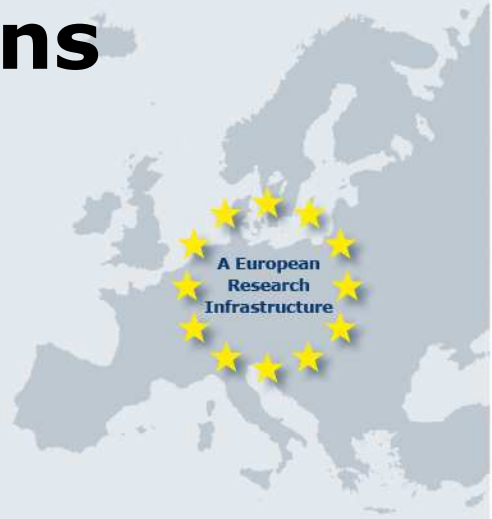
Linguistic processing chains as Web Services: Initial linguistic considerations



2010-01-18 Version: 1.0.1

Editors: Maciej Ogrodniczuk, Adam Przepiórkowski

The ultimate objective of CLARIN is to create a European federation of existing digital repositories that include language-based data, to provide uniform access to the data, wherever it is, and to provide existing language and speech technology tools as web services to retrieve, manipulate, enhance, explore and exploit the data. The primary target audience is researchers in the humanities and social sciences and the aim is to cover all languages relevant for the user community. The objective of the current CLARIN Preparatory Phase Project (2008-2010) is to lay the technical, linguistic and organisational foundations, to provide and validate specifications for all aspects of the infrastructure (including standards, usage, IPR) and to secure sustainable support from the funding bodies in the (now 23) participating countries for the subsequent construction and exploitation phases beyond 2010.

Linguistic processing chains as Web Services: Initial linguistic considerations

CLARIN-2009-D5R-3a

EC FP7 project no. 212230

Deliverable: D5R-3a – Deadline: 31.12.2009

Responsible: Adam Przepiórkowski

Contributing Partners: ICS PAS, ILC, ILSP, RACAI, ULisbon, UPF, UTübingen, WROCUT
Contributing Members: BBAW, ULeipzig, UStuttgart

Scope of the document

This document concentrates on the review of a number of web services implementing linguistic processing chains and the specification of linguistic requirements on web services.

This document will be discussed in the appropriate working groups and in the Executive Board. It will be subject of regular adaptations dependent on the progress in CLARIN.

CLARIN references

• Language Resource and Technology Federation	CLARIN-2008-4	February 2009
• Metadata Infrastructure for Language Resource and Technology	CLARIN-2008-5	February 2009
• Report on Web Services	CLARIN-2008-6	March 2009
• Requirement Specification Web Services and Workflow systems	CLARIN-2009-1	June 2009

Frequently used acronyms

Abbreviation	Explanation
LRT	Language Resources and Technologies
MSD	Morphosyntactic Description
NE	Named Entity
NER	Named Entity Recognition
NLP	Natural Language Processing
WSDL	Web Service Definition Language

Contents

Introduction

This document focuses on obtaining representative examples of the LRTs (Language Resources and Technologies) available as web services and getting an understanding about their status. Exploration results presented below will facilitate selection of appropriate standards for the resources and tools to be integrated in the course of further CLARIN activities.

Chapter 1 describes several web service-based processing chains and individual services in the form of showcases delivered by consortium members. Each framework is presented in a standardized manner, starting with some general background, availability, authorities responsible for preparation and running the web service infrastructure, status of the tools and list of supported languages. Each implemented service is then presented, showing their individual qualities and providing WSDL (Web Service Definition Language) references whenever possible. Web service protocols used by the reviewed tools are showed in narrow scope, in contrast to language resource standards and linguistic data encoding information, described in greater detail. Encoding examples are also shown frequently.

Chapter 2 makes an attempt to draw comparisons between selected properties of registered tools. NLP (Natural Language Processing) functionalities offered by reviewed frameworks are summarized briefly, followed by an overall analysis of the language resource and linguistic data encoding standards within reviewed environments.

Chapter 3 concentrates on reaching preliminary generalizations which might facilitate drawing conclusions and lead towards future recommendations. Preliminary findings concerning requirements for the registries of linguistic resources and tools for the representational standards for the various types of resources are also included.

This document will be followed by another deliverable of WP5R task R3 (Integration of LR into web service infrastructure) containing final conclusions on the subject of harmonized access to resources via published interfaces to enable the interoperable domain. This final deliverable is planned for the end of the third year of the project.

1 Reviewed processing chains and individual web services

1.1 WebLicht

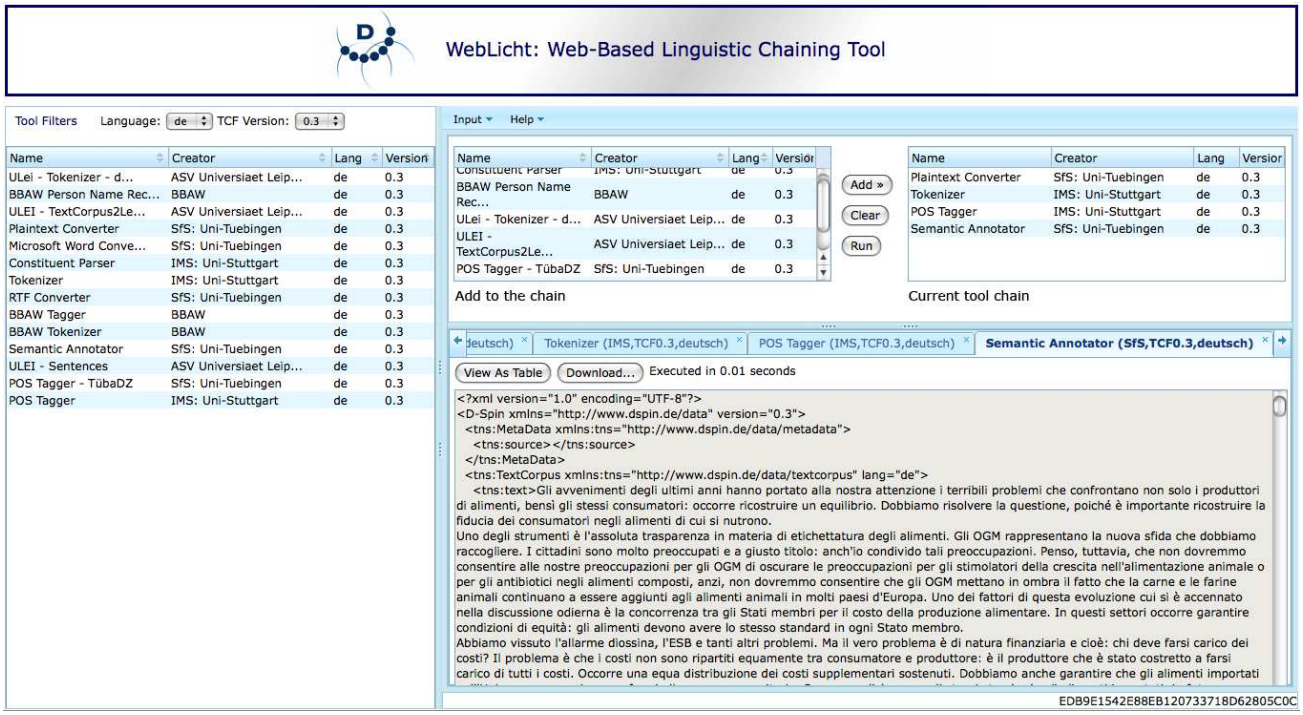
General information

Name of the service/project

WebLicht: Web Based Linguistic Chaining Tool

General URL

http://weblicht.sfs.uni-tuebingen.de:8080/WebLicht1/



Availability

Due to copyright issues, WebLicht is password protected. An overview of WebLicht can be found at http://www.d-spin.org.

Bodies responsible

Seminar für Sprachwissenschaft, Universität Tübingen (UTübingen), Germany
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart (UStuttgart), Germany
Abteilung Automatische Sprachverarbeitung, Universität Leipzig (ULeipzig), Germany
Berlin-Brandenburgische Akademie der Wissenschaften (BBAW), Berlin, Germany

Status of the tools

Stable prototype

Supported languages

German, English, Italian, French, Finnish, several more in preparation.

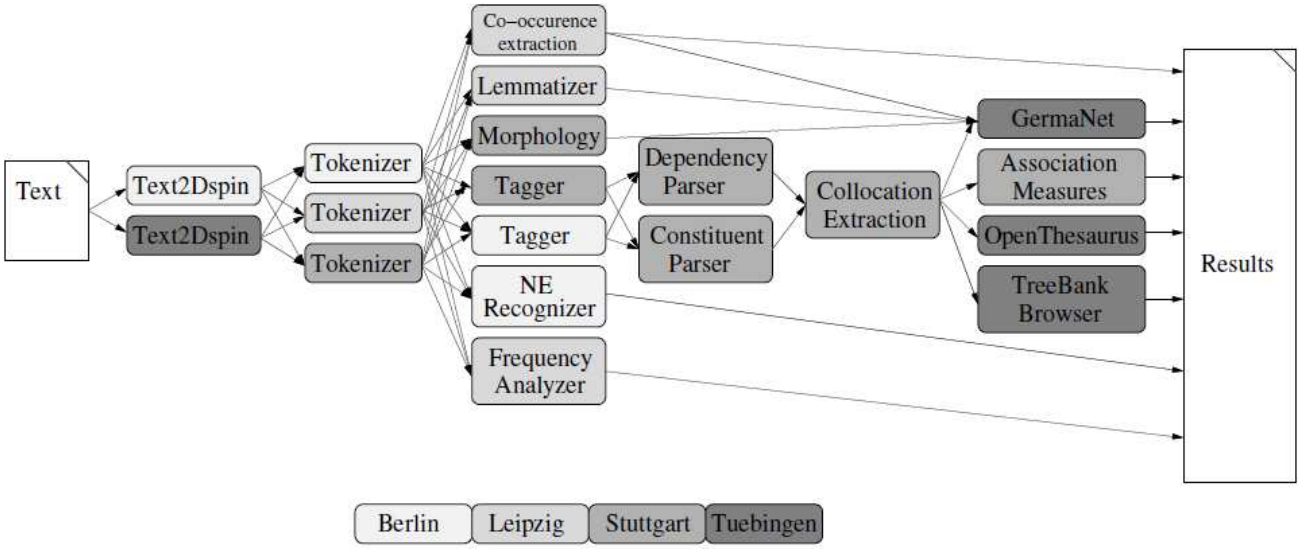
Implemented NLP services

Alltogether, ca. 25 webservices are available:

- several tokenizers,
- detection of sentence borders,
- several part-of-speech taggers,
- named entity recognition,
- lemmatization,
- constituent parsing,
- coocurrence annotation,

- semantic annotator (GermanNet),
- several data format converters (including MS Word/PDF/RTF to plain text and plain text to internal XML).

Other types of tools can be easily integrated.



Web service protocols

REST

Encoding of linguistic resources

All WebLicht web services output files in TCF (Text Corpus Format). This format is highly compatible with other standards. Converters are already available for:

- Negra,
- PAULA (Potsdamer AUstauschformat für Linguistische Annotation; Dipper 2005),
- MAF,
- TüBa/DZ.

The BitPar constituent parser produce TIGER-XML-style analyses (Mengel and Lezius 2000, http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TigerXML.html), but they are also encoded in TCF.

Linguistic data categories

For POS tagging, language-dependent tagsets are used, e.g. STTS (http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html) for German, the Penn Treebank tagset (UPenn) for English, etc.

1.2 GATE Web Services

General information

Name of the service/project

GATE (General Architecture for Text Engineering, see http://gate.ac.uk/science.html for main features) is a widely-used infrastructure for language processing software development. Altough its tools are being developed as plug-ins for the downloadable architecture, an increasing number of GATE tools are now being converted into web services.

Availability

In preparation.

Body responsible

GATE group (http://www.gate.ac.uk), Department of Computer Science, University of Sheffield, UK

Status of the tools

In preparation.

Supported languages

- English – ANNIE, GATE Noun/Verb Phrase Chunker, GATE Lemmatizer, GATE English POS tagger,
- Bulgarian – GATE Bulgarian POS tagger,
- Dutch – GATE Dutch POS tagger.

Implemented NLP services

ANNIE

ANNIE is an open-source, robust Information Extraction (IE) system. Its output relies on finite state algorithms. ANNIE consists of the following main language processing tools: tokeniser, sentence splitter, POS tagger, named entity recogniser and classifier.

The named entity recogniser identifies and categorizes entity names (such as persons, organizations, and location names), temporal expressions (dates and times), and certain types of numerical expressions (monetary values and percentages). For this purpose, it uses three types of processing resources: a gazetteer, a part of speech tagger and a rule grammar module. The gazetteer consists of lists such as cities, organizations, days of the week, etc. It not only consists of entities, but also of names of useful indicators, such as typical company designators (e.g. 'Ltd.'), titles, etc. The gazetteer lists are compiled into finite state machines, which can match text tokens. The part of speech tagger attaches morpho-syntactic labels ("noun", "verb", "adjective" etc.) to text elements. The rule grammar component allows the encoding of rules that operate on the output of both the gazetteer and the pos tagger in order to annotate text spans with the relevant named entity types. The text spans and annotations are exported into an RDF ontology, in which the named entity types such as Organization and Person constitute classes, and the text spans instances of these classes.

GATE Noun/Verb Phrase Chunker

The chunker produces text annotated at phrase level in XML format.

For producing this annotation output it depends on the linguistic preprocessing of the text input (for required text format see below) with domain- and application-independent techniques.

- Tokenization: the tokeniser splits text into simple tokens, such as numbers, punctuation, symbols, and words of different types (e.g. with an initial capital, all upper case, etc.).
- The sentence splitter segments the text into sentences.
- The part-of-speech tagger adds morphosyntactic information to each token.

GATE Lemmatizer

The lemmatizer produces text annotated with lemma information for nouns and verbs in XML format.

For producing this annotation output it depends on the same linguistic preprocessing as described above.

GATE POS taggers

The taggers produce a part-of-speech tag as an annotation on each word or symbol.

Producing annotation output again depends on linguistic preprocessing of the text input.

Web service protocols

SOAP

Encoding of linguistic resources

For all GATE web services the input texts may be encoded in several formats: plain text, HTML, SGML, XML, RTF, PDF (not all), Microsoft Word (not all); no language resource standards are required for input.

The output is always in the form of XML annotated text. Compliance with standard representations is the following:

- GATE Noun/Verb Phrase Chunker – the output is SynAF compliant.
- GATE Lemmatizer and GATE POS taggers – the output is MAF compliant.

Linguistic data categories

The POS tags are Penn Treebank compliant.

1.3 IULA Web Services

Further information and the full description of IULA Web Services can be obtained at http://gilmere.upf.edu/WS/.

General information

Name of the service/project

- Statistical Web Services (statistics and corpus analysis of raw text),
- CQP (corpus analysis of annotated text),

- AAILE WS (Automatic Acquisition of Lexical Information by extracting syntactic patterns and contexts of concordances in a corpus),
- Freeling WS (deployment of the Freeling package of language analysis services as WS),
- Upload web services (uploading of corpora to a server),
- XSLT Transformer WS (transformation of XML content using XSLT).

Availability

Publicly available at http://gilmere.upf.edu/WS/.

Body responsible

Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra (IULA-UPF), Barcelona, Spain

Status of the tools

Stable prototypes

Supported languages

- Statistical Web Services, CQP, Upload, XSLT Transformer – all (language independent),
- AAILE WS – English and Spanish,
- Freeling WS – Spanish, Catalan, Galician, Italian, English, Welsh, Portuguese and Asturian.

Implemented NLP services

Statistical Web Services

The IULA Statistical Web Services (WS) family performs statistical tasks on a specific corpus. This corpus must be provided by the user by means of the Upload Web Service. Once a corpus is uploaded, it is assigned a unique and persistent identifier. This corpus identifier is used by each statistical task WS.

The following functions are provided by the currently available web services:

- DescribeCorpus WS – used to calculate some lexicometric measures in a corpus,
- DescribeCorpusByLength WS – used to calculate some lexicometric measures in a corpus,
- kwic WS – used to extract concordances,
- Ngrams WS – used to calculate word co-occurrences,
- TfIdf WS – used to calculate word relevance,
- Distribution WS – used to calculate word distribution.

Further information: http://gilmere.upf.edu/WS/#Statistical%20Web%20Services

WS Access: http://gilmere.upf.edu/WS/statistical/v1/invoke

WSDL: http://gilmere.upf.edu/WS/statistical/v1/wsdl

CQP

These web services allow (a) querying the IULA's technical corpora and (b) indexing and eventually querying new corpora. Additionally, some chains can also be used here, as the BagOfWords WS takes as input the results of the CQP WS and its output is taken as input for the BagOfWordsClustering WS.

The CQP WS family consists of:

- CQP WS – offers a way to query the IULA technical corpora available at http://bwananet.iula.upf.edu; the web service takes a CQP query expression and a reference corpus as input and returns an XML file with the occurrences.

  More information about IULA's Technical Corpora: (Vivaldi J. 2009).

  WS Access: http://gilmere.upf.edu/WS/cqp/v4/invoke.

  WSDL: http://gilmere.upf.edu/WS/cqp/v4/wsdl.
- corpus_resources WS – indexes a given corpus in CQP format,
- queries WS – offers a way to query an indexed corpus,
- CQP BagOfWords WS – makes a CQP WS query and returns a data matrix that collects the words that go with the lemma.

  More information: (Villegas et al. forth).

  WS Access: http://gilmere.upf.edu/WS/bag_of_words/v3/invoke.

  WSDL: http://gilmere.upf.edu/WS/bag_of_words/v3/wsdl.
- CQP BagOfWordsClustering WS – performs clustering on the matrix returned by CQP BagOfWords.

  WS Access: http://gilmere.upf.edu/WS/bag_of_words_clustering/v2/invoke.

  WSDL: http://gilmere.upf.edu/WS/bag_of_words_clustering/v2/wsdl.

AAILE

This service groups concordances according to the syntactic contexts the key word occurs in. Given a lemma, a corpus and a predefined set of admissible syntactic contexts (expressed in terms of regular expressions), the system (i) looks for all the occurrences of the lemma in the corpus, (ii) constructs the corresponding vectors (taking into account the set of regular expressions) and (iii) groups the vectors. The idea is that, when looking for occurrences in corpus, lexicographers get an organized set of examples. The system is restricted to nouns and adjectives (Bel et al. 2006).

Futher information: http://gilmere.upf.edu/WS/#AILE.

Version 1:

- WS Access: http://gilmere.upf.edu/WS/aaile/aaile/invoke.
- WSDL: http://gilmere.upf.edu/WS/aaile/aaile/wsdl.

Version 2:

- WS Access: http://gilmere.upf.edu/WS/aaile/v2/invoke
- WSDL: http://gilmere.upf.edu/WS/aaile/v2/wsdl.

Freeling

All Freeling applications have been deployed as web services using the REST protocol. The potential of Freeling and its main features are described at http://www.lsi.upc.edu/~nlp/freeling/ and there is also a demo available at http://garraf.epsevg.upc.es/freeling/demo.php (Atserias *et al.* 2006).

The main services offered by the Freeling library include:

- Text tokenization,
- Sentence splitting,
- Morphological analysis,
- Suffix treatment, retokenization of clitic pronouns,

- Flexible multiword recognition,
- Contraction splitting,
- Probabilistic prediction of unkown word categories,
- Named entity detection,
- Recognition of dates, numbers, ratios, currency, and physical magnitudes (speed, weight, temperature, density, etc.),
- PoS tagging,
- Chart-based shallow parsing,
- Named entity classification,
- WordNet based sense annotation and disambiguation,
- Rule-based dependency parsing,
- Nominal correference resolution.

For further references on the different Freeling applications you can also refer to: http://www.lsi.upc.edu/~nlp/freeling/index.php?option=com_content&task=view&id=20&Itemid=49

Upload/IsUploaded package

Upload WS makes it possible to upload corpora to the server. Use Upload WS when uploading a single file, and Uploadzip when uploading a zip corpus. Corpus uploading is asynchronous. Thus, upload services return a ticket number that identifies the uploaded corpus. This ticket is stored in a dB at the server side. When the uploading is finished, the ticket is marked as 'available'. IsUpload and IsUploadzip services are used to check the status of uploaded corpora. They return 'true' when the uploading is finished and 'false' otherwise. When a ticket number is available, it can be used as a corpus ID (http://gilmere.upf.edu/WS/#File%20upload).

- Upload:

  WS Access: http://gilmere.upf.edu/WS/jaguar/v2/invoke_method_params?method=Upload&service=v2.

  WSDL: http://gilmere.upf.edu/WS/jaguar/v2/wsdl.
- Uploadzip:

  WS Access: http://gilmere.upf.edu/WS/jaguar/v2/invoke_method_params?method=UploadZip&service=v2

  WSDL: http://gilmere.upf.edu/WS/jaguar/v2/wsdl.
- IsUploaded:

  WS Access: http://gilmere.upf.edu/WS/jaguar/v2/invoke_method_params?method=IsUploaded&service=v2.

  WSDL: http://gilmere.upf.edu/WS/jaguar/v2/wsdl.
- IsUploadedzip:

  WS Access: http://gilmere.upf.edu/WS/jaguar/v2/invoke_method_params?method=IsUploadedZip&service=v2

  WSDL: http://gilmere.upf.edu/WS/jaguar/v2/wsdl.

Upload WS in REST

REST web service for uploading files accessible by other web services. It has a web interface that can be used at the same URL (http://gilmere.upf.edu/WS/upload).

XSLT Transformer

A REST web service that given an XML content and XSL content, performs the XSL transformation (http://gilmere.upf.edu/WS/xslt/v1/transformation).

Web service protocols

SOAP

- IULA Statistical Web Services,
- CQP WS,
- CQP BagOfWords WS,
- CQP BagOfWordsClustering WS,
- AAILE,
- Upload/IsUploaded package.

REST

- CQP corpus_resources WS,
- CQP queries WS,
- Freeling web services,
- Upload WS in REST,
- XSLT Transformer.

Encoding of linguistic resources

- Statistical Web Services: not applicable – input is raw text.
- CQP (corpus analysis of annotated text): EAGLES.
- AAILE WS (Automatic Acquisition of Lexical Information by extracting syntactic patterns and contexts of concordances in a corpus): EAGLES.
- Freeling WS (deployment of the Freeling package of language analysis services as WS): EAGLES / PAROLE.
- Upload web services (uploading of corpora to a server): not applicable.
- XSLT Transformer WS – transformation of XML content using XSLT.

Linguistic data categories

- Statistical Web Services, CQP, Upload/IsUploaded package and XSLT Transformer – not applicable, as they do not perform any annotation.
- CQP – queries have no restriction on the linguistic data encoding.
- AAILE – IULA tagsets for Spanish (http://www.iula.upf.edu/corpus/etqfrmes.htm) and English (http://www.iula.upf.edu/corpus/etquk.htm).
- Freeling – EAGLES / PAROLE.

1.4 ILSP Text Processing Chain

General information

Name of the service/project

ILSP Text Processing Chain (ILSP TPC)

Availability

Restricted – interested parties should contact ILSP.

Body responsible

Institute for Language and Speech Processing (ILSP), Athens, Greece

Status of the tools

Stable, have been integrated in the framework of many national and European projects.

Reimplementation of the basic TPC tools as Java components based on the Apache UIMA framework (http://incubator.apache.org/uima/) has recently been completed, and is discussed in this document. The tools have been developed and/or trained and evaluated on a pool of annotated resources compiled at ILSP and focus on the Modern Greek language. An overview of an earlier version of the chain of tools is provided in (Papageorgiou et al. 2002) while an update will be included in (Prokopidis & Georgantopoulos, submitted).

Supported languages

Greek

Implemented NLP services (selection)

ILSP Tokenizer and Sentence Splitter

This tool tokenizes and a sentence splitter identifies word and sentence boundaries, on the basis of the ICU4J (http://site.icu-project.org/) RuleBasedBreakIterator, a set of post-processing heuristics and gazetteers of abbreviations.

ILSP FBT POS Tagger

A part-of-speech transformation-based tagger has been trained on a corpus of 455K words. The tagger assigns initial tags by following simple heuristics and looking up words in a precompiled lexicon. For unknown words, lexicons of suffix-tag combinations are used. A set of contextual rules learned from the training corpus are then applied to improve word and suffix lexicons output. As an alternative to the ILSP FBT Tagger, an open source decision tree tagger is used (http://www.ims.uni-stuttgart.de/projekte/corplex/RFTagger/) trained on the same corpus, getting similar evaluation results.

ILSP Lemmatizer

Following POS tagging, a lexicon-based lemmatizer retrieves lemmas from ILSP's Greek Morphological Lexicon. This resource contains 66K lemmas, which in their expanded form extend the lexicon to approximately 2M different entries.

ILSP Chunker

ILSP Chunker is a tool that recognizes non-recursive chunks and clauses. Its main resource is a grammar consisting of non-recursive regular expressions that has been compiled into a cascade of finite state transducers.

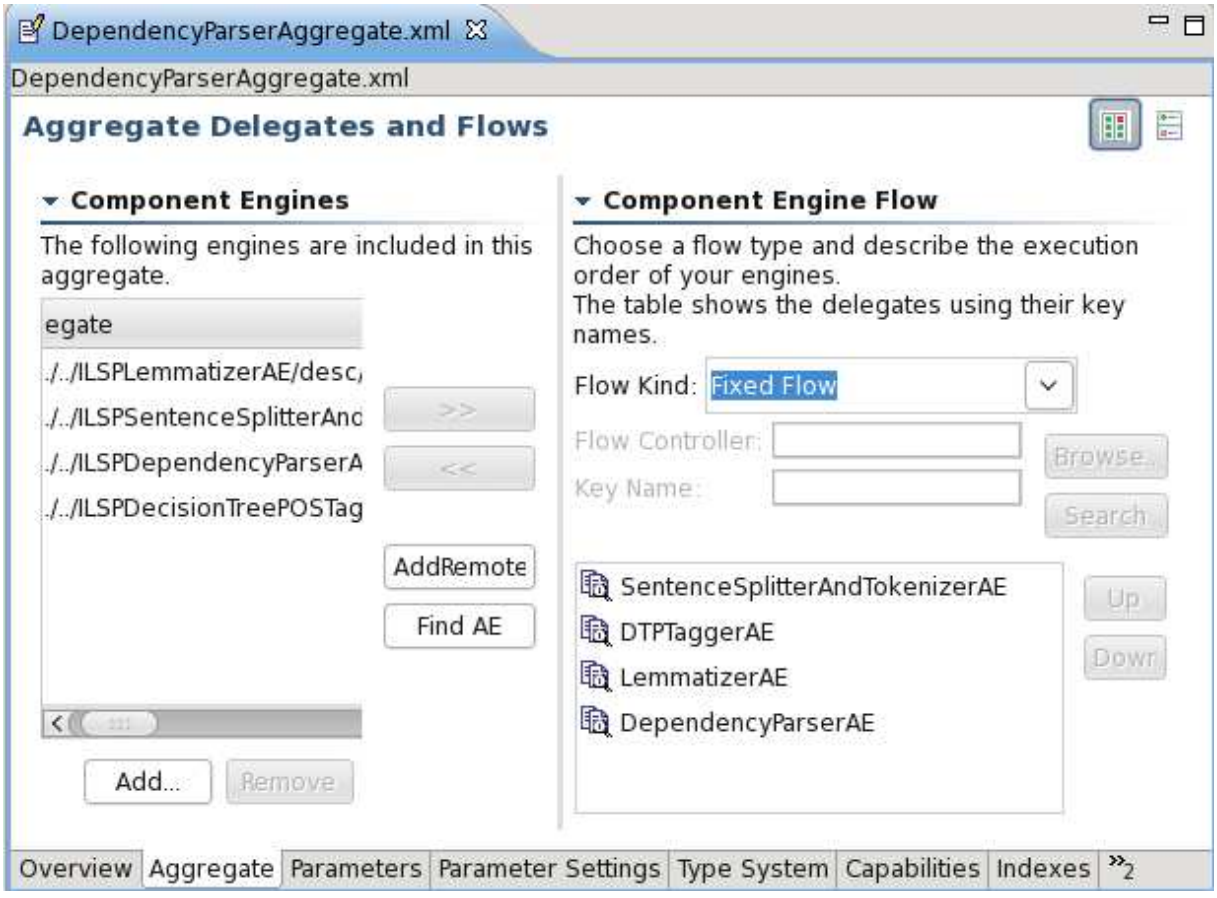
ILSP Dependency Parser

For parsing, open source dependency parsers (http://www.maltparser.org, http://sourceforge.net/projects/mstparser/) have been trained on the Greek Dependency Treebank, a manually annotated resource comprising ~70K words of news documents and European parliament sessions.

Web service protocols

The UIMA components are being made available via UIMA Asynchronous Scaleout (AS) services.

On the server side, the UIMA AS framework includes capabilities that allow wrapping and management of either primitive components, or components aggregated in workflows via specific descriptors. As an example, editing such a descriptor to generate a workflow for dependency parsing is shown in the figure below. Instances of the component can be scaled out by being deployed in different hosts. Thus, more than one input queue can be processed in parallel.



An aggregate AE for dependency parsing

Encoding of linguistic resources

No widely recognized standard is currently used for all annotated resources. However, all resources have been converted in XML files that include original text, document metadata and stand-off annotation. All processing tools mentioned above generate annotations compatible with a UIMA multi-layered annotation type system, which is an extension of the one provided by the JULIE Lab (http://www.julielab.de/JULIE_Lab.html).

The UIMA services described above can export results to editor-specific formats, like, for example, the ones used for dependency tree annotation in the Tree Editor tool (http://ufal.mff.cuni.cz/~pajas/tred/), or for annotation editing in the GATE environment (http://gate.ac.uk/). Results can also be optionally exported to XCES (Ide *et al.* 2000) compatible formats as in the following example:

```
<s id="seg.EL.1">
  <tok id="tok_1_1">
    <orth>Η</orth>
    <base>ο</base>
    <ctag>AtDfFeSgNm</ctag>
    <msd>Tdfsn</msd>
  </tok>
  <tok id="tok_1_2">
    <orth>οικιστική</orth>
    <base>οικιστικός</base>
    <ctag>AjBaFeSgNm</ctag>
    <msd>A_pfs__n</msd>
  </tok>
  ...
</s>
```

Linguistic data categories

Linguistic information is encoded using data categories that are easily mappable to similar encodings for the majority of widely-spoken European languages. As an example, the POS taggers assume a PAROLE-compatible tagset of 584 tags.

The table below presents briefly basic POS tags together with their subcategorizations (without mentioning sub-features regarding case, aspect, gender, etc.):

POS	Description
Ad	Adverb
Aj	Adjective
AsPpPa	Preposition + Article combination
AsPpSp	Simple preposition
AtDf	Definite article
AtId	Indefinite article
CjCo	Coordinating conjunction

POS	Description
OPUNCT	Opening punctuation
PnDm	Demonstrative pronoun
PnId	Indefinite pronoun
PnIr	Interrogative pronoun
PnPe	Personal pronoun
PnPo	Possessive pronoun
PnRe	Relative pronoun

POS	Description
CjSb	Subordinating conjunction
COMP	A composite word form
CPUNCT	Closing punctuation
DATE	Date
DIG	Digit
ENUM	Enumeration element
INIT	Initial
NmCd	Cardinal numeral
NmCt	Collective numeral
NmMl	Multiplicative numeral
NmOd	Ordinal numeral
NoCm	Common noun
NoPr	Proper noun

POS	Description
PnRi	Relative indefinite pronoun
PTERM	Terminal punctuation
PtFu	Future particle
PtNg	Negative particle
PtOt	Other article
PtSj	Subjunctive particle
PUNCT	Other punctuation
RgAbXx	Abbreviation
RgAnXx	Acronym
RgFwOr	Foreign word in its original form
RgFwTr	Transliterated foreign word
VbIs	Impersonal verb
VbMn	Main verb

The table below presents the set of dependency relations used. It is based on the one used in the Prague Dependency Treebank:

Dependency relation	Description
Pred	Main sentence predicate
Sb	Subject
Obj	Direct object
IObj	Indirect object
Pnom	Predicative dependent
Adv	Adverbial dependent
Atv	Adverbial predicative dependent
Atr	Attribute
AuxP	Prepositional node
AuxC	Conjunction node

Dependency relation	Description
Coord	A node governing coordination
Apos	A node governing apposition
*_Co	A node governed by a Coord
*_Ap	A node governed by an Apos
*_Pa	Head node of a parenthetical structure
AuxX	Comma
AuxV	Auxiliary node attached to a verb
AuxK	Terminal punctuation
AuxG	Auxiliary punctuation
ExD	A node whose real parent node is not present in the sentence (ellipsis)

1.5 RACAI Services

General information

Name of the service/project

RACAI Services

Availability

Public.

Body responsible

Research Institute for Artificial Intelligence, Romanian Academy of Sciences (RACAI), Bucharest, Romania

Status of the tools

Stable.

Supported languages

- Language Identification – Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovakian, Slovene, Spanish, Swedish and rare languages (Aweti, Teop),
- TTL, LexPar, TextProcessing, WordNetBrowser – English and Romanian.

Implemented NLP services

Language Identification

Language Identification service ensures automatically identification of the language of a text written in one of the 22 European Union languages. The text may contain a minimal number of 10-15 words (roughly a sentence).

The implementation of the language identification operation involves creating stochastic models of affixes for different languages. When a new text is input, the algorithm computes the probabilities of the affixes and compares them to those computed in the training phase. The language whose model best matches these probabilities is the language of the input text.

Language Identification WSDL is located at http://nlp.racai.ro/webservices/LangIdWebService.asmx?WSDL.

Sample application using the service is located at http://nlp.racai.ro/webservices/LanguageId.aspx.

TTL

TTL (Tokenisation, Tagging and Lemmatisation) web service offers the following remote procedures:

- *SentenceSplitter* – takes as parameters the language of the text to process (currently either “en” or “ro”) and a SGML entity encoded text and returns another string which is a list of sentences separated by CR/LF sequence,
- *Tokenizer* – has as parameters the language code and a sentence and returns a list of tokens separated by CR/LF each token possibly carrying its NE (named entity) tag (added to the token with the tab character) given by the NER (named entity recognition) module of the

SentenceSplitter in the case the token is a NE (i.e., a real or integer number, a roman number, percents, abbreviations, dates, clock times, etc.),

- *Tagger* – takes the language code and a tokenized sentence from Tokenizer and returns a MSD (morphosyntactic description) tagged sentence which is a string with triples of token, Tab character, MSD separated by CR/LF,
- *Lemmatizer* – uses the POS tagged sentence along with the language code and returns a lemmatized sentence which resembles the one from the Tagger's output except that the token annotation is enriched with its lemma which is separated again from the MSD tag by a Tab,
- *Chunker* – is the final operation of TTL and, beside the language code, it takes a lemmatized sentence and returns the same sentence with chunk information added after the lemma annotation,
- *XCES* – is a helper function which calls all the previously mentioned operations in order and returns an XCES (XML Corpus Encoding Standard) representation of the result.

TTL WSDL is located at http://ws.racai.ro/ttlws.wsdl.

LexPar

LexPar web service provides only one function *LinkSentence* which generates the dependency of the tokenized, tagged and chunked sentence.

LexPar WSDL is located at http://ws.racai.ro/lxpws.wsdl.

TextProcessing

TextProcessing web service provides only one function *Process* which combines TTL processing (tokenization, sentence splitting, POS-tagging and morpho-syntactical annotation) in a single action.

TextProcessing WSDL is located at http://nlp.racai.ro/WebServices/TextProcessing.asmx?WSDL.

WordNet Browser

Wordnet browser (http://nlp.racai.ro/wnbrowser/) allows hyperbolic browsing through aligned Princeton 2.0 and the Romanian wordnets.

A common usage scenario for the current wordnet web service is to translate a word to and from Romanian/English: (i) the client applications queries the web service for all the synsets ids of a given literal in the target language; (ii) the client queries for all the synsets of the corresponding ids in the source language; (iii) the client application extracts the literals from the source.

Development facilities, such as getting the synset unique identifiers for a given word (either in Romanian or in English), finding the semantic distance between arbitrary synsets (both monolingually and, via the Interlingual index, crosslingually), getting the translation equivalents for a given word sense, its SUMO, DOMAIN or subjectivity annotation etc. are planned to be added.

WordNetBrowser WSDL is located at http://nlp.racai.ro/wnbrowser/Wordnet.asmx?wsdl.

Factored Statistical Machine Translation

Factored Translation is a processing flow based on other web services and provides translation services for Romanian to English and English to Romanian for legal documents. The system has been trained on JRC-Acquis and therefore its best performance is for texts belonging to this register.

Web service protocols

SOAP

Encoding of linguistic resources

TTL

TTL operates with SGML entities (not UTF-8 representation); a helper function is available to transform the input text from UTF-8 to SGML.

Subsequent steps of the processing chain collect tokenized data in CR/LF separated rows, each row containing Tab-separated token, tagging information, lemma and chunk information in the following form:

```
This        Pd3-s       this
is          Vmip3s      be          Vp#1
a           Ti-s        a           Np#1
simple      Afp         simple      Np#1,Ap#1
example     Ncns        example     Np#1
of          Sp          of          Pp#1
a           Ti-s        a           Pp#1,Np#2
web         Ncns        web         Pp#1,Np#2
service     Ncns        service     Pp#1,Np#2
remote      Afp         remote      Pp#1,Np#2,Ap#2
execution   Ncns        execution   Pp#1,Np#2
.           PERIOD      .
```

Additionally, XCES helper function provides XML representation of the result:

```
<seg lang="en">
  <s id="example.1">
    <w lemma="this" ana="Pd3-s">This</w>
    <w lemma="be" ana="Vmip3s" chunk="Vp#1">is</w>
    <w lemma="a" ana="Ti-s" chunk="Np#1">a</w>
    <w lemma="simple" ana="Afp" chunk="Np#1,Ap#1">simple</w>
    <w lemma="example" ana="Ncns" chunk="Np#1">example</w>
    <w lemma="of" ana="Sp" chunk="Pp#1">of</w>
    <w lemma="a" ana="Ti-s" chunk="Pp#1,Np#2">a</w>
    <w lemma="web" ana="Ncns" chunk="Pp#1,Np#2">web</w>
    <w lemma="service" ana="Ncns" chunk="Pp#1,Np#2">service</w>
    <w lemma="remote" ana="Afp" chunk="Pp1,Np#2,Ap#2">remote</w>
    <w lemma="execution" ana="Ncns"
                         chunk="Pp#1,Np#2">execution</w>
    <c>.</c>
  </s>
</seq>
```

Text Processing

The results are returned in a concise, proprietary format (space- and vertical bar-separated):

```
This|this|DMS|Pd3-s is|be|VERB3|Vmip3s a|a|TS|Ti-s simple|simple|
ADJE|Afp example|example|NN|Ncns of|of|PREP|Sp a|a|TS|Ti-s web|
web|NN|Ncns service|service|NN|Ncns remote|remote|ADJE|Afp
execution|execution|NN|Ncns .|.|PERIOD|PERIOD
```

LexPar

LinkSentence function takes as parameters the XCES encoding of the sentence to be processed and the language code and returns the XML encoding enriched with the dependency information such as:

```
<seg lang="en">
  <s id="example.1">
    <w lemma="this" ana="Pd3-s" head="1"> This</w>
    <w lemma="be" ana="Vmip3s" chunk="Vp#1">is</w>
    <w lemma="a" ana="Ti-s" chunk="Np#1" head="5"> a</w>
    <w lemma="simple" ana="Afp"
       chunk="Np#1,Ap#1" head="5">simple</w>
    <w lemma="example" ana="Ncns"
       chunk="Np#1" head="1">example</w>
...
```

WordNet Browser

The data of WordNets is stored in a database in proprietary XML with records like:

```
<SYNSET>
  <ID>ENG20-12977363-n</ID>
  <POS>n</POS>
  <SYNONYM>
    <LITERAL>cvintilion<SENSE>1</SENSE></LITERAL>
  </SYNONYM>
  <DEF>un milion de cvadrilioane</DEF>
  <ILR>ENG20-12969974-n<TYPE>hypernym</TYPE></ILR>
  <DOMAIN>number</DOMAIN>
  <SUMO>PositiveInteger<TYPE>@</TYPE></SUMO>
  <SENTIWN><P>0.0</P><N>0.0</N><O>1</O></SENTIWN>
</SYNSET>
```

Output format of data retrieval is similar, wrapped up in `<Result>` element.

Linguistic data categories

Tagsets

Multext-East (Erjavec 2004) compliant lexical tagset (614 tags for Romanian and 133 tags for English) and a reduced tagset (according to the tiered tagging model: 92 tags for Romanian and 95 tags for English) is used (Tufiş 2000).

Dependency information

Additional `head` attribute indicates the position in the sentence (0-based numbering) to which the token is linked (the naming of the attribute does not imply that the token with the `head` information is actually the head of the relation). The token without this attribute (in our example the verb *be*) is the root of the dependency. The dependency of the sentence is a connected planar and acyclic graph. The cases in which the graph is not connected might appear because the syntactic filter occasionally rejects good links which otherwise (in the vast majority of cases) are not correct.

1.6 WS-LexicalPlatform

General information

Name of the service/project

WS-LexicalPlatform

Availability

Web services are protected by the x509 certificates. A simple authentication for the whole Simple database based on the Apache Web Server will be prepared shortly.

Search application for browsing the lexicon built on top the WSs is freely available at http://www.clarin-it.it/Simple/SimpleGUI.html.

Body responsible

Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale (CNR-ILC), Pisa, Italy

Status of the tools

Internal to CNR-ILC (experimental)

Supported languages

Italian

Implemented NLP services

Services and functionalities offered by WS-LexicalPlatform can be classified into two main categories:

- functions to elaborate and present the user data from a legacy data source (SIMPLE Italian lexicon),
- functions that provide standard mechanisms for the interoperability among software agents.

PhonoMorpho

Deals with retrieving information concerning phonology and morphology.

PhonoMorpho WSDL is located at http://www.clarin-it.it/Simple/services/PhonoMorphoSOAP?wsdl.

Syntax

Deals with syntactic level of the lexicon.

Syntax WSDL is located at http://www.clarin-it.it/Simple/services/SyntaxSOAP?wsdl.

Semantic

Deals with semantic level and relationship of the lexicon entries.

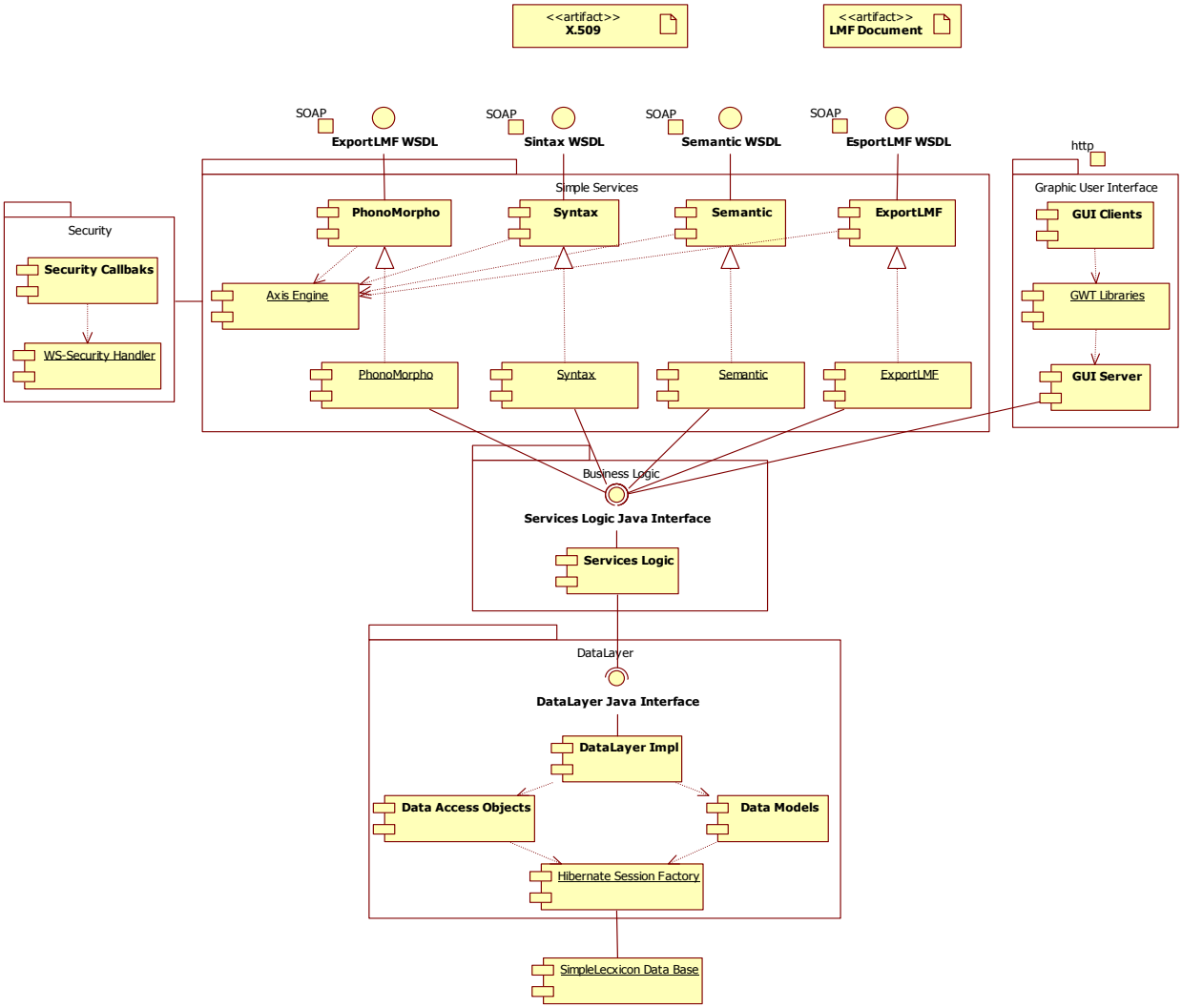
Semantic WSDL is located at http://www.clarin-it.it/Simple/services/SemanticSOAP?wsdl.

ExportLMF

Exports a whole LMF entry.

ExportLMF WSDL is located at http://www.clarin-it.it/Simple/services/ExportLMFSOAP? wsdl.

The following figure shows the WS-LexicalPlatform Architecture:



Web service protocols

For all web-services SOAP protocol is used. Basic RESTful web services are under development.

Encoding of linguistic resources

LMF standard is used as the interchange/standard representation output format.

Linguistic data categories

The data categories used at the moment are proprietary to the SIMPLE lexicons, but in general they have been derived from the EAGLES-ISLE initiatives. They are mappable to ISO DCR categories, especially for the morphosyntactic profile, and most of them are likely to be promoted to the future ISO standardization of data categories and, therefore, be present in ISOCat.

1.7 LXService

General information

Name of the service/project

LXService – a web service for language technology of Portuguese

Availability

Public, after the authorization from the Department of Informatics, University of Lisbon

Bodies responsible

University of Lisbon, Department of Informatics, Natural Language and Speech Group (NLX), Lisbon, Portugal

Status of the tools

Stable

Supported languages

Portuguese

Implemented NLP services

LX-Chunker

Segments the text into sentences and paragraphs.

A f-score of 99.94% was obtained when testing on a 12,000 sentence corpus accurately hand tagged with respect to sentence and paragraph boundaries.

LX-Tokenizer

Segments the paragraphs into lexemes (Branco and Silva 2003).

This tool achieves a f-score of 99.72%.

LX-Tagger

Annotates tokenized text with POS tags (Branco and Silva 2004, Silva 2007).

This tagger was developed with TnT software over 90% of a small, 260k token, accurately hand tagged corpus. Accuracy of 96.87% was obtained with the tagger being trained over 90% of the 260K tokens and evaluated over the held out 10%, this being repeated over 10 different test runs and the results averaged.

LXService WSDL is located at http://nlxserv.di.fc.ul.pt/axis/services/LXService?wsdl.

LXService (Branco *et al.* 2008) is being expanded so that further methods are included to grant access to tools concerned with lemmatization, morphological analysis (LX-Lemmatizer and LX-Inflector – Branco and Silva 2006, Nunes 2007, Martins 2008) and parsing (Silva *et al.* forth.).

Web service protocols

SOAP

Encoding of linguistic resources

LX Chunker

Sentences split over different lines are unwrapped.

Sentence boundaries are marked with `<s>...</s>`, paragraph boundaries with `<p>...</p>`.

LX Tokenizer

Token boundaries (indicated with whitespaces) are marked with vertical bar (`|`):

```
um exemplo -> |um|exemplo|
```

Contractions are expanded; the first element of an expanded contraction is marked with an _ (underscore) symbol:

```
do -> |de_|o|
```

Spacing around punctuation or symbols is marked; `\*` and the `*/` symbols indicate a space to the left and a space to the right:

```
um, dois e três -> |um|,*/|dois|e|três|
5.3 -> |5|.|3|
1. 2 -> |1|.*/|2|
8 . 6 -> |8|\*.*/|6|
```

Clitic pronouns are detached from the verb. The detached pronoun is marked with a `-` (hyphen) symbol. When in mesoclisis, a `-CL-` mark is used to signal the original position of the detached clitic. Additionally, possible vocalic alterations of the verb form are marked with a # (hash) symbol:

```
dá-se-lho -> |dá|-se|-lhe|-o|
afirmar-se-ia -> |afirmar-CL-ia|-se|
vê-las -> |vê#|-las|
```

Ambiguous strings are resolved. Depending on their particular occurrence, these strings can be tokenized in different ways. For instance:

```
deste -> |deste| (when occurring as a Verb)
deste -> |de|este| (when occurring as a contraction (Preposition
+ Demonstrative)
```

LX Tagger

A single morpho-syntactic tag is being assigned to every token. The tag is attached to the token using a `/` (slash) symbol as separator:

```
um exemplo -> um/IA exemplo/CN
```

Each individual token in multi-token expressions gets the tag of that expression prefixed by "`L`" and followed by the number of its position within the expression:

```
de maneira a que -> de/LCJ1 maneira/LCJ2 a/LCJ3 que/LCJ4
```

Linguistic data categories

The tables below present basic tags used by the service.

POS tags

Tag	Description
ADJ	Adjectives
ADV	Adverbs
CARD	Cardinals
CJ	Conjunctions
CL	Clitics
CN	Common Nouns
DA	Definite Articles
DEM	Demonstratives
DFR	Denominators of Fractions
DGTR	Roman Numerals
DGT	Digits
DM	Discourse Marker
EADR	Electronic Addresses
EOE	End of Enumeration
EXC	Exclamative
GER	Gerunds
GERAUX	Gerund "ter"/"haver" in compound tenses
IA	Indefinite Articles
IND	Indefinites
INF	Infinitive
INFAUX	Infinitive "ter"/"haver" in compound tenses
INT	Interrogatives
ITJ	Interjection
LTR	Letters

Tag	Description
MGT	Magnitude Classes
MTH	Months
NP	Noun Phrases
ORD	Ordinals
PADR	Part of Address
PNM	Part of Name
PNT	Punctuation Marks
POSS	Possessives
PPA	Past Participles not in compound tenses
PP	Prepositional Phrases
PPT	Past Participle in compound tenses
PREP	Prepositions
PRS	Personals
QNT	Quantifiers
REL	Relatives
STT	Social Titles
SYB	Symbols
TERMN	Optional Terminations
UM	um or "uma"
UNIT	Abbreviated Measurement Units
VAUX	Finite "ter" or "haver" in compound tenses
V	Verbs (other than PPA, PPT, INF or GER)
WD	Week Days

Tags for multiword expressions

Tag	Description
LADV1…LADVn	Multi-Word Adverbs
LCJ1…LCJn	Multi-Word Conjunctions
LDEM1…LDEMn	Multi-Word Demonstratives
LDFR1…LDFRn	Multi-Word Denominators of Fractions
LDM1…LDMn	Multi-Word Discourse Markers

Tag	Description
LITJ1…LITJn	Multi-Word Interjections
LPRS1…LPRSn	Multi-Word Personals
LPREP1…LPREPn	Multi-Word Prepositions
LQD1…LQDn	Multi-Word Quantifiers
LREL1…LRELn	Multi-Word Relatives

Other tags

Tag	Description
m	Masculine
f	Feminine
s	Singular
p	Plural
dim	Diminutive
sup	Superlative
comp	Comparative
1	First Person
2	Second Person
3	Third Person

Tag	Description
pi	Presente do Indicativo
ppi	Pretérito Perfeito do Indicativo
ii	Pretérito Imperfeito do Indicativo
mpi	Pretérito Mais que Perfeito do Indicativo
fi	Futuro do Indicativo
c	Condicional
pc	Presente do Conjuntivo
ic	Pretérito Imperfeito do Conjuntivo
fc	Futuro do Conjuntivo
imp	Imperativo

1.8 WROCUT/ICS PAS services

General information

Name of the service/project

- TaKIPI WS,
- SuperMatrix WS,
- plWordNet WS,
- Spejd WS.

General URL

http://plwordnet.pwr.wroc.pl/clarin/ws/

Availability

- TaKIPI – GPL-licenced, available at http://plwordnet.pwr.wroc.pl/clarin/ws/,
- plWordNet – freely accessible for online browsing at http://www.plwordnet.pwr.wroc.pl,
- Spejd – GPL-licenced, available at http://nlp.ipipan.waw.pl/Spejd/.

Bodies responsible

Institute of Informatics, Wrocław University of Technology (WROCUT), Wrocław, Poland
Institute of Computer Science, Polish Academy of Sciences (ICS PAS), Warsaw, Poland

Status of the tools

- Tools at the backend of web services – stable,
- TaKIPI WS – stable,
- SuperMatrix WS – still in development,
- Spejd WS – prototype.

Supported languages

- TaKIPI WS (tools) and plWordNet WS (resource) – Polish,
- SuperMatrix and SuperMatrix WS – language independent, however, in the present version, due to the tools integrated, SuperMatrix offers its full functionality for Polish and English,
- Spejd WS – language independent, however, in the present version, due to the integrated grammar and tagset, Spejd offers its full functionality for Polish.

Implemented NLP services

The up to date technical description and documentation for all WSs of WROCUT can be found at http://plwordnet.pwr.wroc.pl/clarin/ws/.

TaKIPI WS

TaKIPI (WROCUT; Piasecki and Radziszewski 2009) is a set of morphosyntactic tools for Polish including the morphosyntactic tagger called TaKIPI. The set comprises a complete chain of the basic morphosyntactic processing and gives access to the whole processing chain as well as to all steps separately. The chain utilises a morphological analyser called *Morfeusz* (Woliński 2006).

The web service provides functionality for text tagging, lematization, segmentation, morphologic analysis and tokenization.

WSDL URL: http://plwordnet.pwr.wroc.pl/clarin/ws/takipi/takipi.wsdl

SuperMatrix WS

SuperMatrix (Broda and Piasecki 2008) supports automatic acquisition of lexical semantic relations from corpora for Polish and English. It enables extraction of coincidence matrices from large amount of text. The words in the matrices can be described by the whole range of means: from simple co-occurrences to instances of lexico-syntactic relations identified with the help of lexico-morphosyntactic constraints, in the case of Polish, or shallow syntactic processing, in the case of English. The constructed matrices can be next filtered and transformed (according to several

different algorithms). Finally, different measures of semantic relatedness can be obtained by the means of several well known and unique algorithms. SuperMatrix can be combined with the clustering tool called CLUTO (Karypis 2002) and wordnets, e.g., Princeton WordNet and *plWordNet* (Piasecki *et al.* 2009) – a wordnet for Polish (http://www.plwordnet.pwr.wroc.pl).

SuperMatrix WS will give access to full functionality of the SuperMatrix system. Users will be able to work with existing matrices by browsing various matrix statistics and inspecting semantic relatedness of selected words according to the selected matrix and algorithm (this part is already implemented). Next, users will be able to upload their own corpora and define the process of the co-incidence matrix construction, build the matrix and extract the measures. The definition of the process will include: the list of words to be described, types of features to be extracted, and the type of filtering and transformation to be performed.

The features can be simply defined by a list of words (for co-occurrence counting), but also by the specification of complex lexicalised morphosyntactic constraints.

Planned high level functions (final list not closed yet):

- construction of words co-occurrence matrices from large corpora,
- counting of measures of semantic relatedness between words,
- construction of language profiles for selected flag words on the basis of the corpora and list delivered by the user,
- re-implementation of the HAL technique for the needs of psychological experiments performed on corpora delivered by the user,
- extraction of associations between expressions representing certain concepts, proper names, trademarks and common words.

WSDL URL (in development): http://plwordnet.pwr.wroc.pl/clarin/ws/supermatrix/supermatrix.wsdl.

plWordNet WS

WS delivers means for browsing and retrieving lexical units and their relations in plWordNet (WROCUT).

WSDL URL: http://plwordnet.pwr.wroc.pl/clarin/ws/plwordnet/plwordnet.wsdl.

Spejd WS

Spejd WS (http://code.google.com/p/spejdws/) provides shallow parser and disambiguation for Polish. It uses Spejd (Buczyński and Przepiórkowski 2009) engine for shallow parsing using cascade grammars. It also may also use TaKIPI WS (http://plwordnet.pwr.wroc.pl/clarin/ws/takipi/) for tokenization, segmentation, lemmatization and morphologic analysis, if requested.

Parsing rules are defined using cascade regular grammars, which match against the orthographic form 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.

XMLRPC URL: http://chopin.ipipan.waw.pl:8081/spejdws/xmlrpc.

WSDL URL: http://chopin.ipipan.waw.pl:8081/spejdws/services/SpejdService?wsdl.

Web service protocols

SOAP

Encoding of linguistic resources

- Wordnet-LMF in plWordNet,
- XCES (xcesAnaIPI, the version of the XCES standard used in the ICS PAS Corpus) in TaKIPI WS, SuperMatrix and Spejd.

Linguistic data categories

TaKIPI WS

ICS PAS tagset (used in the ICS PAS Corpus and, in slightly modified form, in the National Corpus of Polish). See http://korpus.pl/en/cheatsheet/node2.html for details.

SuperMatrix WS

ICS PAS tagset; SuperMatrix has been also adapted to CLAWS5 (tagset used in the British National Corpus) and the format of the MiniPar parser output.

plWordNet WS

The results are returned in one of two formats: WordNet-LMF (Aliprandi *et al.* 2009) – an XML-based lexical data format for wordnets (selected here to increase the interoperability of the WS) – and a composition of nested programming language objects that can be easily manipulated in other applications.

Spejd WS

ICS PAS tagset.

2 Summary of linguistic properties

2.1 NLP functionalities

Noticeably, most advanced processing chains, such as WebLicht, offer the broadest scope of functionality, addressing various NLP fields. At the same time, other frameworks addressing narrower specific fields, such as parsing-related activities or statistical functionality, nevertheless tend to provide complete approach within their scope.

The table on the next page summarizes coverage of LRT functionality available in reviewed frameworks.

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

2.2 Encoding of linguistic resources

Reviewed tools use different data encoding formats – most of them proprietary, most of them XML based.

In RACAI Services SGML is used for internal encoding of data, but a helper function is available to provide means of decoding UTF-8 (most likely, XML-encoded) data into SGML entities. Similarly, parallel to proprietary (although easy to handle) Tab-separated format, XML output (XCES-encoded) is also provided by a supplementary function.

WebLicht uses proprietary TextCorpus (TCF), Lexicon and Metadata formats. TCF strives to be compatible with established standards, especially the data formats of the ISO TC37 SC4 group:

- LAF: Linguistic Annotation Framework,
- LMF: Lexical Markup Framework,
- MAF: Morpho-Syntactic Annotation Framework.

In case of proprietary formats, availability of converters transforming existing language resources into standard formats is essential. WebLicht, for instance, offers converters for PAULA and MAF. It should be noted that the border between acknowledged standards and proprietary formats is fluid. Proprietary extensions of recognized formats may retain the advantages of the latter while preserving supplementary properties unavailable in the core standard, serving as a golden mean for difficult representations.

The table below summarizes available output formats of reviewed services:

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

2.3 Linguistic data categories

The table below summarizes information about tagsets used to encode linguistic annotation by 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					x					x
GATE					x					
IULA		x								
ILSP		x		x						
RACAI			x							
WS-LexPl									x	

	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)**
LXService							x			
WROCUT/ ICS PAS	x					x				

Similarly to encoding of linguistic resources, the border between tagsets considered standard and proprietary is vague. Some tagsets (such as STTS for German or ICS PAS tagset for Polish), although not always being familiar to the whole LRT community, are universally used for specific languages or constitute regional standards.

Still, proprietary tagsets are most widely used, but tendency to standardize becomes apparent since the number of standard- or semi-standard tagsets used among reviewed services (simultaneously or by specific linguistic field) is significant.

3 Preliminary Conclusions

The presence of such a broad spectrum of different solutions seems to show that the necessity of using widely-accepted standards, especially at the level of linguistic data categories, may be still underestimated by many NLP developers and resource providers, probably due to costs of conversion of proprietary formats, as well as, in case of data categories, because of the inherent complexity of linguistic issues involved.

At the same time, a standardized approach is a prerequisite to convert standalone applications into cooperating services. The role of CLARIN to create and promote standards is therefore of far-reaching significance.

In many cases, discernible efforts in the direction of ensuring compliance or mappability to the current ISO or *de facto* standards are made. At present, most efforts go in the direction of structural or formal compliance to the more general standard models for the various resource types (i.e., lexicon encoding and text annotation formats). Significant effort must, however, still be made to ensure interoperability at the semantic level, i.e., at the level of linguistic data categories (incl. tagsets).

Given the present situation in the LRT and NLP community, CLARIN puts forward the vision of standards not so much as imposed formats and data categories for proprietary lexica, corpora and NLP tools, but as interchange formats to be used, at least in a short-term interoperability scenario. Thus, CLARIN would require proprietary formats to be mapped to standards recommended by CLARIN (cf. *Standards for Text Encoding: A CLARIN ShortGuide* at http://www.clarin.eu/documents for a preliminary overview).

3.1 Standards for the interoperability of linguistic tools

Interoperability of reviewed linguistic tools is currently very limited. Data encoding formats vary, both in terms of character encoding, as well as in terms of representation of linguistic structures and annotation.

Nevertheless, ability to accept and deliver linguistic data in standard representational format is absolutely necessary to enable data comparison and merging, as well as for processing the data by common tools. To accomplish this, either a standard format must be used from the very beginning, in the whole process of data encoding, storage and processing, or transducers must be made available to enable lossless conversion of data from and to proprietary formats.

In general, interoperability of language resources can be obtained at several levels and by various means, e.g.:

- using standardized data exchange format,
- using common language resource data model.

Both issues are briefly addressed below – the most important observation to be pointed out here is that the tendency to adopt XML as the basic data exchange format can be easily noticed. Definitely no proprietary format and not even SGML can currently offer such processing possibilities as XML in terms of available tools, frameworks and support.

Technical interoperability

Technical issues are of little concern in this deliverable, although one important remark can be made on the basis of the reviewed frameworks. Currently two almost equally popular web service protocols are used: REST and SOAP. From the linguistic point of view, the protocol is nothing more than a means to interface resources (through services) with the outside world, similar to what web services are offering as compared to standalone applications. Although important for the practical interoperability of LRTs, the particular choice of the WS protocol is not a concern of the current deliverable (but see CLARIN D2R-6 deliverable, Requirement Specification Web Services and Workflow systems).

Formal interoperability vs. semantic interoperability

Current standards, especially the official ones, give a good way of establishing formal or syntactic interoperability, which is a fundamental prerequisite for interoperability at large. Assuming XML interchange formats following official representation standards, interoperability at resource-structure level can be achieved.

However, the problem of semantic interoperability still remains open: even ensuring isomorphic structures, the meaning of the descriptors may still be unknown. With this respect, the idea of formally mapping (proprietary) descriptors to (some) standard concepts, as those in ISOCat, appears to be, at present, a practical and tangible solution.

3.2 Standards for the encoding of linguistic data and the representation of linguistic annotation

As already suggested in the previous section, XML-based formats seem to satisfy most aspects of the linguistic encoding. The verbosity of XML and the cost of its processing do not appear to act as a deterrent in this process, because the benefits are more than ample.

Major standardization initiatives (EAGLES, ISLE, ISO, etc.) never promoted the vision of forcing the use of a single dominant representation format for in-house use, which could be seen as impeding research creativity. Rather, they have always promoted the view that providing some standardized representation of data and program outputs is fundamental when sharing and re-using data. And this is even more crucial for an infrastructure like CLARIN. Therefore, the adoption of general metamodels could be a viable solution, because they can accommodate many different representation

conventions. However, clear guidelines and best practices are needed to establish a common strategy for converting different resource types into interchange formats compliant with those standards (i.e., examples and best practices of how to convert, for example, Penn Treebank-style corpora into a LAF representation need to be established, in order to ensure real interoperability).

Obviously, the usefulness of such interchange formats must be proven in practice. If such formats are very abstract and permissive metamodels, they will only provide a thin coat over the original resource, and the intimate knowledge of the original encoding will be necessary to deal with such "interchange" representation. If they are too specific, they may overtly constrain the kinds of information that may be represented with the use of such formats. Unless the right balance is achieved, LRT practitioners may opt to work with a greater number of specific and well-defined standards such as TIGER-XML and the (original version of) XCES, rather than with conceptually more attractive standard interchange formats. It is one of the tasks of CLARIN – and the focal point of the second D5R-3 deliverable, to be completed by the end of 2010 – to make recommendations on which data exchange formats and language resource models should be adopted by the LRT community.

4 Bibliography

Aliprandi, Carlo, 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.

Atserias, Jordi, 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 fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA*. Genoa, Italy. May 2006.

Bel, Núria, 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*. Paris: European Language Resources Association. 1362-1367, Génova, Italia.

Branco, António and João Silva, 2003. Contractions: breaking the tokenization-tagging circularity. *Lecture Notes in Artificial Intelligence* 2721. Berlin, Springer-Verlag, ISSN 0302-9743, pp. 167-170.

Branco, António and João Silva, 2004. Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. In *Proceedings LREC2004*.

Branco, António and João Silva, 2006. Dedicated Nominal Featurization of Portuguese. Lecture Notes in Artificial Intelligence, 3960, Berlin, Springer-Verlag.

Branco, António, Francisco Costa, Pedro Martins, Filipe Nunes, João Silva and Sara Silveira, 2008. LXService: Web Services of Language Technology for Portuguese. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, Piperidis, D. Tapias (eds.), *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC2008)*, Paris, ELRA.

Broda, B. and M. Piasecki, 2008. SuperMatrix: a General Tool for Lexical Semantic Knowledge Acquisition. In G. Demenko, K. Jassem and S. Szpakowicz (ed.) *Speech and Language Technology*. Vol. 11, pp. 239-254, Polish Phonetics Association.

Buczyński, A. and A. Przepiórkowski, 2009. Spejd: A Shallow Processing and Morphological Disambiguation Tool. In Z. Vetulani and H. Uszkoreit (ed.) *Human Language Technology: Challenges of the Information Society*. Vol. 5603, pp. 131-141, Berlin, Springer-Verlag.

CLARIN, 2009. Standards for Text Encoding: A CLARIN ShortGuide. http://www.clarin.eu/documents/.

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

Erjavec, Tomaz, 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'2004*, ELRA, Paris.

Francopoulo, G., M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, C. Soria, 2006. Lexical Markup Framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (LREC 2006), Genova, Italy, pp. 233-236.

Francopoulo, G., N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, C. Soria, 2009. Multilingual Resources for NLP in the Lexical Markup Framework (LMF). Guest Editors: G. Sérasset, A. Witt, U. Heid, F. Sasaki. Language Resources and Evaluation Journal, Vol. 43:1, pp. 57-70.

Ide, Nancy, Patrice Bonhomme, and Laurent Romary, 2000. XCES: An XML-based standard for linguistic corpora. In LREC (2000), pages 825–830.

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

Lenci A., N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowsky, I. Peters, W. Peters, N. Ruimy, M. Villegas, A. Zampolli, 2000. SIMPLE: A General Framework for the Development of Multilingual. Lexicons. International Journal of Lexicography XIII (4). pp. 249-263.

Lin, D., 1993. Principle-based parsing without overgeneration. In: Proc. 31st Meeting of the ACL. pp. 112-120.

Martins, Pedro, 2008. Desambiguação Automática da Flexão Verbal em Contexto. Msc Dissertation, University of Lisbon.

Mengel, Andreas and Wolfgang Lezius, 2000. An XML-based encoding format for syntactically annotated corpora. In *LREC (2000)*, pages 121–126.

Nunes, Filipe, 2007. Verbal Lemmatization and Featurization of Portuguese with Ambiguity Resolution in Context. MSc Dissertation, University of Lisbon.

Papageorgiou, Harris, Prokopis Prokopidis, Iason Demiros, Voula Giouli, Alexis Konstantinidis, and Stelios Piperidis, 2002. Multi-level XML-based Corpus Annotation. In *Proceedings of LREC 2002*.

Piasecki, M.; Szpakowicz, S. and Broda, B., 2009. A Wordnet from the Ground Up. Oficyna Wydawnicza Politechniki Wrocławskiej. http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf.

Piasecki, M. and Radziszewski, A., 2009. Morphosyntactic Constraints in Acquisition of Linguistic Knowledge for Polish. In Mykowiecka, A. and Marciniak, M. (ed.) *Aspects of Natural Language Processing (a festschrift for Professor Leonard Bolc)*, Springer, LNCS 5070, pp. 163-190.

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

Przepiórkowski, A., 2004. The ICS PAS Corpus, Preliminary Version. Institute of Computer Science PAS. http://nlp.ipipan.waw.pl/~adamp/Papers/2004-corpus/.

Ruimy, N., M. Monachini, R. Distante, E. Guazzini, S. Molino, M. Ulivieri, N. Calzolari, and A. Zampolli, 2002. Clips, a multi-level Italian computational lexicon: A glimpse to data. In *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC'02), Las Palmas de Gran Canaria, Spain.

Silva, João, 2007. Shallow Processing of Portuguese: From Sentence Chunking to Nominal Lemmatization. MSc Dissertation, University of Lisbon.

Silva, João, António Branco, Sérgio Castro and Ruben Reis, forth. Out-of-the-box Robust Parsing for Portuguese. In *Proceedings of PROPOR'2010 – 9th International Conference on the Computational Processing of Portuguese*, Porto Alegre, PUCRS.

Toral, A. and M. 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).

Tufiş, D., 2000. Using a Large Set of EAGLES-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging. International Conference on Language Resources and Evaluation LREC'2000, Athens, pp. 1105-1112.

Villegas, Marta, Núria Bel and Santiago Bel; Alemany, Francesca; Martínez, Hèctor, 2009. Lexicography in the grid environment. In *Proceedings of e-lex 2009*. Lovaina: Cahiers du Cental. Pàg (in print).

Vivaldi Palatresi, Jorge, 2009. Corpus and exploitation tool: IULACT and bwanaNet dins Cantos Gómez, Pascual; Sánchez Pérez, Aquilino (eds.) *A survey on corpus-based research = Panorama de investigaciones basadas en corpus [Actas del I Congreso Internacional de Lingüística de Corpus (CICL-09), 7-9 Mayo 2009, Universidad de Murcia]*. Murcia: Asociación Española de Lingüística del Corpus. Pàg. 224-239. ISBN 978-84-692-2198-3.

Woliński Marcin, 2006. Morfeusz – a practical tool for the morphological analysis of Polish. In M. A. Kłopotek, S. T. Wierzchoń, and K. Trojanowski, editors. In *Proceedings of the International IIS: IIPWM'06 Conference*. Wisła, Poland. June 2006, pp. 511–520.

Wright, S. E., 2004. A global data category registry for interoperable language resources. *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC 2004), Lisbon, Portugal, CD-ROM.