COGNITIVE STUDIES | ÉTUDES COGNITIVES, 14: 75–84
SOW Publishing House, Warsaw 2014

DOI: 10.11649/cs.2014.008

JANUSZ S. BIEŃ
Formal Linguistics Department, University of Warsaw, Poland
jsbien@uw.edu.pl

THE IMPACT PROJECT POLISH GROUND-TRUTH TEXTS AS A DjVu CORPUS[1]

Abstract

The purpose of the paper is twofold. First, to describe the already implemented idea of DjVu corpora, i.e. corpora which consist of both scanned images and a transcription of the texts with the words associated with their occurrences in the scans. Secondly, to present a case study of a corpus consisting of almost 5 000 pages of Polish historical texts dating from 1570 to 1756 (it is practically the very first corpus of historical Polish). The tools described have universal character and are freely available under the GNU GPL license, hence they can be used also for other purposes.

Keywords: Polish language, corpora, DjVu, OCR, PAGE, Page Analysis and Ground-Truth Elements, GNU GPL.

1 Introduction

The IMPACT project lasted from January 2008 to June 2012, it brought together twenty-six national and regional libraries, research institutions and commercial suppliers; it has been superseded by a Center of Competence (http://www.digitisation.eu/). In the project framework, sample texts were prepared for historical texts in the national languages of the project partners, following the common rules formulated by the project leaders. The texts, intended to train and evaluate Optical Character Recognition programs and called therefore Ground-Truth texts, were annotated using the PAGE (Page Analysis and Ground-truth Elements) XML format, cf. (Pletschacher and Antonacopoulos, 2010). Preparation of the Polish texts was outsourced to DIGI-TEXX (http://www.digi-texx.com.vn/) and supervised by the Poznań Supercomputing and Networking Center Digital

[1] The Ground-Truth data has been prepared in the framework of the FP7-ICT IMPACT project (*Improving Access to Texts*, http://www.impact-project.eu) under the supervision of Poznań Supercomputing and Networking Center. The DjVu corpus tools were developed within the Ministry of Science and Higher Education's grant (no. N N519 384036, http://bc.klf.uw.edu.pl/297/).

Libraries Team. The results were made publicly available in January 2012 under the terms of the liberal Creative Commons Attribution License, cf. `http://dl.psnc.pl/activities/projekty/impact/results/`.

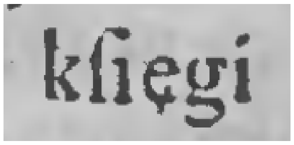
From May 2009 to May 2012, the Formal Linguistics Department of the University of Warsaw, which participated in the IMPACT project, was also working on its own project called *Digitalization tools for philological research*, supported by a grant from the Polish government; the project results are available now at `https://bitbucket.org/jsbien/ndt`. The main idea was to take full advantage of the DjVu format (Le Cun et al., 1998), in particular by extending appropriately the Poliqarp corpus software (Przepiórkowski et al., 2004), developed in the Institute of Computer Science of the Polish Academy of Sciences and now used to support the National Corpus of Polish (`http://nkjp.pl`). Poliqarp stands for *POLyinterpretation Indexing Query and Retrieval Processor*, and the extension is named simply Poliqarp for DjVu. It was designed by the present author and implemented by Jakub Wilk. The reasons for selecting the DjVu format were presented on several occasions, such as (Bień, 2009) and (Bień, 2011). In this paper we focus on problems specific to the IMPACT data.

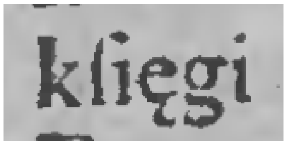
The IMPACT corpus is available at `http://corpora.klf.uw.edu.pl`.

2 Text encoding

2.1 Transcriptions used

The primary transcription used in the corpus, which may be called *strict diplomatic* or just *facsimile* transcription, preserves the distinctions which are irrelevant for many users. For example, the two occurrences presented below of the same word (spelled now *księgi*) are encoded differently:





The texts are encoded according to the Unicode standard, which in principle encodes abstract characters without taking into account their actual shapes called glyphs. The distinction between characters and glyphs is not always clear. Some glyphs are already encoded in Unicode for compatibility with other important coded character sets, and there is a possibility to use the so called Private Use Area (PUA for short) for local extensions.

Not all distinctions which we wanted to preserve could be encoded directly in Unicode, so the use of PUA was necessary. Sometimes however we used a similar, already existing character. In the above example the variants of the letter *e* had been encoded respectively as `LATIN SMALL LETTER E WITH STROKE`, which is just an approximation of the original shape, and `LATIN SMALL LETTER E WITH OGONEK` (the character, with practically the same shape, is used in comtemporary Polish). However to encode the ligature `LATIN SMALL LIGATURE LONG S I` we had to resort to the use of PUA, following the recommendations of Medieval Unicode Font Initiative (`http://www.mufi.info/`).

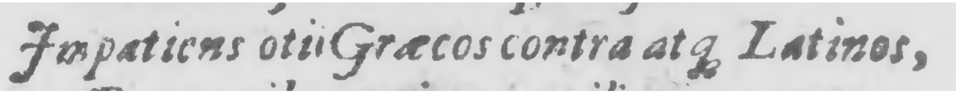
If needed, the distinction between `LATIN SMALL LETTER E WITH STROKE` and `LATIN SMALL LETTER E WITH OGONEK` can be ignored in a query by using an appropriate regular expression. As it is much more cumbersome to account in an analogical way for the use of ligatures, we decided to introduce an auxiliary transcription called tentatively *textel* transcription (by *textel* we understand a text element, but its precise definition has not been yet formulated).

2.2 Improving encoding of the source data

The changes made to the original IMPACT data are described at `http://bc.klf.uw.edu.pl/288/`. Here we only summarize them briefly.

A histogram of the characters contained in the PAGE files was created (using a tool by Piotr Findeisen available at `https://bitbucket.org/jsbien/unihistext`) and all suspicious occurrences were looked up.

Some of the questionable characters were just simple mistakes, like the unjustified occurrence of `RIGHT-TO-LEFT MARK`, but some of them were more or less plausible interpretations of the glyphs in question which we found inappropriate in the context. For example, 11 occurrences of the character `BULLET` were replaced by the `MIDDLE DOT` character. Sometimes the decision was quite difficult, which can be exemplified by the following passage

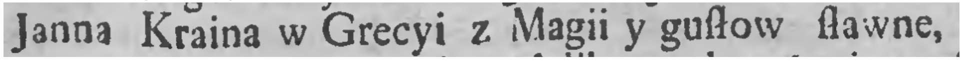


Jmpatiens otiiGræcos contra atq Latinos,

in which the last character of the last-but-one word was initially encoded as `LATIN SMALL LETTER Q WITH DIAGONAL STROKE`. After a discussion on the MUFI (Medieval Unicode Font Initiative) mailing list and consultation with Polish specialists, it was decided that the character actually is `LATIN SMALL LETTER Q LIGATED WITH FINAL ET`, a Private Use Area character assigned to the `E8BF` code point by the MUFI recommendation.

We tried to eliminate all the PUA characters not supported by MUFI-compatible fonts, so e.g. 664 occurrences of code point `F51D`, in the IMPACT project meaning `LATIN SMALL LETTER Z WITH HOOK ABOVE`, were replaced by a sequence of standard Unicode `LATIN SMALL LETTER Z` (`U+007A`) and `COMBINING HOOK ABOVE` (`U+0309`) characters.

We had however to accept as an exception a single PUA character, namely `F51E`, which we submitted to MUFI for a possible inclusion in the next version of the specification (`http://folk.uib.no/hnooh/mufi/pipeline/`). The character is `LATIN SMALL LIGATURE LONG S L WITH STROKE`. In the IMPACT corpus, it appears in over two thousand different words, in particular in the two last words of the following passage:



Janna Kraina w Grecyi z Magii y guſłow ſławne,

3 Creating the DjVu corpus

3.1 Text segmentation

The PAGE files used as the input for the corpus describe the page structure explicitly only on the level of the so called segments, which can be of several types:

paragraph (23771 occurrences in the corpus), heading (2 766 occurences), caption, page (4 646 occurences), page header (3465 occurences), page footer, page-number, drop-capital (1 017 occurrences), credit, floating, signature-mark (1 391 occurrences), catch-word (4 515 occurences), marginalia (963 occurences), footnote (6 occurences), footnote-continued or TOC-entry (502 occurences); for most regions the reading order is defined. The text content of elements preserves the line segmentation of the original, but the description of the location on the page is provided only for the region as a whole. This was done for financial reasons, as verifying the borders of smaller units was too expensive. At the very end of the IMPACT project some savings on other tasks allowed to provide such detailed information for a small part of the Ground-Truth texts, but they became available too late to be included in the corpus.

Poliqarp requires the texts to be provided in one of the following formats: XCES (*XML Corpus Encoding Standard*, `http://www.xces.org/`) or TEI4KJP (*TEI P5 Encoding of the National Corpus of Polish*, `http://nlp.ipipan.waw.pl/TEI4NKJP`); both formats are based on the recommendations of the Text Encoding Initiative consortium (`http://www.tei-c.org`). At present DjVu corpora are created using XCES with the hOCR format as an intermediary; hOCR is an open standard allowing HTML files to store additional OCR information, cf. (Breuel, 2007). The hOCR files provide also appropriate links to the scans for the Poliqarp for DjVu extension. The conversion from PAGE to hOCR is done using the pageparser, a script by Jakub Wilk. The segment types are converted (with some exceptions) into chunks, which can be referenced in the Poliqarp queries. However, the straightforward conversion is not sufficient for our purposes.

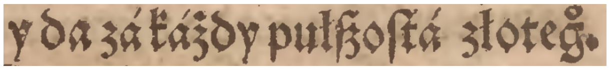
The most serious problem is to find a better approximation of the word bounding boxes than just whole regions; for the present a rather primitive algorithm is used to split the regions into lines. Moreover, it is desirable to optionally reconstruct hyphenated words; `HYPHEN-MINUS` (`002D`), `NON-BREAKING HYPHEN` (`U+2011`) and `DOUBLE OBLIQUE HYPHEN` (`U+2E17`) are taken into account for this purpose. These additional operations are performed by specialized scripts. For technical reasons, we create two versions of the corpus, called one-dimensional and two-dimensional: the former does not contain hyphenated words, while in the latter hyphenation breaks are preserved.

At the lowest level the corpus texts are segmented into tokens: words, punctuation marks and other symbols. Usually the algorithm described in the Unicode Standard Annex #29 *Unicode Text Segmentation* (`http://unicode.org/reports/tr29/`) provides desired results and it is therefore used in other DjVu corpora, but for the IMPACT corpus some *ad hoc* modifications were necessary:

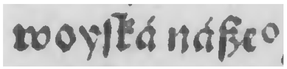
- All PUA characters are treated as letters, which turned out to be quite satisfactory (in the general case, the character properties should be provided somehow for the algorithm).

- The `REPLACEMENT CHARACTER` (`U+FFFD`, used for unreadable characters) is treated as a letter, because otherwise the words with a single unreadable letter would be split into several tokens.

- An interesting problem is posed by the use of `COMBINING LATIN SMALL LETTER O` as an abbreviation sign, cf. the following figures:

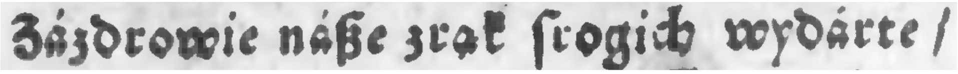
  y da zá káżdy pulßoſtá złotegͦ.

  woyſká náßeͦ

  In the first case, the last word is *złotego*, so the abbreviation has purely graphical character: all the letters are there, but *o* is printed over *g*. In the second case the word is *naszego*, the raised letter *o* meaning the ending *go*. Encoding the letter as `COMBINING LATIN SMALL LETTER O` seems appropriate from the semantic point of view, but as it is a combining character, there is a need for a base character. We decided to use `NON-BREAK SPACE` only in the very function of the base character for `COMBINING LATIN SMALL LETTER O`; in consequence `NON-BREAK SPACE` is also treated as a letter, otherwise the word would be broken into two tokens. The search for a more elegant solution requires the matter to be investigated further on a larger corpus of texts; perhaps introducing a spacing variant of the character will be necessary.

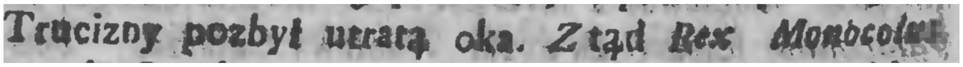
It should be noted that rules and customs of splitting the text into typographical words used to be different, e.g. the 5-word fragment



Záżdrowie náße zrąk ſrogich wydárte /

would be written now in 7 words (*Za zdrowie* instead of *Zazdrowie* and *z rąk* instead *zrąk* — for technical reasons we do not use here the facsimile transcription).

On the other hand the word spelled contemporarily as *stąd* occurs often as two words, *z* and *tąd*, e.g:



Trucizny pozbył utratą oka. Z tąd Rex Monocolus

In principle such problems can be handled by using regular expression queries, but some facilities helping user to search for such phenomena are highly desirable.

3.2 DjVu documents and metadata

The content of the corpus is described in details at `http://bc.klf.uw.edu.pl/289/`, so here it is only summarized briefly.

The core of the corpus is a 4-volume 18th century encyclopedia *Nowe Ateny* (*New Athens*) of almost 3,000 pages, known in particular as the source of a quite famous definition of a *horse*: *Koń, jaki jest, każdy widzi*, i.e. *Horse is as everyone can see* (this very good translation is provided by Wikipedia, cf. `http://en.wikipedia.org/wiki/Nowe_Ateny`).

Next come three 17th century books and an 18th century one, of total size of about 1,000 pages. All the books, including *Nowe Ateny*, are used as the source for the Late Middle Polish dictionary (*Słownik języka polskiego XVII i pierwszej połowy XVIII wieku*, work in progress, cf. `http://sxvii.pl/`), and it was one of the reasons to include them in the Ground-Truth set.

The rest of the corpus consists of 25 news pamphlets dated from 1570 to 1728 which range from 6 to 32 pages.

All the texts come from 4 digital libraries:

- Elbląska Digital Library (`http://dlibra.bibliotekaelblaska.pl/`), hosted by Elbląg library,
- Digital Library of Wielkopolska, serving in particular the scans from the Kórnik Library of the Polish Academy of Sciences (`http://www.wbc.poznan.pl`),
- Digital Library of Polish and Poland-Related News Pamphlets (`http://cbdu.id.uw.edu.pl/`), hosted by the Institute of Journalism, University of Warsaw.
- Lower Silesian Digital Library (`http://www.dbc.wroc.pl/`), hosted by the Wrocław University of Environmental and Life Sciences.

All the publications included in the corpus are already available in the respective libraries in the DjVu format. However, for the IMPACT project master scans had been provided, which were then subjected to some post-processing (e.g. splitting double pages) before being used in the project. Obviously, a DjVu corpus needs DjVu versions of the documents, but they were created from the graphic files used in the project, so they differ to some extent from those served by the libraries. The crucial difference is that in the corpus every page is a separate DjVu document with a separate set of metadata which include the URL of the equivalent page in the appropriate library. Other important metadata field is the title of the document; the remaining fields have a more or less technical character (in particular, they identify the PAGE file and the scan in the IMPACT database).

3.3 Corpus representation of tokens

The corpus representation of tokens was originally designed to store the lemmata of words, their parts of speech and grammatical properties. From the technical point of view, the representation consists of two free text fields and at least one field (the part of speech) with values from a pre-defined repertoire (specified in the corpus configuration file).

In standard Poliqarp clients, the two text fields differ in when and how they are displayed. The `orth` field, intended for the written representation of the token, is always displayed. The `base` field, intended for the lemma, is displayed only optionally and only in some situations. Therefore we decided to use the `orth` field for the textel transcription and the `base` fields for the facsimile transcription, although the reverse would be more logical.

We do not use any other fields (the mandatory part of speech field contains just a placeholder).

4 Using a DjVu corpus

4.1 Poliqarp for DjVu clients

The Poliqarp system is based on a client-server architecture. The Poliqarp for DjVu server can be accessed in four ways, with one of two available local clients, with a Web client or with a relatively recent remote client.

The local clients have some features not available for remote users. The command line client has no limits for the size of the results and can save them as CSV files for further processing. Another local client has a graphical user interface and additional facilities for statistical operations like collocation finding.

The Poliqarp Web client, called marasca-wbl, created by Jakub Wilk (primarily for Poliqarp for DjVu, but used now also by the National Corpus of Polish) is in principle the simplest way to use a DjVu corpus, but from time to time there are some problems with installing the DjVu plugin for the preferred Web browser. It is also possible to use it without a DjVu plugin by enabling the graphical concordances option, but it is not recommended because of some limitations of this mode and the substantially greater use of the server resources.

The most convenient way to access the corpus remotely is by using a specialized client, tentatively called djview for Poliqarp, designed by the present author and implemented by Michał Rudolf. It is a fork of the open source djview4 viewer developed by Léon Bottou (one of the inventors of the DjVu format). The code is known to compile on several popular Linux distribution such as Debian and Ubuntu, on OS/X for Macintosh and of course also on MS Windows. Additionally a MS Windows installer and a GNU/Linux Debian package are available for download at `https://bitbucket.org/mrudolf/djview-poliqarp`. It is still maintained despite the end of the project which supported its development.

4.2 Querying Poliqarp

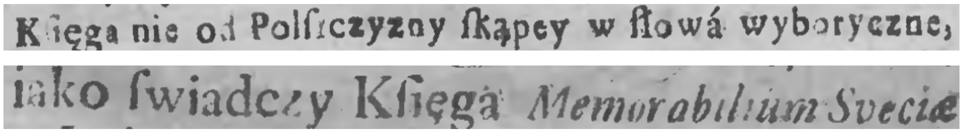
As described e.g. in (Przepiórkowski et al., 2004), Poliqarp queries are very powerful. They are based on two-level regular expressions. The first level describes classes of tokens by referring to their properties, the second level describes the sequences of those token classes.

When describing a class of tokens by referring to the text fields of their representation (`orth` or `base`), all the forms of the extended regular expressions can be used, e.g. POSIX character classes like `[:digit:]` and equivalence classes like `[=a=]`. Moreover, the regular expression syntax was extended making it possible to refer to a Unicode character by its code point, e.g. `\uf51e` refers to the character `LATIN SMALL LIGATURE LONG S L WITH STROKE` mentioned earlier. There are also some other minor extensions available.

The example below shows a case-insensitive (notice the `/i` flag) query for the word `księga`, independent of how the contemporary letters `s` and `ę` are actually represented.

```
'k[[=s=]]i.a'/i
```

The 48 hits returned by the query include in particular the following:



Kſięga nie od Polſzczyzny ſkąpey w ſłowá wyboryczne,

iako ſwiadczy Kſięgá *Memorabilium Sveciæ*

The problem of `s` is handled by using POSIX equivalence classes inside a "bracket expression", i.e. `[[=s=]]`, which allows us to match also a long *s* (by default the textel transcription is used with decomposed ligatures). The POSIX equivalence classes depend on a system setting called *locale*, which describes the language used, its collating sequence, etc. In the standard locale for Polish, `LATIN SMALL`

`LETTER E WITH STROKE` and `LATIN SMALL LETTER E WITH OGONEK` belong to different equivalence classes, so instead of such a class we can use an alternative. However in such a specific context it appeared sufficient to allow just any character as the fourth element of the word.

5 Future work

The most obvious and straight-forward goal is to put more information into the token representations. DjVu corpora can be automatically augmented by classifying the tokens at least into two categories: words and punctuation marks. More detailed classifications, accounting e.g. for numbers and symbols, can also be produced automatically.

A more ambitious goal is to allow different transcriptions in the same corpus by introducing an additional tag for transcription identifiers. At this level various ambiguities could be resolved, e.g. that of diacritical marks — although the Polish letter *z* accepts only an acute or a dot above, in the corpus those diacritics occur also in the shapes of a caron and a macron. Creating such a corpus is possible already, the problem is displaying the search results in a convenient way.

Last but not least, the tokens can be subject to standard linguistic analysis including lemmatization. In the framework of the IMPACT project, a small sample of the corpus was lemmatized using a morphological analyser of Polish and the CoBaLT system provided by Institute for Dutch Lexicology (an IMPACT project partner); cf. e.g. (Kenter et al., 2012). It would be interesting to apply those methods to the whole corpus.

However the future work will be affected by several important factors.

Recently at other institutions two projects have been started aiming at full-fledged historical corpora. The goals of the first project, which is to be implemented by the Institute of the Polish Language and the Institute of Computer Science of the Polish Academy of Sciences, are described as follows (quoted from `http://clip.ipipan.waw.pl/KORBA`):

> The aim of the project is the creation of a corpus of 17th and 18th century Polish texts (up to 1772) and tools for its processing (searching, filtering, summarizing statistical data, etc.). The entire corpus will feature annotation for text structure and language (all foreign elements, e.g. Latin intrusions, will be distinguished), and a portion of it will also feature morphological annotation.

The second project is to be implemented by the Institute of Literary Research of the Polish Academy of Sciences and is named *The corpus of 16th century Polish texts. Phase 1: text digitalization, creating software tools and making test data available.*

This means that the IMPACT corpus will probably not be developed further but directly or indirectly included in the new corpora.

Another important fact is the intention to develop a new version of Poliqarp, called Poliqarp2 (cf. `http://sourceforge.net/projects/poliqarp2/`), as part of the CLARIN-PL project (`http://www.clarin-pl.eu`). The project is a part of CLARIN ERIC (Common Language Resources and Technology Infrastructure, European Research Infrastructure Consortium), which in turn belongs to ESFRI (European Roadmap for Research Infrastructures, European Strategy Forum on Research Infrastructures). The new version will be designed to handle problems encountered in historical corpora, and to include page coordinates of tokens required by Poliqarp for DjVu (cf. `http://nlp.ipipan.waw.pl/NLP-SEMINAR/131021.pdf`). It is expected that the IMPACT corpus or its fragments will serve as test data on different stages of Poliqarp2 development and perhaps in due time the whole corpus will be migrated to the new system, which would however require developing a completely new client program.

6 Conclusions

Thanks to Poliqarp for DjVu, the Ground-Truth texts created for the IMPACT project were quickly made available for the Polish and international community in a relatively convenient way. This method can be applied also to other scans supplemented by OCR or transcription results available directly or indirectly in appropriate formats.

Acknowledgements:

The IMPACT project Polish Ground-Truth DjVu corpus was created, using the data and tools described above, by Krzysztof Szafran as a part of intensive testing of Poliqarp for DjVu. Earlier several quick-and-dirty versions of the corpus were created by Tomasz Olejniczak to facilitate some tasks of the IMPACT project.

Radosław Moszczyński and Mateusz Sarnecki helped with proofreading the article.

References

Bień, J. S. (2009). Facilitating access to digitalized dictionaries in DjVu format. *Cognitives Studies / Études cognitives*, *9*, 161–170. Retrieved from `http://bc.klf.uw.edu.pl/160/`.

Bień, J. S. (2011). 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*, volume 6699 of *Lecture Notes in Computer Science* (pp. 1–14). Berlin/Heidelberg: Springer. Retrieved from `http://dx.doi.org/10.1007/978-3-642-23160-51,http://bc.klf.uw.edu.pl/177/`.

Breuel, T. (2007). The hOCR microformat for OCR workflow and results. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition* (pp. 1063–1067). IEEE Computer Society. Retrieved from `http://madm.dfki.de/publication&pubid=4373`.

Kenter, T., Erjavec, T., Žorga Dulmin, M., & Fišer, D. (2012). Lexicon construction and corpus annotation of historical language with the CoBaLT editor. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 1–6). Avignon: Association for Computational Linguistics. Retrieved from `http://www.aclweb.org/anthology/W/W12/W12-1001.pdf`.

Le Cun, Y., Bottou, L., Haffner, P., & Howard, P. G. (1998). DjVu: a compression method for distributing scanned documents in color over the internet. In *Sixth Color Imaging Conference: Color Science, Systems and Applications* (pp. 220–223). Scottsdale, Arizona: IST. Retrieved from `http://leon.bottou.org/papers/lecun-98c`.

Pletschacher, S. & Antonacopoulos, A. (2010). The PAGE (Page Analysis and Ground-Truth Elements) format framework. In *International Conference on Pattern Recognition* (pp. 257–260). Los Alamitos, CA: USA. IEEE Computer Society. Retrieved from `http://www.impact-project.eu/fileadmin/Editorial/Documents/ICPR2010_The_PAGE_Format_Framework_USAL.pdf`

Przepiórkowski, A., Krynicki, Z., Dębowski, Ł., Woliński, M., Janus, D., & Bański, P. (2004). A search tool for corpora with positional tagsets and ambiguities. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC2004* (pp. 1235–1238). Retrieved from `http://nlp.ipipan.waw.pl/~adamp/Papers/2004-lrec/fcqp.pdf`.

Publisher: Institute of Slavic Studies PAS & University of Silesia in Katowice