Parallel and spoken corpora in an open repository of Polish language resources

Piotr Pęzik[1], **Maciej Ogrodniczuk**[2], **Adam Przepiórkowski**[2]

[1]University of Łódź
[2]Institute of Computer Science, Polish Academy of Sciences

Abstract

The aim of this paper is to present current efforts towards the creation of a comprehensive open repository of Polish language resources and tools (LRTs). The work described here is carried out within the CESAR project, member of the META-NET consortium. It has already resulted in the creation of the Computational Linguistics in Poland site containing an exhaustive collection of Polish LRTs. Current work is focused on the creation of new LRTs and, esp., the enhancement of existing LRTs, such as parallel corpora, annotated corpora of written and spoken Polish and morphological dictionaries to be made available via the META-SHARE repository.

Keywords: META-SHARE, parallel corpora, spoken corpora, morphological dictionaries, annotated corpora, TEI, XLiFF, Polish

1 Introduction

One of the main aims of the CESAR project, part of the META-NET consortium (`http://www.meta-net.eu/projects/cesar`), started in February 2011, is to make existing language resources and tools (LRTs) for Central and East European languages more readily available and more widely used. This aim is being achieved via three main means:

1. increasing the awareness of existing LRTs,
2. increasing their availability, esp., via standard and liberal licensing,
3. increasing their reusability, esp., via ensuring adherence to standards and by increasing their quality.

The aim of this paper is to report on some early successes and plans with respect to these points. While the work described here is carried out in the context of Polish LRTs, we concentrate here on multilingual resources, esp., parallel corpora.

The outline of the paper is as follows: §2 presents the idea of the open repository of Polish LRTs, §§3–5 briefly describe some of the main Polish resources enhanced within CESAR, while §6 covers one of them – parallel corpora – in more detail. Subsequently, §7 mentions other resources under consideration in the project and, finally, §8 concludes the paper.

2 Towards an open repository

To fulfil the need for increasing the awareness of existing LRTs for Polish and reinforcing relations between the key players in Polish natural language processing (NLP), a new Web portal "Computational Linguistics in Poland" (CLIP, `http://clip.ipipan.waw.pl`) was established in mid-April 2011, following the idea of the page operated between 2000 and 2004 at the Institute of Computer Science, Polish Academy of Sciences. The site aims at containing exhaustive information about LRTs, research centres, projects and linguistic engineering courses related to Polish. Furthermore, it intends to bring language-related initiatives, institutions and people from research, government and industry communities together, offering them comprehensive information on available language technology. One of the main design principles of the site was to maintain a wiki-like mode of operation, allowing the authorised representatives of all LRT groups in Poland to edit the content directly. This assumption proved very fruitful and several modifications and additions have already been made by external editors. According to our best knowledge the site is currently the largest repository of references to publicly available Polish LRTs.

Along with creating synergies within the national language community, the Polish partners are playing an active role in META-SHARE – the open language resource exchange infrastructure created by META-NET and operated at the European level. Its main function is sustainable sharing and dissemination of LRTs on a global scale. The operational level of META-SHARE is a network of distributed repositories providing a multi-layer infrastructure for OAI-PMH[1]-enabled exchange of LRTs and related metadata, as well as interfaces for remote indexing of LRs. However, the initiative goes far beyond the repository setup: it promotes the use of widely acceptable LR standards ensuring their maximum interoperability and sustainability, advertises its own CC-based licensing models and IPR provisions, offering legal and organizational support in the form of licensing templates, language resource sharing forms, ready-to-use agreement declarations and various other LR-related recommendations.

Regarding its technological impact, CESAR targets specific Polish language processing resources with a view to improving their *availability*, *interoperability* and *representativeness*. In the remaining part of this paper we introduce a number of key resources whose availability and interoperability will be improved within the CESAR project. We start with **morphological dictionaries** and **annotated corpora**, which are a basic prerequisite for most NLP solutions. Due to technical difficulties, spoken discourse corpora and speech databases are sparsely distributed across different languages. A separate subtask of the CESAR project is thus concerned with the release of **a corpus of casual spoken** Polish including a subset of time-aligned transcriptions, as well as **a speech database** of telephone

[1] Open Archives Initiative Protocol for Metadata Harvesting, see `http://www.openarchives.org/OAI/openarchivesprotocol.html`.

conversations. The next section of this paper outlines CESAR's contribution to improving the availability of cross-linguistic NLP resources through the acquisition of new and existing **parallel corpora** annotated in widely accepted text encoding and translation memory formats such as TEI (Text Encoding Initiative; Burnard and Bauman 2008) and XLiFF[2].

These various language resources together with a Polish Wordnet, dictionaries of named entities and a treebank of Polish will be gradually released as part of the open repository in three batches planned for November 2011, June 2012 and January 2013 respectively.

3 Dictionaries

Morphological dictionaries are about the most basic language resources, and most NLP tasks require their existence and availability. Until recently, there has only been one morphological dictionary available for Polish under an open source licence (LGPL and Creative Commons), namely, Morfologik (`http://morfologik.blogspot.com/`; not to be confused with the Hungarian NLP company Morphologic). Another morphological analyser, Morfeusz (`http://sgjp.pl/morfeusz/`; Woliński 2006, Saloni et al. 2007), whose quality is widely believed to be higher than that of Morfologik, was available under a closed – albeit free for non-commercial applications – licence. These two tools seem to be the most widely used morphological analysers for Polish; actually, both are used in the National Corpus of Polish (`http://nkjp.pl/`; Przepiórkowski et al. 2010, 2011).

Largely due to the efforts at the very initial stages of CESAR, the owners of the data of both dictionaries agreed to release them on a very liberal open source licence (the FreeBSD licence, also known as the 2-clause BSD licence). Moreover, again within CESAR, cooperation between the maintainers of the dictionaries has been initiated, leading to the creation of a single large morphological dictionary for Polish, comprising and extending both Morfologik and Morfeusz. A dedicated tool for extending the dictionary with new lexemes is currently in the final stages of development. The tool will allow linguists to add lexemes and their morphological specification in a distributed fashion, over the Internet. Various quality control mechanisms have been implemented, to minimise errors in the resulting dictionary.

The first version of the new morphological dictionary resulting from the automatic merger of Morfeusz and Morfologik will be available as early as November 2011; the complete and supplemented version will be compiled by January 2013.

4 Annotated Corpora

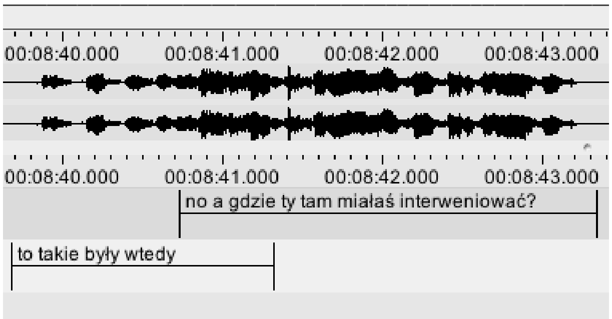
Manually annotated corpora are important resources, used for training various language processing tools. One of the most basic such tools are morphological taggers, used for disambiguating the results of morphological analysers. The most comprehensive resource of this kind for Polish is the 1-million-word subcorpus of the National Corpus of Polish (Pol. *Narodowy Korpus Języka Polskiego*; NKJP), manually annotated at various linguistic levels, including the morphosyntactic level. However, for a morphologically rich language, 1 million words is not sufficient to attain the same tagging accuracy as, for example, for English (over 97%); in fact, current Polish taggers perform at the level of 92–93%.

In order to improve these results, two kinds of activities are undertaken in CESAR. First, although a very careful annotation procedure was adopted in NKJP (Przepiórkowski and Murzynowski 2010), annotation errors may readily be found in the corpus, so known issues are corrected manually and semi-automatically within CESAR. Additionally, statistical methods (of the kind described in Dickinson and Meurers 2003) are employed to discover unknown errors.

Second, an additional corpus of 500 thousand words is annotated within CESAR, with the aim of creating a high-quality 1.5-million-word training corpus. However, in order to minimise costs, an existing corpus is used for this purpose, namely, the "Polish language of the 1960s" corpus (`http://clip.ipipan.waw.pl/PL196x`; Ogrodniczuk 2003). The corpus was originally manually annotated with a much more limited tagset than that currently used for Polish, so the work consists in the semi-automatic conversion the annotation of that corpus to the current standards and – most importantly – in its independent re-annotation. These two annotations are compared and any differences are sent for adjudication, thus increasing the annotation quality.

5 Spoken Corpora

Corpora of casual spoken discourse are a rather rare resource for many languages. The largest collection of transcriptions of naturally occurring conversational Polish has been compiled by the PELCRA team[3] at the University of Łódź since 2000, initially as part of the PELCRA reference Corpus and later within the National Corpus of Polish (Pęzik 2011). In total, the corpus contains almost 2 million words of transcriptions of conversations recorded in an informal setting, often without some of the speakers knowing they were being taped (although they had been informed about and agreed to the possibility of being recorded and later granted their permission to transcribe the recordings).



no a gdzie ty tam miałaś interweniować?

to takie były wtedy

Figure 1: A sample of the time-aligned corpus of conversational Polish.

So far this data has been only available through on-

[2] `http://docs.oasis-open.org/xliff/xliff-core/xliff-core.html`

[3] See `http://pelcra.pl`.

line search interfaces, but within CESAR a subset of this data will be made available in the TEI P5 format following some privacy considerations. Furthermore, a selection of the transcriptions are being time-aligned with the original recordings at the level of utterances and made available under a CC-like license through the META-SHARE repository.

Another multimedia speech corpus planned to be included into META-SHARE repository is the TEI-encoded corpus of transliterated complex spontaneous human-human telephone conversations acquired in the course of LUNA (Spoken Language UNderstanding in multilinguAl communication systems; `http://www.ist-luna.eu`; Marciniak 2010) project. The source data have been collected at the call centre of the Public Transport Authority of Warsaw and annotated in terms of semantic constituents and semantic structures (Mykowiecka and Waszczuk 2008).

6 Parallel Corpora

Parallel corpora are among the most important resources used in multilingual language processing. On the one hand, they serve as training data sets in machine translation systems, cross-linguistic information retrieval and in the construction of bilingual dictionaries. Depending on their annotation format, they can also be used, more or less readily, as translation memories, as well as an empirical basis in comparative linguistic and translation studies. Although a number of freely available public domain and open license parallel resources exist for Polish, they generally suffer from problems which seriously affect their usability and interoperability. First of all, they are available from a relatively large number of different sources, which often makes it difficult to identify the right set of corpora to use for a particular purpose. Secondly, when it comes to annotation standards, Polish parallel corpora and translation memories come in many shapes and sizes. Some resources are available as translation memories without any text structure annotation. Some parallel corpora are little more than plain text collections which encode segment boundaries using simple line breaks, while others make use of sophisticated annotation schemas which make it possible to express non-trivial cases of equivalence between segments, such as non-sequential cross-links, deletions, insertions, or segment splits and mergers. Apart from technical problems, the representativeness of openly available parallel resources leaves much to be desired. The majority of freely available parallel corpora and translation memories are public domain legal collections and open license software and technical documentation localization memories. This in turn means that non-technical and non-legal text genres and language registers tend to be poorly represented in openly available corpora.

In order to illustrate the contribution of the CESAR project to the availability, interoperability and representativeness of parallel corpora of Polish let us consider the first batch or such resources to be delivered in November 2011 (Table 1).

Collection	Lang. pairs	Alignment	Original format	Documents
PAS Academia	1	Sentence, manual	PDF, DOC	500
CORDIS	1 (5)	Sentence, automatic	HTML	10 000
RAPID	1 (21)	Sentence, automatic	HTML	4 900
JRC Acquis Communautaire	1 (21)	Sentence, automatic	TEI	26 000

Table 1: The first batch of Polish parallel corpora due in November 2011.

The first source made available contains some 500 scientific articles in Polish and English from *Academia – the Magazine of the Polish Academy of Sciences*. The articles were first converted from the PDF format and aligned semi-automatically at the sentence level using the memoQ CAT environment. The initial sentence-level alignment was then manually verified and the texts were further annotated with bibliographic information. Since this data is a completely new parallel resource for Polish, it can be considered as an example of CESAR's contribution to improving the coverage and representativeness of Polish parallel corpora. The CORDIS collection contains over 10 000 articles published at `http://cordis.europa.eu/` – the Community Research and Development Information Centre in Polish and 5 other EU languages. The CORDIS and RAPID collections were web-crawled parsed for contents using dedicated web-crawlers and aligned with mALIGNa (Jassem and Lipski 2008). Compared with the 2010 version of these two collections available from the Information Systems Laboratory at the Adam Mickiewicz University, the CESAR version features structural and bibliographic annotation adhering to the TEI and XLiFF formats described below. We have also decided to include the Polish-English component of the JRC version of Acquis Communautaire in the first batch of resources for the sake of its increased availability in the output formats. In the process of converting the JRC Acquis corpus additional bibliographic information was added to the metadata headers.

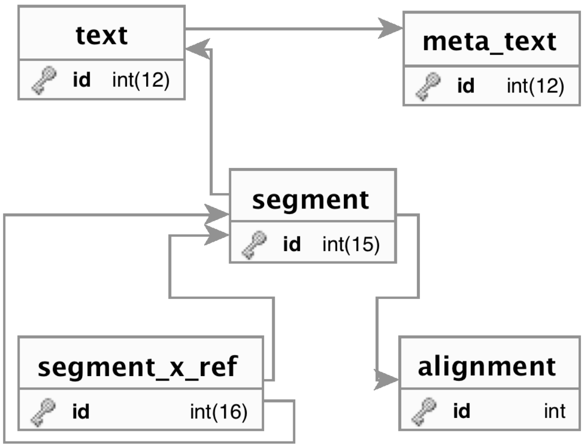


Figure 2: Core tables of the parallel corpus database.

The process of converting, processing and exporting parallel resources encoded in a variety of formats is facilitated by the use of a central relational database system (named *Paralela*) to which texts collections are imported in the first phase of the acquisition process. The use of RDB systems for the management of large collections of texts is motivated by our experiences in the National Corpus of Polish, in which a relational database was used to manage a collection of four million texts. The *Paralela* database is used to store bibliographic, structural and alignment information, and it has been designed to handle multiple alignments for the same collection. For example, the JRC Acquis Communautaire collection originally comes with two alignments both of which are retained in the database. The core tables of the database are presented in Fig. 2.

General bibliographic information about texts is stored in the `meta_text` table, while language specific metadata can be found in the `text` table. Arbitrary-length segments of texts are stored in the `segment` table and aligned through the references in `segment_xref` table. Further information about the type of alignment of each segment is stored in the `alignment` table. This simple schema combined with the expressive power of SQL and programming interfaces to RDBs makes the database a very efficient parallel text management system.

Once the variously encoded collections are converted and normalised in the database, they can be processed and exported into more uniform and standard formats used for the exchange of parallel corpora and translation memories. We have decided to provide the parallel data in two main formats TEI and XLiFF. The first format is a widely recognised standard of annotating corpus data with good support for encoding structural, bibliographic and alignment annotation. We expect this format to be a more natural choice in corpus analysis and NLP contexts as it can be used to mark up information about alignment splits, mergers as well as many-to-many segment linking, which proves useful when manually aligning texts. A fragment of bilingual bibliographic and alignment annotation encoded in TEI is shown in List. 1.

The XLiFF format, on the other hand, although much less expressive, is supported by all major CAT environments as an increasingly popular way of exchanging translation memories. Any subset of the parallel collections can thus be used directly as a translation memory in a modern CAT environment. An overview of the process of conversion is shown in Fig. 3.

Apart from the collections mentioned above as part of the first batch, specific resources are being targeted now for the inclusion in the next two batches of META-NET resources to be released in mid-2012 and early 2013. We believe that the total pool of Polish parallel corpora released in the CESAR project will be an important contribution to the availability, representativeness and interoperability of such resources.

7 Other resources

Apart from the above-mentioned core resources, the processing of which seems the most time-consuming and labour-intensive, another set of equally important resources will be made available through META-SHARE channels. The most prominent of them is the Polish Wordnet (Piasecki et al. 2009), still actively developed and therefore planned to be issued in all three CESAR batch editions, in November 2011, July 2012 and January 2013.

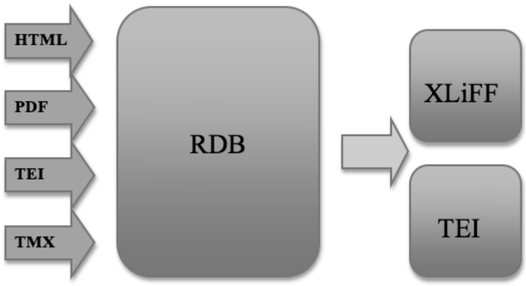


Figure 3: Source texts are imported into a relational database and exported in XLiFF and TEI formats. Some of the source formats (e.g. HTML, PDF) required manual and/or automatic alignment.

Another important resource, scheduled for January 2013, is the merger of existing dictionaries of Polish Named Entities. Various resources are planned to be gathered (e.g. from Tours, Poznań, Warszawa and Wrocław) and standardised within this task by encoding them in the LMF (Lexical Markup Framework; ISO:24613) format.

Last but not least, Marcin Woliński's treebank of Polish constructed using automatic syntactic analysis will be made available in mid-2012.

8 Conclusion

Being part of META-NET creates a good opportunity to work out the long-term sustainability plan for the important LRTs, which must be extended, linked, preferably multilingually aligned, but first of all upgraded to recommended representation standards. Starting with technical interoperability provided by Unicode and XML it is vital to maintain the standardisation principle also at the syntactic and semantic levels. Although the latter problem still remains open, even though tackled by several ongoing initiatives such as ISOCat Data Category Registry[4] and its instantiations (such as the one described in Patejuk and Przepiórkowski 2010), keeping the resource-structure layer seems a much more straightforward task. For Polish resources the recommendations of FLaReNet and CLARIN are being followed, including LMF for the representation of dictionaries, XLIFF for parallel corpora and TEI for various textual resources. The conversion and maintenance of resources scheduled for META-SHARE inclusion in these formats is an important mission of the META-NET / CESAR project.

Acknowledgements

Research funded in 2011–2013 within CESAR (CEntral and South-east europeAn Resources; `http://www.meta-net.eu/projects/cesar`), a European (CIP ICT-PSP) project (grant agreement 271022), part of META-NET.

[4] See `http://www.isocat.org/interface/index.html`.

```
<bibl xml:id="bibl-3">
  <ptr target="#text-3"/>
  <relatedItem xml:lang="en" type="original">
    <bibl>
      <title level="a">EU funds methane in sea floor research</title>
      <date type="published" when="2005-01-04">2005-01-04</date>
    </bibl>
  </relatedItem>
  <relatedItem xml:lang="pl" type="translation">
    <bibl>
      <title level="a">UE finansuje badania nad metanem na dnie morskim</title>
      <date type="published" when="2005-01-04">2005-01-04</date>
    </bibl>
  </relatedItem>
</bibl>
<!-- Stand-off segment alignment: -->
<linkGrp>
    <link target="#div-11 #div-15" type="merge"/>
    <link target="#div-12 #div-16" type="split"/>
    <link target="#div-13 #div-17" type="complex"/>
    <link target="#div-14 #div-18" type="simple"/>
</linkGrp>
```

Listing 1: A snippet of the bilingual annotation in the output TEI format.

References

Lou Burnard and Syd Bauman, editors. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford, 2008. http://www.tei-c.org/Guidelines/P5/.

Markus Dickinson and W. Detmar Meurers. Detecting errors in part-of-speech annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, pages 107–114, Budapest, 2003.

ISO:24613. Language resource management – lexical markup framework (LMF), 2008. ISO/FDIS 24613, ISO TC 37/SC 4 document N 45 of 2008-03-21.

Krzysztof Jassem and Jarosław Lipski. A new tool for the bilingual text aligning at the sentence level. In *Intelligent Information Systems*, Warsaw, 2008. Akademicka Oficyna Wydawnicza EXIT.

Małgorzata Marciniak, editor. *Anotowany korpus dialogów telefonicznych*. Akademicka Oficyna Wydawnicza EXIT, Warsaw, 2010.

Agnieszka Mykowiecka and Jakub Waszczuk. Semantic annotation of city transportation information dialogues using CRF method. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue: 12th International Conference, TSD 2009, Pilsen, Czech Republic, September 2009*, volume 5729 of *Lecture Notes in Artificial Intelligence*, pages 411–419, Berlin, 2008. Springer-Verlag.

Maciej Ogrodniczuk. Nowa edycja wzbogaconego korpusu słownika frekwencyjnego. In Stanisław Gajda, editor, *Językoznawstwo w Polsce. Stan i perspektywy*, pages 181–190. Komitet Językoznawstwa, Polska Akademia Nauk and Instytut Filologii Polskiej, Uniwersytet Opolski, Opole, 2003. http://www.mimuw.edu.pl/~jsbien/MO/JwP03/.

Agnieszka Patejuk and Adam Przepiórkowski. ISOcat definition of the National Corpus of Polish tagset. In *LREC 2010 Workshop on LRT Standards*, Valletta, Malta, 2010. ELRA.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wroclawskiej, Wrocław, 2009.

Adam Przepiórkowski and Grzegorz Murzynowski. Manual annotation of the National Corpus of Polish with Anotatornia. In Stanisław Goźdź-Roszkowski, editor, *The proceedings of Practical Applications in Language and Computers PALC 2009*, Frankfurt am Main, 2010. Peter Lang.

Adam Przepiórkowski, Rafał L. Górski, Marek Łaziński, and Piotr Pęzik. Recent developments in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta, 2010. ELRA.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw, 2011. Forthcoming.

Piotr Pęzik. Język mówiony w NKJP. In *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw, 2011. Forthcoming.

Zygmunt Saloni, Włodzimierz Gruszczyński, Marcin Woliński, and Robert Wołosz. *Słownik gramatyczny języka polskiego*. Wiedza Powszechna, Warsaw, 2007.

Marcin Woliński. 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*, Advances in Soft Computing, pages 511–520. Springer-Verlag, Berlin, 2006.