Abstract
The ATLAS project, started in March 2010, intends to create the multilingual language processing framework integrating the common set of linguistic tools for a group of European languages, among them Polish. The chained tools are producing multi-level UIMA-encoded annotation of texts which can be used by higher-level Web applications for complex language-intensive operations such as automated categorization, information extraction, machine translation or summarization.

This paper concentrates on properties of language of processing chains integrated for Polish and findings specific to this integration. Inflectional characteristics of Polish offers the possibility to present a few more advanced functions such as multiword unit lemmatisation, vital for real-life presentation of extracted phrases.

1. Introduction
The ATLAS project (Belogay et al., 2011; Ogrodniczuk and Karagiozov, 2011; Karagiozov et al., 2011) offers the possibility to test integration of several NLP tools for Polish together with other project languages — Bulgarian, English, German, Greek and Romanian. To demonstrate capabilities of the framework, three linguistically-aware online services have been built on top of it: i-Publisher (Web-based content management platform), i-Librarian (a digital library of scientific works) and EUDocLib (site for browsing and searching through EUR-LEX documents).

2. Polish Language Tools
A number of existing tools for processing Polish were used by the project. According to the agreed annotation model, supporting the target application of linguistic data, particularly their display to the user, the tools were in certain cases reconfigured and extended to provide additional information such as ready-to-display normalized versions of base forms of identified multiword noun phrases.

2.1. Morfeusz and Pantera
Morfeusz (Woliński, 2006) is a morphological analyzer for Polish, also used for sentence- and token-level segmentation and lemmatisation of texts before the morphological part is applied. It uses positional tags starting with POS information followed by values of morphosyntactic categories corresponding to the given part of speech (Przepiórkowski and Woliński, 2003). Current version of the tool, Morfeusz SGJP, is based on linguistic data coming from The Grammatical Dictionary of Polish (Saloni et al., 2007).

Pantera (Acedański, 2010; Acedański and Gołuchowski, 2009) is a recently developed morphosyntactic rule-based Brill tagger of Polish. It uses an optimized version of Brill’s algorithm adapted for specifics of inflectional languages. The tagging is performed in two steps, with a smaller set of morphosyntactic categories disambiguated in the first run (part of speech, case, person) and the remaining ones in the second run. Due to free word order nature of Polish the original set of rule templates as proposed by Brill has been extended to cover larger contexts.

2.2. Spejd and Multi-Word Expression Lemmatiser
Spejd (Przepiórkowski, 2008; Przepiórkowski and Buczyński, 2007) is an engine for shallow parsing using cascade grammars, able to co-operate with TaKIPI for tokenization, segmentation, lemmatisation and morphologic analysis. Parsing rules are defined using cascade regular grammars which match against orthographic forms or morphological interpretations of particular words. Spejd’s specification language is used, which supports a variety of actions to perform on the matching fragments: accepting and rejecting morphological interpretations, agreement of entire tags or particular grammatical categories, grouping (syntactic and semantic head may be specified independently). Users may provide custom rules or may use one of the provided sample rule sets.

Spejd is the basis for the nominal groups lemmatiser (Degórski, 2011) developed throughout the project which combines the lemmatisation task with shallow parsing. The parsing structures are used as bases for lemmatisation schemata written separately for each grammar rule and operating on the strings and structure matched by that rule.

2.3. Named Entity Recognizer
Polish NER tool (Savary et al., 2010; Waszczuk et al., 2010) is a statistical CRF-based named entity recognizer trained over 1-million manually annotated subcorpus of the National Corpus of Polish (Przepiórkowski et al., 2008) and
3. Towards Advanced Processing of Polish

The project targets at practical approach to numerous advanced linguistic issues such as coreference resolution, summarization, machine translation or categorization. Some of them have been recently tackled in other initiatives which can create useful synergies with ATLAS.

3.1. Coreference Resolution

The first attempts of coreference resolution for Polish are being currently carried out within the Computer-based methods for coreference resolution in Polish texts project financed by the Polish National Science Centre\(^1\) which intends to create a useful baseline for future experiments with this topic, but also to facilitate comparisons with general coreference resolvers planned for the project such as RARE (Postolache, 2004; Cristea et al., 29 31 May 2002).

The resulting implementation is designed to run either on true mention boundaries (Ogrodniczuk and Kopeć, 2011b) or in an end-to-end manner (Ogrodniczuk and Kopeć, 2011a). The current system uses a few rich rules, corresponding to syntactic constraints (elimination of nested nominal groups), syntactic filters (elimination of syntactic incompatible heads), semantic filters (wordnet-derived compatibility) and selection (weighted scoring).

3.2. Text Summarization

Despite intensive worldwide research in the field, just a few general summarization systems for Polish have been implemented so far. The first of them was PolSumm2 (Ciura et al., 2004), a modular, text extraction-based system monitoring inter-sentence relations, anaphors and ellipses. It was then complemented with Lackon (Dudczak, 2007; Dudczak et al., 2008b; Dudczak et al., 2008a; Dudczak et al., 2010), a heuristic summarizer used for testing various sentence selection methods and the extraction-based machine learning system of Joanna Świetlicka (Świetlicka, 2010). The most mature of these approaches can be adapted to ATLAS.

4. References


Adam Dudczak. 2007. Application of selected data exploration methods to summarization of Polish newspaper articles (Zastosowanie wybranych metod eksplozacji danych do tworzenia streszczeń tekstów prasowych dla języka polskiego, in Polish). MSc thesis.


---

\(^1\)Contract number 6505/B/T02/2011/40.


