński, D. Janus, Piotr Bański, *A Search Tool for Corpora with Positional Tagsets and Ambiguities*, [In:] *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1235–1238, 2004.

[9] A. Przepiórkowski, M. Woliński, *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 2003)*, pages 109–116, 2003.

[10] A. Przepiórkowski, *The IPI PAN Corpus in Numbers*, [In:] Z. Vetulani, editor, *Proceedings of the 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2005)*, pages 27–31, Poznań, Poland, 2005.

[11] Ł. Kobyliński, *PoliTa: A multitagger for Polish*, [In:] N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, eds, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2949–2954, Reykjavík, Iceland European Language Resources Association.

[12] A. Radziszewski, T. Śniatowski, *Maca – a configurable tool to integrate Polish morphological data*, [In:] Felipe Sánchez-Martínez and Juan Antonio Pérez-Ortiz, editors, *Proceedings of the 2nd International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 29–36, 2011.

[13] J.S. Bień, *Efficient Search in Hidden Text of Large DjVu Documents*, [In:] R. Bernardi, S. Chambers, B. Gottfried, F. Segond, I. Zaihrayeu, eds, *Advanced Language Technologies for Digital Libraries (NLP4DL 2009)*, volume 6699 of *Lecture Notes in Computer Science*, pages 1–14 Springer, 2009.

[14] M. Łaziński, Korpusy w programach badawczych i dydaktyce Instytutu Języka Polskiego Uniwersytetu Warszawskiego, In I. Bundza, J. Kowalewski, A. Kravčuk, and O. Slivinskij, editors, *Język polski i polonistyka w Europie wschodniej: przeszłość i współczesność: praca zbiorowa z okazji dziesięciolecia Katedry Filologii Polskiej Narodowego Uniwersytetu Lwowskiego im. Iwana Franki*, pages 584–585 Fìrma ÌNKOS, Kijów, 2015.

[15] M. Ogrodniczuk, *The Polish Sejm Corpus*, [In:] N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, eds, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2219–2223, Istanbul, Turkey, 2012 European Language Resources Association.

Witold Kieraś graduated from Inter-Faculty Individual Studies in Humanities at the University of Warsaw and received his PhD in Philosophy in 2012. Interested in computational and corpus linguistics with special regard to morphology. Currently works at the Institute of Computer Science, Polish Academy of Sciences in projects concerning corpora of historical Polish. He is also interested in bringing new technologies to traditional branches of linguistics and humanities in general.

Łukasz Kobyliński received his PhD in Computer Science from the Warsaw University of Technology in 2012, specializing in data mining and pattern-based machine learning methods. Currently, he is an Assistant Professor at the Institute of Computer Science, Polish Academy of Sciences, where he focuses on applying machine learning methods to problems in the domain of natural language processing. He is also interested in expanding the availability of NLP tools to a broader audience of users, to narrow the gap between academia and business.

Maciej Ogrodniczuk graduated from the Faculty of Mathematics, Informatics and Mechanics at the University of Warsaw and received the PhD in linguistics in 2006. Currently he is leading the Linguistic Engineering Group at the Institute of Computer Science, Polish Academy of Sciences and is involved in numerous national and international projects related to computational linguistics. His research interests include semantic description of Polish, in particular automated analysis of reference relations.