LINGUISTIC PROCESSING CHAINS AS WEB SERVICES: INITIAL LINGUISTIC CONSIDERATIONS

Maciej Ogrodniczuk, Adam Przepiórkowski

Institute of Computer Science
Polish Academy of Sciences
ul. Ordona 21, Warsaw, Poland
maciej.ogrodniczuk@ipipan.waw.pl, adamp@ipipan.waw.pl

Abstract

At the end of 2009 the review of a number of available Web services implementing linguistic processing chains (CLARIN deliverable D5R-3a, 2009) was prepared as part of Common Language Resources and Technology Infrastructure (CLARIN) Working Group 5.6 (LRT integration) activities. Basing on the showcases contributed by WG members, the summary of features of both chained and individual Web Services was compiled, preparing the ground for comparisons between selected linguistic properties of registered frameworks. The article aims at presenting preliminary generalizations regarding functionalities, communication standards and representation of linguistic resources being adopted as web services, which were initially put forward in the CLARIN paper. The major features of the tools are summarized to provide starting point for discussion over interchange formats and tagsets, standards of encoding of linguistic resources and linguistic data categories. Apart from concentrating on representation of linguistic annotation, very preliminary conclusions concern technical, formal and semantic interoperability of language resources.

1. Introduction

Working Group 5.6 fulfils CLARIN mission of *creating, coordinating and making language resources and technology available and readily useable for scholars in the humanities and social sciences*[1] by concentrating on interoperability issues, mainly at the linguistic level (e.g., the problem of mapping between tagsets).

Within the Work Package 5 (Language Resources and Technologies Exploration) the group intended to provide the consortium with a broad overview of the LRTs available as web service chains and get an understanding of their status. This has been achieved by studying examples of the LRTs obtained as showcases from contributing partners (CLARIN consortium members) and compiled into initial summary of their status, properties, adopted standards and individual qualities.

2. Web service showcases

The call for contribution resulted in gathering descriptions of 8 frameworks, summarized according to the template delivered in the beginning of the process. On account of potential grave differences among submissions, the questions asked allowed some latitude in providing the general information on described solutions while remaining strict about their linguistic properties (languages covered, implemented NLP services, web service protocols, language resource standards and linguistic data encoding). The obtained materials were characterized by good quality and all partners showed advanced responsiveness while presenting and clarifying their solutions.

The next subsections attempt to summarize the showcases in a concise form, providing brief information on linguistic properties, performed functions, available web services (in form of WSDL[2] references, wherever available) and organizations involved in their preparation.

2.1. WebLicht

WebLicht (Web Based Linguistic Chaining Tool) is a SOA[3] framework of 25 web services performing specialized NLP[4] tasks for German, English, Italian, French and Finnish, such as sentence border detection, tokenization, POS[5] tagging, named entity recognition, lemmatization, constituent parsing, co-ocurrence annotation and semantic annotation. The open architecture allows for stacking existing services into processing chains as well as incorporating external tools and web services into existing solution.

The common representation of texts and annotations within the WebLicht processing chain is TCF (Text Corpus Format), an XML-based format supporting stand-off annotation and compatible with ISO LAF[6]. Converters for Negra[7], Paula (Dipper, 2005), MAF[8] and TüBa-D/Z[9] are available; the constituent parser output is TIGER-XML[10] (Mengel and Lezius, 2000), also TCF-encoded. Linguistic data is represented by means of language-dependent tagsets

[1]See the CLARIN Web page, `http://www.clarin.eu/`.

[2]Web Service Definition Language

[3]Service-Oriented Architecture

[4]Natural Language Processing

[5]Part-of-Speech

[6]Linguistic Annotation Framework, ISO/DIS 24612, see `http://www.tc37sc4.org/`.

[7]See `http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html`.

[8]Morpho-Syntactic Annotation Framework

[9]Tübinger Baumbank des Deutschen / Zeitungskorpus (Tübingen Treebank of Written German), see `http://www.sfs.uni-tuebingen.de/tuebadz.shtml`.

[10]See `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/`.

such as STTS[11] for German or the Penn Treebank tagset (UPenn)[12] for English.

WebLicht results from cooperation of linguistic departments of major German research institutions (Berlin Brandenburgische Akademie der Wissenschaften, University of Leipzig, University of Stuttgart and University of Tübingen).

2.2. GATE Web Services

GATE (General Architecture for Text Engineering) is open source software offering a wide range of language processing functionalities to be organized in maintainable workflows. Initially offered as plugins for the downloadable architecture, GATE subsystems are being gradually transformed into web services with information extraction (tokenizer, sentence splitter, POS tagger, named entity recogniser and classifier), phrase chunking, lemmatization and POS tagging tools leading the way.

Input data for the services may be encoded in a variety of text formats (plain text, HTML, SGML/XML, RTF/MS Word, PDF). The output is SynAF[13] (for noun/verb phrase chunker) and MAF-compliant XML (for lemmatizer and English/Bulgarian/Dutch POS taggers). Linguistic data are categorized by means of Penn Treebank tags.

GATE Web Services have been developed by the GATE group[14] at the University of Sheffield, UK.

2.3. IULA Web Services

The IULA Web Services family (Vivaldi Palatresi, 2009; Bel et al., 2006; Atserias et al., 2006; Villegas et al., 2009) allows for uploading and indexing text corpora to perform statistical queries (such as calculation of several lexicometric measures, word co-occurrences, relevance, distribution, extract and group concordances etc.) and various NLP tasks (e.g., tokenization, sentence splitting, morphological analysis, named entity detection and classification, POS tagging, chart-based shallow parsing, rule-based dependency parsing, nominal correference resolution or WordNet-based sense annotation and disambiguation), also in a chained manner. All services are available for English and Spanish, some of them (Freeling[16]) also for Catalan, Galician, Italian, Welsh, Portuguese and Asturian.

Input format for statistical processing is plain text while corpus analysis of annotated text requires EAGLES[17]/PAROLE[18] compliance. AAILE web service (Automatic Acquisition of Lexical Information by extracting syntactic patterns and contexts of concordances in a corpus) employs IULA tagsets for Spanish[19] and English[20].

The Web Services are maintained by Institut Universitari de Lingüística Aplicada at University Pompeu Fabra (IULA-UPF) in Barcelona, Spain.

2.4. ILSP Text Processing Chain

The main tools integrated by ILSP TPC are tokenizer and sentence splitter, POS tagger, lemmatizer, chunker and dependency parser.

All processing tools from the chain generate annotations compatible with UIMA annotation type system, an extension of JULIE Lab annotation scheme[21]. The services can also export results to other structured formats, e.g., GATE XML or XCES[22] (Ide et al., 2000). POS information is represented using PAROLE-compatible tagset, while dependency relations are described using Prague Dependency Treebank syntax.

The tools are provided by Institute for Language and Speech Processing (ILSP) from Athens, Greece. For more information see (Papageorgiou et al., 2002; Prokopidis and Georgantopoulos, 2010).

2.5. RACAI Services

The RACAI framework offers multiple linguistic tools for language identification (all EU languages), tokenization, tagging and lemmatization (TTL service, also containing remote procedures for sentence splitting and chunking), dependency parsing or wordnet browsing (remaining tools for Romanian and English).

Along with several proprietary formats, the tools encode results in XCES format. Lexical tagsets used is MULTEXT-EAST[23]-compliant (Erjavec, 2004; Tufiş, 2000).

The services are maintained by Research Institute for Artificial Intelligence, Romanian Academy of Sciences (RACAI), Bucharest, Romania.

2.6. WS-LexicalPlatform

The platform provides web service interface to the Italian SIMPLE lexicon, assisting in retrieving information concerning phonology, morphology, syntax and semantics.

The interchange data format is LMF[24] with ISO DCR[25]-mappable data categories basing on EAGLES-ISLE[26] (to be promoted to the future ISO standardization of data categories and, therefore, ISOCat).

---

[11] Stuttgart-Tübingen Tagset, see http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html.

[12] See http://www.cis.upenn.edu/~treebank/.

[13] Syntactic Annotation Framework, see http://www.tc37sc4.org/new_doc/ISO_TC37_4_N244_SynAF_WD_draft.pdf.

[14] See http://www.gate.ac.uk/.

[16] See http://www.lsi.upc.edu/~nlp/freeling/.

[17] Expert Advisory Group on Language Engineering Standards, see http://www.ilc.cnr.it/EAGLES96/home.html.

[18] See http://www.elda.org/catalogue/en/text/doc/parole.html.

[19] See http://www.iula.upf.edu/corpus/etqfrmes.htm.

[20] See http://www.iula.upf.edu/corpus/etquk.htm.

[21] See http://www.julielab.de/JULIE_Lab.html.

[22] XML Corpus Encoding Standard

[23] Multilingual Text Tools and Corpora for Central and Eastern European Languages, see http://nl.ijs.si/ME/.

[24] Lexical Markup Framework

[25] ISO 12620, Data Category Registry, see http://www.isocat.org/.

[26] International Standard for Language Engineering, see http://www.mpi.nl/ISLE/.

	Language identification	Sentence border detection	Tokenization	POS tagging / MSD[15]	Named Entity recognition	Lemmatization	Parsing	TreeBank browsing	Co-ocurrence annotation	Collocation extraction	Frequency analysis	Association measures	Semantic annotation	WordNet –related functionality	Thesaurus-related functionality	Lexicon access	Machine translation
WebLicht		•	•	•	•	•	•	•	•	•	•	•	•	•	•		
GATE		•	•	•	•	•							•				
IULA		•	•	•	•	•	•		•	•	•	•	•	•			
ILSP		•	•	•		•	•										
RACAI	•	•	•	•		•	•							•			•
WS-LexPl																•	
LXService		•	•	•													
WROCUT/ICS PAS		•	•	•		•	•		•	•	•	•		•			

Table 1: LRT functionality available in reviewed frameworks

The services are provided by Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale (CNR-ILC), Pisa, Italy.

2.7. LXService

The web service offers chunking, tokenization (Branco and Silva, 2003) and tagging (Branco and Silva, 2004; Silva, 2007) functionality for Portuguese. More tools, such as morphological analyser (Branco and Silva, 2006; Nunes, 2007; Martins, 2008) or parser (Silva et al., 2010) are being currently integrated. Proprietary formats are used both for encoding resources and linguistic data categories.

The body responsible for the services is University of Lisbon, Department of Informatics, Natural Language and Speech Group (NLX), Lisbon, Portugal. For more information see (Branco et al., 2008).

2.8. WROCUT/ICS PAS services

The tool set (language independent, although currently used with a grammar and tagset for Polish) comprise a tagger (Piasecki and Godlewski, 2006), a lemmatizer, tokenizer and morphologic analyser (Woliński, 2006), a shallow parser and disambiguation tool (Buczyński and Przepiórkowski, 2009), as well as an automatic harvester of lexical semantic relations from corpora for Polish and English (Broda and Piasecki, 2008; Piasecki et al., 2009).

Resources are represented is XCES and Wordnet-LMF (Aliprandi et al., 2009), while linguistic data is encoded using proprietary (currently de facto standard for Polish) ICS PAS tagset (Przepiórkowski and Woliński, 2003)[27] and CLAWS5 (British National Corpus tagset).

The services are the result of co-operation between Institute of Informatics, Wrocław University of Technology (WROCUT) and Institute of Computer Science, Polish Academy of Sciences (ICS PAS), Warsaw, Poland.

3. Summary of linguistic properties

3.1. NLP-specific functions

Table 1 presents the scope of LRT functionalities offered by the reviewed frameworks. The most complex web service-enabled processing chains seem to provide the widest linguistic coverage which obviously results from their background — due to increasing popularity of the remote service approach, existing tools are often being converted into web services. This tendency should be considered a good sign for small-size providers of linguistic material and services since their individual tools may effectively compete in the global network with their large-scale equivalents.

3.2. Encoding of linguistic resources

Table 2 presents the encoding formats of reviewed services. The first observation is that no common input/output format can be distinguished, neither any format is clearly standing out. The lowest common denominator for all reviewed formats seems to be XML — even the tools using text proprietary formats are, to some extent, XML-compatible or use XML as a variant representation (e.g., RACAI Services use internal Tab-separated SGML format along with XCES-encoded output).

Another dimension while evaluating formats is „standard or proprietary", with similar findings: proprietary formats tend to exist along with established standards or even gradually become standards, on local or multinational level.

[27] A slightly modified version of the tagset (Przepiórkowski, 2009) is used in the National Corpus of Polish (`http://nkjp.pl/`) and defined in ISOcat as a public data category set "NKJP" (cf. `http://www.isocat.org/interface/`).

	Acknowledged standards						Proprietary formats		
	XML-based formats								
	LMF-XML	LMF-WordNet	MAF	SynAF	TIGER-XML	XCES	XCES proprietary extension	XML proprietary format	Plain text proprietary format
WebLicht			•		•			•	
GATE			•	•				•	
IULA						•			
ILSP						•		•	
RACAI						•	•	•	•
WS-LexPl	•								
LXService								•	•
WROCUT/ICS PAS		•					•		

Table 2: Output formats of reviewed services

	Standard tagsets					Proprietary tagsets				
	CLAWS5	EAGLES/ PAROLE	MULTEXT-EAST	Prague Dependency Treebank	UPenn	ICS PAS (PL)	LX tagset (PT)	RACAI tagset (EN, RO)	SIMPLE-based tagset (IT)	STTS (DE)
WebLicht					•					•
GATE					•					
IULA		•								
ILSP		•		•						
RACAI			•							
WS-LexPl									•	
LXService							•			
WROCUT/ICS PAS	•					•				

Table 3: Tagsets used to encode linguistic annotation

WebLicht TCF is a good example here: being proprietary, it retains compatibility with ISO LAF/LMF/MAF standards.

In many cases proprietary extensions of recognized formats can supplement them with project-specific properties which makes the border between standard and non-standard even more vague. The need for compatibility is (and should be) in such cases satisfied by providing converters between internal and widely accepted formats (such as TCF-to-PAULA and MAF formats for WebLicht).

3.3. Linguistic data categories

Table 3 presents tagsets used by reviewed services for representing linguistic data categories. Similarly to the previous section, the border between standard and proprietary seems flexible. Some tagsets (such as STTS for German or ICS PAS tagset for Polish), while being non-standard, i.e., not recognized worldwide or approved by official standards development organization, are universally used for certain languages or constitute regional norms. Regardless of the process of emerging new standards-to-be, the tendency to normalize is noticeable since most frameworks tend to adopt well-known tagsets, either exclusively or along with their private formats.

4. Preliminary findings

Before making any generalizations it is worth to point out that neither the overview of text processing chains and web services in the LRT area, nor the initial findings were planned as an exhaustive summary, rather a study of usage scenarios including chains of operation.

Firstly, the presence of such a broad spectrum of different standards, both for encoding of linguistic resources and annotation categories, shows that the unification process is still in its beginnings. The reasons behind such condition do not seem to be the underestimation of the necessity of using widely-accepted standards by NLP community, but rather high costs of conversion of proprietary formats and preparation of mapping tools or, probably, the lack of linguistically mature interchange models. The role of such projects as CLARIN and FLaReNet[28] to create and endorse standards is therefore highly significant. In the long run, the concept of data conversion to impose formats and data categories loses the contest with a vision of ensuring compliance of current representation with some, preferably ISO-related, encoding standard. This scenario is universally adopted by most reviewed environments and remains compatible with CLARIN goals.

4.1. Interoperability issues

In general, interoperability of language resources can be discussed on three major levels: technical, syntactic and semantic. Technical interoperability, regarding e.g., web service protocols, is hardly of any concern here and has been adressed in (CLARIN deliverable D2R-6b, 2009). Formal interoperability, obtained by standardizing data exchange format and common language resource data model is already attainable with XML-based interchange formats following official representation standards. Semantic interoperability issue is still open, but appears to be solvable by providing formal mapping of proprietary categories to standard classes (such as those of ISOCat).

4.2. Linguistic standards

As stated above, the use of different representation standards is not discouraged and therefore the adoption of general metamodels seems the most appropriate solution for accommodating many encoding conventions. However, unambiguous unifying procedures (such as examples and best practices of how to convert, for instance, Penn Treebank-style representation into LAF) are necessary to ensure real interoperability between standards.

Practical assessment of methods and formats seems also necessary to strike a balance between permissiveness and constriction to enable accurate, yet flexible representation. Until then, a wider range of standards may be used to achieve better precision of linguistic description.

5. Closing notes

More and more linguistic processing chains are being available as web services and, however it will still be a long time before the new interfaces reach the quality of separate tools, the need of making their advanced functionalities available according to popular web service protocols is clearly visible and several renowned frameworks (such as the one of DFKI) are currently being amended with or ported to web service frameworks (as for DFKI, it is planned to be completed before the end of 2010).

The investigation of a growing network of linguistic tools available as services is therefore being continually underway, along with research and development in the closely related area of linguistic data interchange. As a result, the initial CLARIN document will be followed by an extended version containing final conclusions on the subject of harmonized access to resources via published interfaces to enable the interoperable domain. This deliverable will be available in the beginning of 2011.

6. References

Carlo Aliprandi, Federico Neri, Andrea Marchetti, Francesco Ronzano, Maurizio Tesconi, Claudia Soria, Monica Monachini, Piek Vossen, Wauter Bosma, Eneko Agirre, Xabier Artola, Arantza Diaz de Ilarraza, German Rigau, and Aitor Soroa. 2009. Database models and data formats. KYOTO Deliverable NR 1/WP NR 2, Version 3.1, 2009-01-31. *http://www2.let.vu.nl/twiki/pub/Kyoto/WP02:SystemDesignD2.1Database_Models_and_Data_Formats_v3.1.pdf*.

Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy. European Language Resources Association (ELRA).

Núria Bel, Sergio Espeja, and Montserrat Marimon. 2006. New tools for the encoding of lexical data extracted from corpus. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1362–1367, Genoa, Italy. European Language Resources Association (ELRA).

António Branco and Joao Silva. 2003. Contractions: breaking the tokenization-tagging circularity. *Lecture Notes in Artificial Intelligence 2721*, pages 167–170.

[28] *Fostering Language Resources Network*; see http://www.flarenet.eu/.

António Branco and Joao Silva. 2004. Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal. European Language Resources Association (ELRA).

António Branco and Joao Silva. 2006. Dedicated Nominal Featurization of Portuguese. *Lecture Notes in Artificial Intelligence 3960*.

António Branco, Francisco Costa, Pedro Martins, Filipe Nunes, Joao Silva, and Sara Silveira. 2008. LXService: Web Services of Language Technology for Portuguese. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Paris. European Language Resources Association (ELRA).

Bartosz Broda and Maciej Piasecki. 2008. SuperMatrix: a General Tool for Lexical Semantic Knowledge Acquisition. *Speech and Language Technology. Vol. 11*, pages 239–254.

Aleksander Buczyński and Adam Przepiórkowski. 2009. Spejd: A Shallow Processing and Morphological Disambiguation Tool. *Human Language Technology: Challenges of the Information Society. Vol. 5603*, pages 131–141.

CLARIN deliverable D2R-6b. 2009. Requirement Specification Web Services and Workflow Systems. `http://www-sk.let.uu.nl/u/D2R-6b.pdf`.

CLARIN deliverable D5R-3a. 2009. Linguistic processing chains as Web Services: Initial linguistic considerations. `http://www-sk.let.uu.nl/u/D5R-3a.pdf`.

Stefanie Dipper. 2005. Stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*, Berlin.

Tomaz Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 233–236, Genoa, Italy. European Language Resources Association (ELRA).

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2009. Multilingual Resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation Journal. Vol. 43:1*, pages 57–70.

Karypis George. 2002. CLUTO — a clustering toolkit. Technical Report Technical Report 02-017, Department of Computer Science, University of Minnesota.

Nancy Ide, Patrice Bonhomme, and Laurent Romary. 2000. XCES: An XML-based standard for linguistic corpora. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 825—-830, Athens, Greece. European Language Resources Association (ELRA).

Violetta Koseska-Toszewa, Ludmila Dimitrova, and Roman Roszko, editors. 2009. *Representing Semantics in Digital Lexicography: Proceedings of MONDILEX Fourth Open Workshop*, Warsaw.

Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors. 2006. *Intelligent Information Processing and Web Mining*. Advances in Soft Computing. Springer-Verlag, Berlin.

Alessandro Lenci, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. 2000. SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography XIII (4)*, pages 249–263.

Dekang Lin. 1993. Principle-based parsing without overgeneration. In *Proceedings of the 31st Meeting of the ACL*, pages 112–120.

Pedro Martins. 2008. Desambiguaçao Automática da Flexao Verbal em Contexto. Master's thesis, University of Lisbon.

Andreas Mengel and Wolfgang Lezius. 2000. An XML-based encoding format for syntactically annotated corpora. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 121–126, Athens, Greece. European Language Resources Association (ELRA).

Filipe Nunes. 2007. Verbal Lemmatization and Featurization of Portuguese with Ambiguity Resolution in Context. Master's thesis, University of Lisbon.

Harris Papageorgiou, Prokopis Prokopidis, Iason Demiros, Voula Giouli, Alexis Konstantinidis, and Stelios Piperidis. 2002. Multi-level XML-based Corpus Annotation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain. European Language Resources Association (ELRA).

Maciej Piasecki and Grzegorz Godlewski. 2006. Reductionistic, tree and rule based tagger for Polish. In Kłopotek et al. (Kłopotek et al., 2006), pages 531–540.

Maciej Piasecki and Adam Radziszewski. 2009. Morphosyntactic Constraints in Acquisition of Linguistic Knowledge for Polish. Aspects of Natural Language Processing (a festschrift for Professor Leonard Bolc). *Lecture Notes in Computer Science, 5070*, pages 163–190.

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

Prokopis Prokopidis and Byron Georgantopoulos. 2010. Extending a Text Processing Pipeline for Greek. Submitted in LREC 2010.

Adam Przepiórkowski and Marcin Woliński. 2003. The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. In *Proceedings of the*

4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003, pages 109–116.

Adam Przepiórkowski. 2004. *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Adam Przepiórkowski. 2009. A comparison of two morphosyntactic tagsets of Polish. In Koseska-Toszewa et al. (Koseska-Toszewa et al., 2009), pages 138–144.

Nilda Ruimy, Monica Monachini, Raffaella Distante, Elisabetta Guazzini, Stefano Molino, Marisa Ulivieri, Nicoletta Calzolari, and Antonio Zampolli. 2002. Clips, a multi-level Italian computational lexicon: A glimpse to data. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain. European Language Resources Association (ELRA).

Joao Silva, António Branco, Sérgio Castro, and Ruben Reis. 2010. Out-of-the-box Robust Parsing for Portuguese. In *Proceedings of the 9th International Conference on the Computational Processing of Portuguese (PROPOR 2010)*, Porto Alegre. Pontifícia Universidade do Rio Grande do Sul.

Joao Silva. 2007. Shallow Processing of Portuguese: From Sentence Chunking to Nominal Lemmatization. Master's thesis, University of Lisbon.

Antonio Toral and Monica Monachini. 2007. SIMPLE-OWL: a Generative Lexicon Ontology for NLP and the Semantic Web. In *Workshop on Cooperative Construction of Linguistic Knowledge Bases (AIIA 2007)*.

Dan Tufiş. 2000. Using a Large Set of EAGLES-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 1105–1112, Athens, Greece. European Language Resources Association (ELRA).

Marta Villegas, Núria Bel, Santiago Bel, Francesca Alemany, and Hector Martínez. 2009. Lexicography in the grid environment. In *Proceedings of e-lex 2009*, Louvain: Cahiers du Cental.

Jorge Vivaldi Palatresi. 2009. Corpus and exploitation tool: IULACT and bwanaNet. A survey on corpus-based research = Panorama de investigaciones basadas en corpus. In Pascual Cantos Gómez and Aquilino Sánchez Pérez, editors, *Actas del I Congreso Internacional de Lingüística de Corpus (CICL-09)*, pages 224–239. Universidad de Murcia, Asociación Espanola de Lingüística del Corpus.

Marcin Woliński. 2006. Morfeusz — a practical tool for the morphological analysis of Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Proceedings of the International IIS: IIPWM'06 Conference*, pages 511–520, Wisła, Poland.

Sue Ellen Wright. 2004. A global data category registry for interoperable language resources. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal. European Language Resources Association (ELRA).