Towards a comprehensive open repository of Polish language resources

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

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

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 website 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. Efforts are made to ensure a high level of reusability of the LTRs by adhering to widely accepted annotation and interoperability standards. Last but not least, since the great majority of the Polish CESAR resources are released under open licenses, special work is required to clarify their Intellectual Property Rights status.

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

1. Introduction

There are significant gaps in the distribution of key language resources and tools (LRTs) across different European languages. 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 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.

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 approach 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, Polish CESAR partners play 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.

This initiative goes far beyond the repository setup: it promotes the use of widely acceptable LR standards ensuring their maximum interoperability and sustain-

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

ability, advertises its own CC-based licensing models and IPR provisions, offering legal and organisational 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 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 & Bauman 2008) and XLiFF[2]. These 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, 2012).

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 PoliMorf (Woliński et al. 2012), 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 allows 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 was made available in 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% (Piasecki 2007, Karwańska & Przepiórkowski 2011, Acedański 2010).

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, annotation errors may readily be found in the corpus, so known issues are corrected manually and semi-automatically within CESAR. Additionally, statistical methods 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.

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

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 2012). 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).

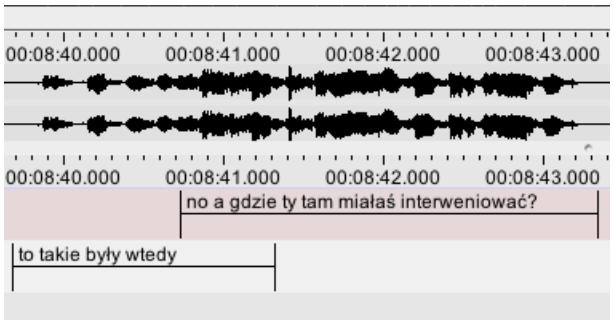

Figure 1: A sample of the time aligned corpus of conversational Polish.

So far this data has been only available through online 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 the GPL 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 & Waszczuk 2008).

The University of Łódź is also currently cooperating with industry partners to provide a specialized Spelling and Number Voice recognition corpus (SNUV) which could improve the performance of Polish voice recognition systems. The corpus will transcribed and time-aligned recordings of hundreds of Polish speakers reading numbers and spelling words, as these two areas are often key in real-life voice recognition systems. We are planning to release this corpus as part of the last batch of CESAR resources in January 2013.

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

6. Parallel Corpora

The Polish branch of CESAR makes an important contribution to the availability and interoperability of parallel corpora of Polish as illustrated in (Table 1). Some of the resources listed in the table are completely new and aligned manually (i.e. Polish Academy of Sciences Academia or Centre for Eastern Studies Corpus). Others have previously been available as parallel corpora in rather minimalistic formats (CORDIS & RAPID) or lacked some bibliographic metadata (JRC Acquis Communautaire) which we considered to be important in advanced parallel corpus search applications.

Collection	Lang. pairs	Alignment	Original format	Documents
PAS Academia	1	Sentence, manual	PDF, DOC	500
CORDIS	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.

The process of converting, processing and exporting parallel resources encoded in a variety of formats (ranging from HTML and PDF to TEI) is facilitated by the use of a central relational database system (named *Paralela*) to which text collections are imported in the first phase of the acquisition process. The *Paralela* database is used to store bibliographic, structural and alignment information, and it has been designed to handle multiple alignments of the same collection.

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, namely 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. 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 paral-

lel collections can thus be used directly as a translation memory in a modern CAT environment. The general workflow of the process of conversion is shown in Fig. 2.

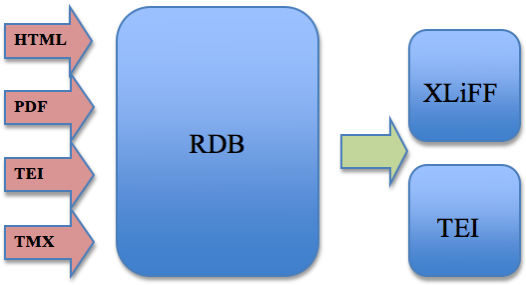

Figure 2: Source texts are imported into a relational database and exported in XLiFF and TEI formats. Some of the source formats (e.g. HTML, PDF) required manual and/or automatic alignment.

More Polish parallel corpora are being prepared for the next two batches of resources covering more than twenty different language pairs. Apart from automatically aligned collections, Polish-English and Polish-Russian corpora are being made available with manual alignment annotation for non-trivial segment corresponded cases, which we believe can be used to evaluate and improve the performance of statistical translation aligners.

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.

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 2008) format.

The last batch of Polish CESAR resources will also include Polish and English dictionaries of collocations containing some 2 million potential collocations each extracted from the National Corpus of Polish and the British National Corpus. For each potential collocation a number of association and dispersion measures are computed and recorded in the dictionary along with annotations of part-of–speech patterns in which they were found. The dictionaries will be available in the form of relational databases and they will hopefully be used to complement paradigmatically oriented lexical databases with syntagmatic information about the phraseological potential of word patterns.

Finally, Składnica (Woliński et al. 2011), a treebank of Polish constructed semi-manually on the basis 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 & 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 2010–2013 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.

References

Acedański, S. (2010). A morphosyntactic Brill tagger for inflectional languages. In H. Loftsson, E. Rögnvaldsson, and S. Helgadóttir, editors, *Advances in Natural Language Processing: Proceedings of the 7th International Conference on Natural Language Processing, IceTAL 2010, Reykjavík, Iceland*, volume 6233 of *Lecture Notes in Artificial Intelligence*, pages 3–14, Heidelberg. Springer-Verlag.

Burnard, L. & Bauman, S., editors (2008). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford. `http://www.tei-c.org/Guidelines/P5/`.

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

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

Karwańska, D. & Przepiórkowski, A. (2011). On the evaluation of two Polish taggers. In S. Goźdź-Roszkowski, editor, *Explorations across Languages and Corpora: PALC 2009*, pages 105–113, Frankfurt am Main. Peter Lang.

Marciniak, M., editor (2010). *Anotowany korpus dialogów telefonicznych*. Akademicka Oficyna Wydawnicza EXIT, Warsaw.

Mykowiecka, A. & Waszczuk, J. (2008). Semantic annotation of city transportation information dialogues using CRF method. In P. Sojka, A. Horák, I. Kopeček, and K. 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. Springer-Verlag.

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

Patejuk, A. & Przepiórkowski, A. (2010). ISOcat definition of the National Corpus of Polish tagset. In *LREC 2010 Workshop on LRT Standards*, Valletta, Malta. ELRA.

Piasecki, M. (2007). Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly*, **11**(1–2), 151–167.

Piasecki, M., Szpakowicz, S., & Broda, B. (2009). *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wroclawskiej, Wrocław.

Przepiórkowski, A., Górski, R. L., Łaziński, M., & Pęzik, P. (2010). Recent developments in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. ELRA.

Przepiórkowski, A., Bańko, M., Górski, R. L., & Lewandowska-Tomaszczyk, B., editors (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw. Forthcoming.

Pęzik, P. (2012). Język mówiony w NKJP. In Przepiórkowski et al. (2012). Forthcoming.

Saloni, Z., Gruszczyński, W., Woliński, M., & Wołosz, R. (2007). *Słownik gramatyczny języka polskiego*. Wiedza Powszechna, Warsaw.

Woliński, M. (2006). Morfeusz — a practical tool for the morphological analysis of Polish. In M. A. Kłopotek, S. T. Wierzchoń, and K. Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 503–512. Springer-Verlag, Berlin.

Woliński, M., Głowińska, K., & Świdziński, M. (2011). A preliminary version of składnica—a treebank of Polish. In Z. Vetulani, editor, *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 299–303, Poznań, Poland.

Woliński, M., Miłkowski, M., Ogrodniczuk, M., Przepiórkowski, A., & Łukasz Szałkiewicz (2012). PoliMorf: a (not so) new open morphological dictionary for Polish. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey. ELRA. These proceedings.