



# Polish Summaries Corpus

Mateusz Kopec

Institute of Computer Science, Polish Academy of Sciences

`m.kopec@phd.ipipan.waw.pl`

August 7, 2015

This document presents the Polish Summaries Corpus (PSC), a resource created to support the development and evaluation of the tools for automated single-document summarization of Polish. It is loosely based on an article about PSC by Ogrodniczuk et al. [2014].

## 1 Corpus Desiderata

As an conclusion from the analysis of the existing corpora of Polish summaries and summaries corpora for other languages, I have designed the following desiderata for the new corpus:

- It should contain as large number of texts as possible, not fewer than a few hundred. For machine learning tools and general evaluation purposes, the bigger the better. This would overcome the main drawback of LAKON's corpus.

- There should be several different ratios of compression for summaries for each text, to allow for testing the behaviour of algorithms in different compression settings. This would improve on both LAKON's and Świetlicka's corpora.
- The corpus should contain extractive summaries, not limited to sentence selection, but rather word selection, to allow for research on human summarization techniques. My experience with manually creating clause-based extraction summaries shows, that clauses (or sentences) are too coarse elements for this task. It often happened, that a clause contained important as well as non-important information, which made it a difficult choice whether to select it into the summary or not.
- It should contain also abstractive summaries, written without any constraints imposed on the annotators, to be able to test the evaluation measures in such setting and also study the human summarization process.
- To overcome a single-annotator bias, each summary should have many versions, written by different, independent annotators.
- Source texts should come from press genre, as it is the most popular type of text consumed by the readers. According to Górski and Łaziński [2012], a balanced Polish corpus should contain about 50% of such genre. At the same time, it would be most useful for practical applications of the automatic summarization tool.
- The length of the source texts should be medium. Very long texts are not so common in the press, while very short ones are most often best summarized by their leading paragraph.
- Corpus texts should represent different press article domains (such as sport or politics), to enable research on differences between article types.
- Most importantly, to enable reproducible research, the corpus should be publicly available.

The corpus presented in this chapter fulfils these desiderata.

## 2 Corpus Source and Preprocessing

The Polish Summaries Corpus contains manual single-document summaries of press articles. This section presents the procedure for obtaining the texts, which were manually summarised.

## 2.1 Source data

Texts of the corpus were derived from the “Rzeczpospolita Corpus” (RC) by Weiss [2002] — a collection of articles from the Web archive of Rzeczpospolita, a nationwide Polish daily newspaper. RC consists of 190 379 pseudo-HTML files (1.9 GB data) dating from 1993 to 2002, with unequal representation of individual years. The data set has been made available by its owners (Presspublica, the publisher of the newspaper) for research and so far they have been used many times in various computational linguistic tasks [Broda et al., 2008, Piasecki and Radziszewski, 2008, Mykowiecka et al., 2007, Piskorski et al., 2008, 2007, 2009].

Every file in RC contains one or more articles (or practically none, when it references some non-textual content, such as a comic strip). Textual data is accompanied by HTML metadata (not always complete), such as the name of the newspaper subsection (DZIAL) in which the article was published, e.g.:

```
<META NAME="DZIAL" CONTENT="gazeta-sport">
```

where `gazeta-sport` can be translated as ‘newspaper-sport’. This subsection information, whenever filled in (empty for 8 165 files) was used to detect text domains (106 variants).

## 2.2 Data selection and conversion

Since the HTML code of the files in RC is not valid (particularly it does not contain an `<html>` tag), article borders have been detected using simple heuristics based on the verified assumption that particular HTML comments, output by the Presspublica archiving system, mark the beginning and end of each text. Aggregate and ‘empty’ texts have been removed from our result set by counting HTML elements representing document title (`<FONT SIZE="5">...</FONT>`). For the sake of our experiment, all texts have been finally converted to plain text and certain HTML content was completely removed (such as `<TABLE>...</TABLE>` or `<MENU>...</MENU>`).

By limiting the resulting data set to domains represented by more than 1000 articles sized between 1000 and 4000 words (arbitrary decision about the medium size of a press article), 7 most frequent domains were selected. Number of selected domains was chosen to have at least 30 texts in each one. In the last step of the data selection the articles were manually investigated to remove aggregates, legal acts or sports results (frequently published in the form of articles, but not suitable for the typical single-document news article summarization task).

<b>Text domain</b>	<b>Abstractive corpus</b>	<b>Extractive corpus</b>
Social and political	22	393
Sport	22	36
Economy	22	34
Cultural news	22	32
Law	22	26
National news	22	24
Science and technology	22	24
<b>Total</b>	<b>154</b>	<b>569</b>

Table 1: Selected domains

Out of these texts 569 were manually summarized: all of them have extractive summaries, 154 out of 569 have also abstractive summaries. The details about the summarization process are presented in the next section. Table 1 gives an insight into the distribution of the selected domains among the summarized texts. Because all the texts with abstractive summaries have also extractive summaries, one may use the corpus of 154 texts if he needs summaries of both kinds, while if only extractive summaries are required, one may benefit from larger, 569-text corpus.

Number of texts annotated was limited to 569 because of time and cost constraints, and the larger number of extractive summaries is due to the fact, that most of the automatic summarization systems are based on extraction techniques. Majority of texts in the extractive corpus is from social and political domain, as the number of texts from other domains in the “Rzeczpospolita corpus” was not enough to maintain the equal ratio of each type, as in the abstractive corpus.

### 3 Manual Summarization

Manual summarization was conducted by 11 annotators, who were randomly assigned texts to summarize. They were using three dedicated applications: for acquiring texts to work on (DISTSYS, available at <http://zil.ipipan.waw.pl/DistSys>), for creating abstractive summaries, and for creating extractive summaries (ABSSUMANNOTATOR and EXTRSUMANNOTATOR, both available at <http://zil.ipipan.waw.pl/SummaryAnnotationTools>).

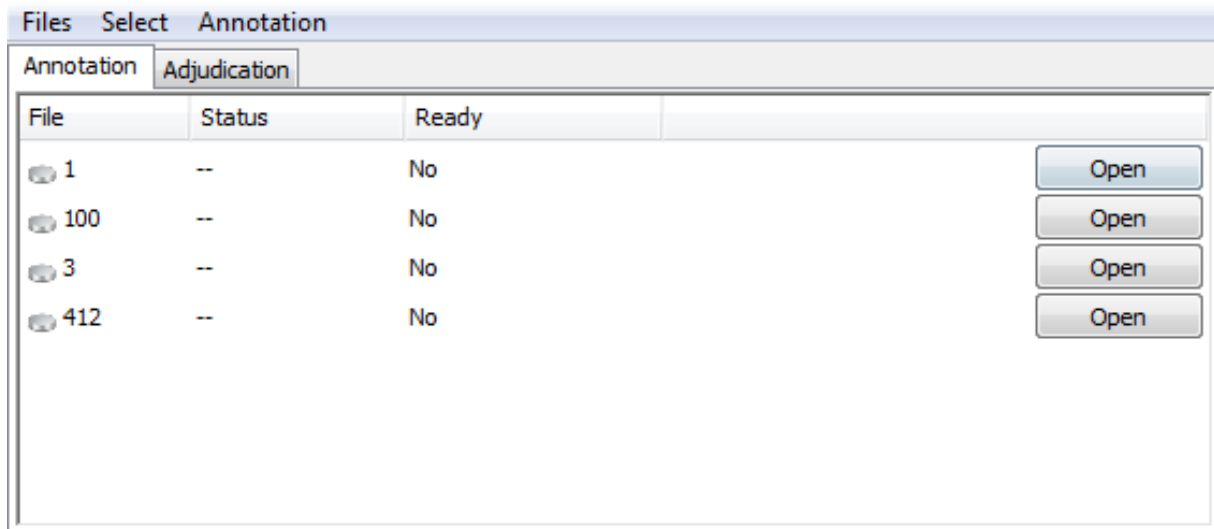


Figure 1: DISTSYS – application for text distribution

The text distribution tool follows a client-server scheme, with command-line server controlling which texts can be downloaded by which annotator. Annotators use client application with graphical interface (presented in Figure 1) to acquire texts, then these texts are summarized in another dedicated application. After work on a given document is finished, it may be uploaded to the central sever again via DISTSYS. This text distribution tool is of general purpose, in fact it has been used during manual annotation of the Polish Coreference Corpus [Ogrodniczuk et al., 2014].

Extractive summary annotation tool is depicted in Figure 2. It shows both the original text and the summary and facilitate selection of fragments as well as counting percentages on the fly. Three tabs allow for annotation of three summaries of different sizes for the text loaded. Similar application was used for the abstractive summary annotation.

### 3.1 Extractive summaries

Annotators were instructed to create three extractive summaries of a given text, each constituting approximately 20%, 10% and 5% of the word count of the original (for a 1000-word source text the resulting summaries should then respectively be 200, 100 and 50 words). Minor (a few word-length) deviations were acceptable to encourage annotators to select the most important fragments — and not the ones which would add up to the desired limit.

Only original words and punctuation in the original order had to be used (so that annotators could e.g. select just the superordinate clause and a finishing dot, removing the less important part of a sentence such as subordinate clauses, interjections, excessing adjectives — but not

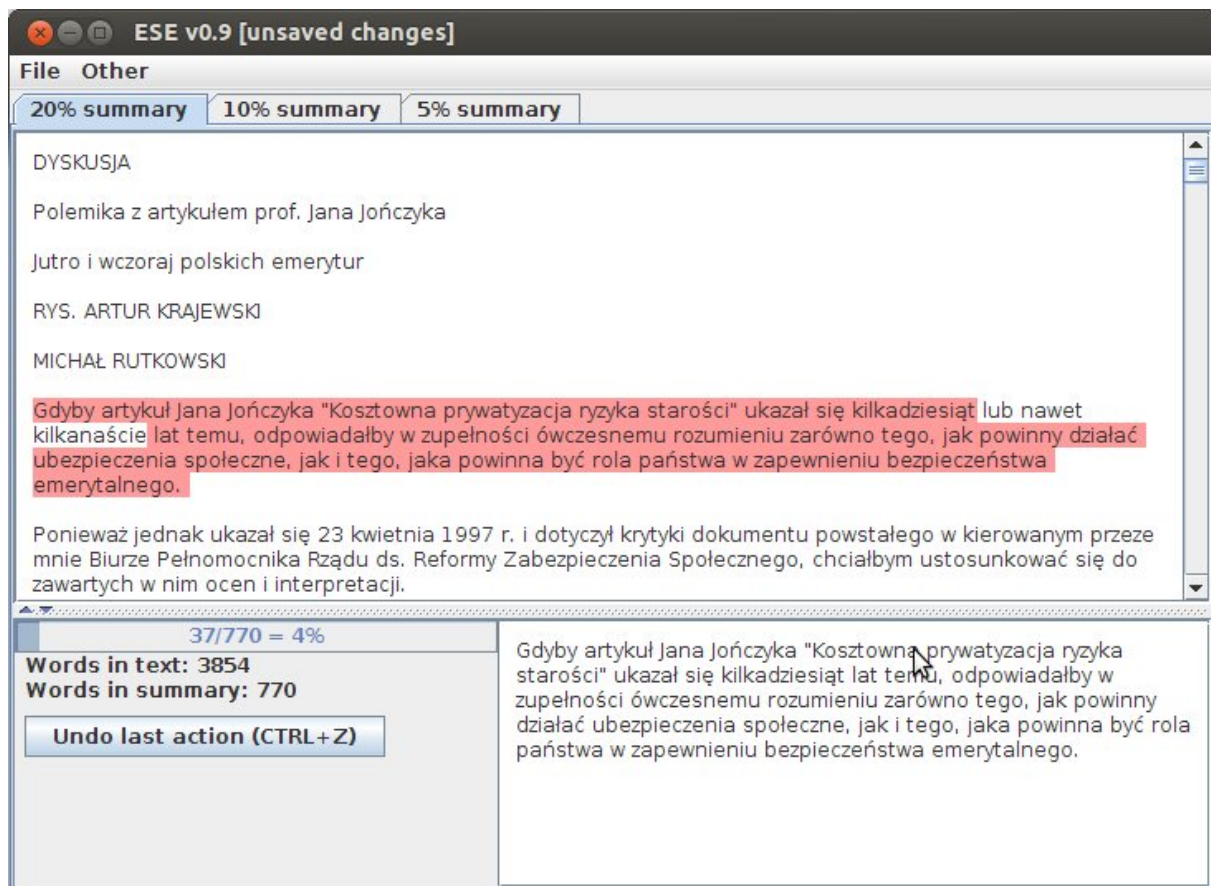


Figure 2: EXTRSUMANNOTATOR – an application for extractive summaries annotation

creating abbreviations from first letters of a proper name MWU). No document title, subtitle or author should be included, neither any information referring to the summarization process (such as “the text explains...”). The resulting summary was supposed to be grammatically correct and coherent, but tricks such as linking two phrases from two sentences with a conjunction coming from a third one were discouraged. As phrases could be selected and sentences combined, lowercase start of the sentence or an uppercase character in the middle of the resulting sentence was acceptable.

The sequence of summaries was forced to be inclusive, i.e. the 10-percent summary had to use only fragments previously selected for a 20-percent summary — and, similarly, the 5-percent summary had to use only fragments previously selected for a 10-percent summary. In this way a partial importance ranking of text spans could be inferred.

## **3.2 Abstractive summaries**

Similarly to the previous task, annotators were instructed to create 3 abstractive summaries of a given text, each constituting approximately 20%, 10% and 5% of the word count of the original, with acceptable minor deviations in word count.

Contrary to extractive summaries, abstractive summaries did not have to contain fragments of original texts and could express the same ideas “in own words” of an annotator. Differently from the extractive, longer summaries did not have to contain fragments of shorter ones, but they could.

### **3.2.1 Independent annotations**

Based on the opinions of many researchers, that there is no single “gold” summary for a given text (see for example one of the seminal works in the domain – Rath et al. [1961]), we decided to provide 5 independent versions of the summaries described above, each one written by a different annotator (yet single annotator always summarized to reach all three sizes: 5%, 10% and 15%). Such approach is supposed to overcome the annotator bias, which is often described as a problem during the evaluation of the summarization algorithms against a single gold standard. We have chosen to annotate 5 independent versions following the research of Nenkova, where 4 to 5 summaries is said to provide an optimal balance of annotation effort and reliability for the Pyramid method evaluation (see for example Nenkova et al. [2007]).

Therefore, because of 3 summary sizes for each text (20, 10 and 5%), our corpus contains altogether  $569 * 3 * 5 = 8535$  extractive summaries and  $154 * 3 * 5 = 2310$  abstractive ones, with makes a total of 10845 summaries.

## 4 Corpus Format

The corpus has been encoded in XML, each source document with all its summaries stored in a separate file, i.e. it consists of 569 files. XML Schema defining the structure of these files is attached to the corpus.

A sample text file is presented in Listing 1. Each document starts with `text` element, with `id` attribute defining the text identifier in the original “Rzeczpospolita Corpus”. Nested XML elements are:

- `date` – indicating date the text was originally published,
- `title` – title of the text,
- `section` – newspaper section the text appeared in (one of the sections indicated in Table 1),
- `authors` – author or authors of the text,
- `body` – original document content,
- `summaries` – element containing all the manual summaries for a given text.

The `summaries` element contain multiple `summary` tags, each containing a manual summary. A `summary` has the following attributes:

- `ratio` – target ratio of the summary (as percent of word count), possible values are: 20, 10, 5,
- `type` – indicates `extract` or `abstract` summary, no other values are present,
- `author` – single letter from A to K, indicating one of the eleven annotators.

In case of `abstract` summary, the only nested tag inside the `summary` tag is `body`, which stores the text of summary. In case of `extract` type, `body` is also present, but in addition we have `spans` element. This element stores the text spans marked by the annotator during summary extraction, each in a `span` tag. The tag has `start` and `end` attributes, marking character positions of the span in the original text, as well as the span content.



<?xml version='1.0' encoding='UTF-8'?>

<text id="199704210012">

<date>1997-04-21</date>

<title>Demokracji nie wolno deptać bezkarnie</title>

<section>Publicystyka , Opinie</section>

<authors>Teresa Stylińska</authors>

<body>GRECJA

21 kwietnia mija 30 lat od dnia, gdy grupa wojskowych obaliła rząd, wprowadziła stan wyjątkowy, a tysiące przeciwników wtrąciła do więzień

Demokracji nie wolno deptać bezkarnie

TERESA STYLIŃSKA

– Rządy pułkowników nigdy nie cieszyły się sympatią społeczeństwa. Dyktatura opierała się wyłącznie na wojsku, no i miała też za sobą poparcie USA i NATO – opowiada Jeorjos Aleksandros Mangakis, niegdyś członek ruchu oporu i więzień polityczny, dziś deputowany socjalistycznej partii PASOK.

[...]

– W dyktaturze – zauważa Jeorjos Mangakis – jest odwrotnie niż w naturze: osad nie opada, lecz idzie do góry.

</body>

<summaries>

<summary ratio="5" type="extract" author="I">

<body>21 kwietnia mija 30 lat od dnia, gdy grupa wojskowych obaliła rząd Grecji, wprowadziła stan wyjątkowy, a tysiące przeciwników wtrąciła do więzień. Sprawcy tych nieszczęść zostali ukarani. Grecy uważają, że darowanie im win byłoby czymś z gruntu niemoralnym. W więzieniu do dziś przebywa ścisła czołówka sprawców zamachu. Czas dyktatury przyniósł Grekom wiele cierpień, ale, paradoksalnie, miał też jeden

skutek pozytywny: gdyby nie potrzeba ugruntowania świeżej demokracji, Grecja, przy swym poziomie gospodarczym, miałyby znikome szanse na wejście do Europejskiej Wspólnoty Gospodarczej, jak wówczas nazywała się UE.

</body>

<spans>

<span start="499" end="541">21 kwietnia mija 30 lat od dnia, gdy grupa</span>

<span start="557" end="588">wojskowych obaliła rząd Grecji,</span>

<span start="634" end="705">wprowadziła stan wyjątkowy, a tysiące przeciwników wtrąciła do więzień.</span>

[...]

<span start="10908" end="11156">ale, paradoksalnie, miał też jeden skutek pozytywny: gdyby nie potrzeba ugruntowania świeżej demokracji, Grecja, przy swym poziomie gospodarczym, miałyby znikome szanse na wejście do Europejskiej Wspólnoty Gospodarczej, jak wówczas nazywała się UE.</span>

</spans>

</summary>

<summary ratio="5" type="abstract" author="I">

<body>Dwudziestego pierwszego kwietnia mija 30 lat od chwili, gdy w Grecji grupa wojskowych obaliła rząd i wtrąciła do więzień tysiące przeciwników. Reżim nie był w stanie zdobyć uznania i poparcia społecznego, nie był też na tyle silny, by długo opierać się na represji. Punktem zwrotnym był listopad 1973 roku, gdy krwawo stłumiono studentów politechniki ateńskiej. W 1974 roku władza junty się rozsypała. Rozpoczęły się procesy jej członków, kilkoro otrzymało wyroki śmierci, które później zmieniono na dożywocie. Wielu uważa, że spuścizną dyktatury jest przede wszystkim rozluźnienie zasad i zniszczenie uczciwości w narodzie.</body>

```
</summary>

[ ... ]

</summaries>
</text>
```

Listing 1: Example corpus text encoding

Alongside the corpus data, an API (Application Programming Interface) for the corpus was created, to allow an easy access to the corpus for applications written in Java. Consult corpus web page for details.

## References

- Bartosz Broda, Maciej Piasecki, and Stanisław Szpakowicz. Sense-based clustering of polish nouns in the extraction of semantic relatedness. In *Computer Science and Information Technology, 2008. IMCSIT 2008. International Multiconference on*, pages 83–89. IEEE, 2008. Cited on page 3.
- Rafał L. Górski and Marek Łaziński. Reprezentatywność i zrównoważenie korpusu. In Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors, *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*, pages 25–36. Wydawnictwo Naukowe PWN, Warsaw, 2012. Cited on page 2.
- Agnieszka Mykowiecka, Anna Kupść, Małgorzata Marciniak, and Jakub Piskorski. Resources for information extraction from polish texts. In *Proceedings of the 3rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics,(LTC’2007), Poznań, Poland*, pages 5–7, 2007. Cited on page 3.
- Ani Nenkova, Rebecca J. Passonneau, and Kathleen McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. *TSLP*, 4(2), 2007. doi: 10.1145/1233912.1233913. Cited on page 7.
- Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. *Coreference: Annotation, Resolution and Evaluation in Polish*. Walter de Gruyter Inc., 2014. Cited on pages 1 and 5.

Maciej Piasecki and Adam Radziszewski. Morphological prediction for polish by a statistical a tergo index. *Systems Science*, 34(4):7–17, 2008. Cited on page 3.

Jakub Piskorski, Marcin Sydow, and Anna Kupść. Lemmatization of polish person names. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 27–34. Association for Computational Linguistics, 2007. Cited on page 3.

Jakub Piskorski, Karol Wieloch, M Pikula, and Marcin Sydow. Towards person name matching for inflective languages. In *WWW 2008 Workshop NLP Challenges in the Information Explosion Era*, 2008. Cited on page 3.

Jakub Piskorski, Karol Wieloch, and Marcin Sydow. On knowledge-poor methods for person name matching and lemmatization for highly inflectional languages. *Information retrieval*, 12(3):275–299, 2009. Cited on page 3.

G. J. Rath, A. Resnick, and R. Savage. The formation of abstracts by the selection of sentences. *American Documentation*, 12(2), 1961. Cited on page 7.

Dawid Weiss. Korpus Rzeczpospolitej. [on-line] <http://www.cs.put.poznan.pl/dweiss/rzeczpospolita>, 2002. Cited on page 3.