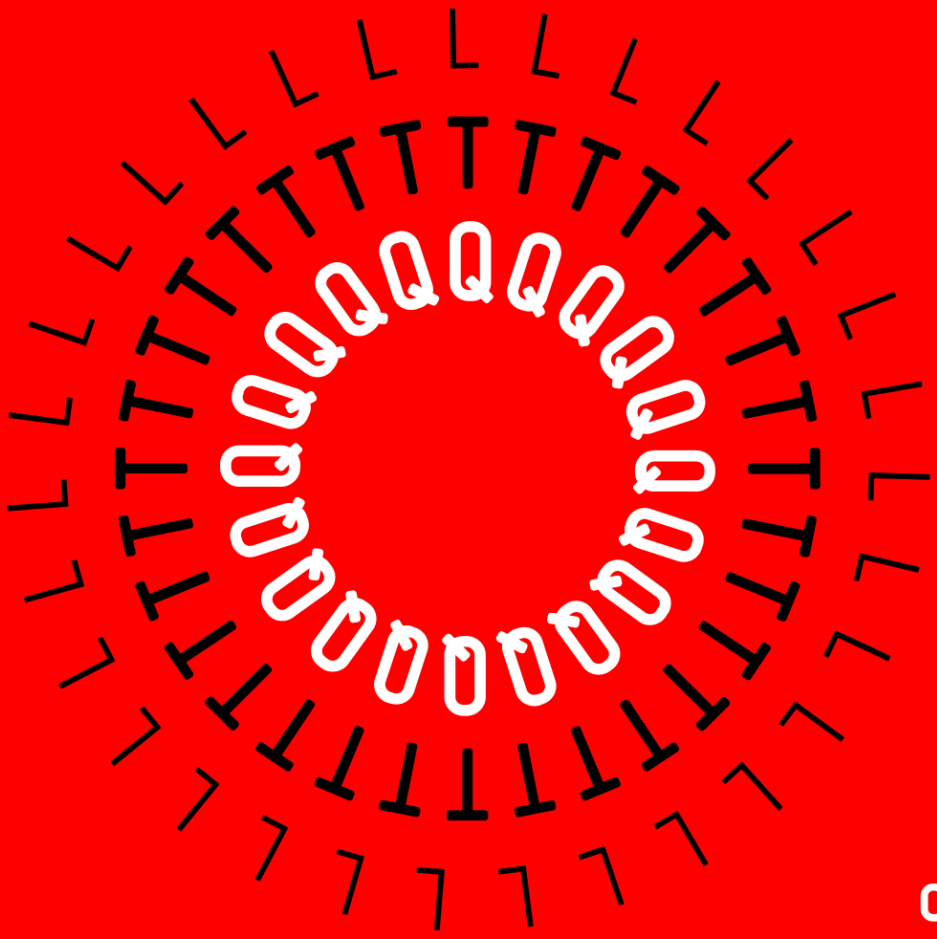


qtleap
quality translation by deep language engineering approaches

Report on First pilot version of LRTs enhanced to support Deep Processing

DELIVERABLE D4.7

VERSION 3.0 | 2015-06-15

QTLeap

Machine translation is a computational procedure that seeks to provide the translation of utterances from one language into another language.

Research and development around this grand challenge is bringing this technology to a level of maturity that already supports useful practical solutions. It permits to get at least the gist of the utterances being translated, and even to get pretty good results for some language pairs in some focused discourse domains, helping to reduce costs and to improve productivity in international businesses.

There is nevertheless still a way to go for this technology to attain a level of maturity that permits the delivery of quality translation across the board.

The goal of the QTLeap project is to research on and deliver an articulated methodology for machine translation that explores deep language engineering approaches in view of breaking the way to translations of higher quality.

The deeper the processing of utterances the less language-specific differences remain between the representation of the meaning of a given utterance and the meaning representation of its translation. Further chances of success can thus be explored by machine translation systems that are based on deeper semantic engineering approaches.

Deep language processing has its stepping-stone in linguistically principled methods and generalizations. It has been evolving towards supporting realistic applications, namely by embedding more data based solutions, and by exploring new types of datasets recently developed, such as parallel DeepBanks.

This progress is further supported by recent advances in terms of lexical processing. These advances have been made possible by enhanced techniques for referential and conceptual ambiguity resolution, and supported also by new types of datasets recently developed as linked open data.

The project QTLeap explores novel ways for attaining machine translation of higher quality that are opened by a new generation of increasingly sophisticated semantic datasets and by recent advances in deep language processing.

www.qtleap.eu

Funded by

QTLeap is funded by the 7th Framework Programme of the European Commission.

Supported by

And supported by the participating institutions:

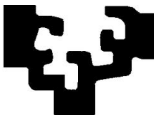
Faculty of Sciences, University of Lisbon

German Research Centre for Artificial Intelligence

Charles University in Prague

Bulgarian Academy of Sciences

Humboldt University of Berlin



University of Basque Country

University of Groningen

Higher Functions, Lda

Revision history

Version	Date	Authors	Organisation	Description
1.0	Oct 06, 2014	Petya Osenova	IICT-BAS	First draft
1.1	Oct 21, 2014	João Silva	FCUL	Integrated text
1.2	Oct 22, 2014	Aljoscha Burchardt	DFKI	Integrated text
1.3	Oct 28, 2014	Gorka Labaka, Eneko Agirre, Nora Aranberri;	UPV-EHU	Integrated text
1.4	Oct 28, 2014	Dieke Oele, Gertjan van Noord	UG	Integrated text
1.5	Oct 28, 2014	Petya Osenova	IICT-BAS	Editing
1.6	Oct 31, 2014	Markus Egg	UBER	Review comments incorporated
2.0	Nov 14, 2014	Kiril Simov	IICT-BAS	Narrative descriptions added
3.0	May 28, 2015	João Silva	FCUL	Integrated text
3.0	May 28, 2015	Aljoscha Burchardt	DFKI	Integrated text
3.0	May 29, 2015	Gorka Labaka, Eneko Agirre, Nora Aranberri;	UPV-EHU	Integrated text
3.0	May 29, 2015	Dieke Oele, Gertjan van Noord	UG	Integrated text
3.0	June 2, 2015	Petya Osenova, Kiril Simov	IICT-BAS	Integrated text
3.0	June 15, 2015	Markus Egg	UBER	Integrated reviewer's comments

Statement of originality
This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Report on First pilot version of LRTs enhanced to support Deep Processing

DOCUMENT QTLEAP-2015-D4.7
EC FP7 PROJECT #610516

DELIVERABLE D4.7

completition
FINAL
status
SUBMITTED
dissemination level
PUBLIC

responsible
Kiril Simov (WP4 Coordinator)
reviewers
Markus Egg, Kostadin Cholakov
contributing partners
FCUL, DFKI, IICT-BAS, UBER, UPV-EHU, UG

authors
Petya Osenova, João Silva, Aljoscha Burchardt, Gorka Labaka, Dieke Oele

Contents

List of Abbreviations

BDT	Basque Dependency Treebank
CoNLL	Conference on Natural Language Learning
ERG	English Resource Grammar
HPSG	Head-driven Phrase Structure Grammar
LRTs	Language Resources and Tools
MT	Machine Translation
MWE	Multiword Expressions
NLP	Natural Language Processing
POS	Part-Of-Speech
PDT	Prague Dependency Treebank
SRL	Semantic Role Labeling

1 Introduction

This deliverable describes the language resources to support the deep language processing that are provided within the deliverable D4.6 "First pilot version of language resources and tools (LRTs) enhanced to support robust deep processing". The resources for each language are uploaded to the QTLeap repository. The language resources are of two types:

- **Treebanks.** Syntactically annotated corpora to be used for the training of deep language processing tools and deep (tree-based) machine translation models.
- **Lexicons.** Dictionaries that provide semantic and valency information to support deep language processing.

Please note that there are no tools reported in D4.7 because it focuses on LRTs for Pilot 3 and until the deadline of D4.7 the work towards preparing Pilot 3 was focused mostly on resources, viz., in implementing the plan set up in D1.3.

Also, in D1.3, in the specification of the curation plan of resources (DOW, pp.70-71) it was decided not to spend effort on resources for Czech or Spanish, only for the other project languages.

The resources provided within D4.6 were described in D1.3 "Language resources and tools (LRTs) management plan". The types, size, and number of resources per language were influenced by the different pre-project availability of appropriate data and the degree to which specific languages have been the object of prior research and compilation of linguistic resources. The deliverable D1.3 outlined the following resources and tools to be provided within Deliverable 4.6 and uploaded onto the project repository. The data is presented in tokens:

Language/Resources	M7	M13	M17
Basque			
Basque-Eng ParDeepBank	2.5K tokens	5.0K tokens	6.0K tokens
Bulgarian			
BOL	600	1200	1600
BVFL	900	1800	2400
BulTreeBank-DP	15000	30000	40000
BulEngTreebank	12000	24000	32000
ParDeepBank	5000	10000	15000
Dutch			
Lassy: manually	500/10K	1000/20K	1300/26K
Lassy: automatic	15K/300K	30K/600K	40K/800K
German			
DeepBankDE	2.5K	5.0K	6.0K
Portuguese			
Lexicon (entries)	300	600	750
Corpora (tokens)	20K	40K	50K

Table 1: Language Resources for Supporting Deep Semantic Processing.

For deliverable D4.6, we have to provide the resources that are ready by the milestone M7 — 15 % of the total size of the resources. The resources provided in deliverable D4.6 cover these 15 % for each language. In some cases more data is delivered.

All the data is equipped with metadata record. The data and metadata for all languages are uploaded on the project repository. For each language there is a separate directory — BG (for Bulgarian), DE (for German), EU (for Basque), NL (for Dutch), and PT (for Portuguese). The root directory for WP4 related resources and tools is /Shared/Project/LRT-M11M12/WP4.

The rest of the deliverable presents in greater details the resouces due for and delivered in D4.6. Chapters 2 and 3 are devoted to treebanks and lexicons, respectively.

2 Treebanks

Treebanks are syntactically annotated corpora. Deepbanks add the level of semantics on top of it. It this way the translation models are enhanced in both treebank types: monolingual and parallel.

2.1 Basque

This resource is part of Deliverable 4.6 of the QTLeap FP7 project. In its current development (15% of the intended goal of the project), it consists of 150 sentences (1,416 English tokens and 1,275 Basque tokens).

The English sentences are part of the IULA Penn Treebank English-Spanish parallel corpus[1]. The sentences are excerpts from journalistic text that have been manually translated into Basque to generate a parallel corpus. In this way, we will additionally have access to a trilingual parallel treebank (English-Spanish-Basque).

The selected English sentences were manually translated, and their Basque counterparts analyzed using automatic tools. This analysis includes several levels of linguistic information for each sentence, including lemmatization and morphological analysis as well as dependency parsing trees. After the automatic analysis, a human correction phase was performed.

For English, Stanford dependency tags are used[2], whereas Basque syntactic annotation follows Basque Dependency Treebank (BDT) guidelines [Aldezabal et al., 2009]. It is important to notice that both tagging styles are already included in HamleDT [Rosa et al., 2014]. Therefore, harmonization rules have already been developed and can be used to convert the current resource's analyses into harmonized parses, if needed.

The main motivation behind the creation of this resource was to build a high quality data set with rich grammatical information that could support the development of linguistically-informed translation tools. This treebank can be used both to guide the development of the linguistic analyzers that will be used in translation or to train, in combination with automatically annotated texts, statistical transfer module that will transform source language parses into target language ones.

This resource is uploaded in the directory /Shared/Project/LRT-M11M12/WP4/EU in subdirectory D4.6-ParDeepBank-batch1. The data is represented in CoNLL format for the actual treebank and in text files for the textual form. The metadata is given as XML document and as PDF document in the directory /Shared/Project/LRT-M11M12/WP4/EU. The resource is available for external use on demand.

[1] http://repositori.upf.edu/handle/10230/20049 which has a CC-BY license

[2] For a detailed description see http://nlp.stanford.edu/software/dependencies_manual.pdf

2.2 Bulgarian

2.2.1 BulTreeBank-DP

BulTreeBank-DP is a monolingual dependency CoNLL-based conversion of the original HPSG-based treebank. The treebank consists of 60 % newspaper texts, 30 % literary texts, and 10 % administrative and other texts. It comprises 11,900 sentences, since sentences with ellipses have been left out. The resource complies with the annotation scheme as well as the input requirements, defined for the CoNLL contest on Dependency Parsing in 2006. The treebank includes the following levels of linguistic information: tokenization, POS, morphosyntactic features, dependency relations, coreferences. The resource is very useful for the creation of adequate language models, which are to be used for the Bulgarian part in translation processes.

This resource is uploaded in the directory /Shared/Project/LRT-M11M12/WP4/BG as an archive file D4.6-BulTreeBankDP.zip. The data is represented in CoNLL format. The metadata is given as XML document and as PDF document in the archive. The final version of the resource will be made available via META-SHARE and CLARIN repositories as well as on the website of the Bulgarian partner.

2.2.2 BulEngTreebank

This resource is part of Deliverable 4.6. It contains 920 sentences (9308 tokens) which are part of Bulgarian English Parallel Treebank. It includes English sentences from datasets distributed with the English Resource Grammar (ERG), whose domain is tourism. These sentences have already been analysed using ERG, and manually disambiguated. The sentences were translated into Bulgarian by professional translators.

Bulgarian and English sentences are aligned manually on the word level. Then they were annotated morphologically and parsed by a dependency parser. The result was manually corrected. Within the project this treebank will be extended with more pairs of sentences. Also the Bulgarian part will be represented in other formats: Universal Dependency Tagset, Minimal Recursion Semantics.

This resource is uploaded in the directory /Shared/Project/LRT-M11M12/WP4/BG as an archive file D4.6-BulEngTreebank.zip. The data is represented in CoNLL format. The metadata is given as XML document and as PDF document in the archive. The resource is available for external use on demand.

2.2.3 ParDeepBankBG

This resource is part of Deliverable 4.6. It consists of 838 sentences (21 949 tokens) from the Bulgarian English Parallel Deepbank. It includes English sentences from the English Deepbank, whose domain is journalism (Wall Street Journal). These sentences have already been analysed using ERG, and manually disambiguated. The sentences were translated into Bulgarian by professional translators. Bulgarian and English sentences are aligned manually on the word level. Then they were annotated morphologically and parsed by a dependency parser. Then the result was partially manually corrected. The dependency analyses are represented in CoNLL 2006 format. Within the project this treebank will be extended with more pairs of sentences. Also the Bulgarian part will be represented in other formats: Universal Dependency Tagset, Minimal Recursion Semantics.

This resource is uploaded in directory /Shared/Project/LRT-M11M12/WP4/BG as an archive file D4.6-ParDeepBank.zip. The data is represented in CoNLL format. The

metadata is given as XML document and as PDF document in the archive. The resource is available for external use on demand. P11

2.3 Dutch

This monolingual resource contains 3000 sentences. They are parsed with the Alpino parser for Dutch and converted to Treex a-trees that are used as input for the Treex MT system. The sentences where taken from the Dutch part of the parallel OPUS-KDE4corpus, the manual of KDE which is a Windowing Manager and Graphical User Interface for the UNIX operating system. These conversion will ultimately be used for the translation of Dutch sentences within the Treex pipeline as it is planned for Pilot 1.

This resource is uploaded in the directory /Shared/Project/LRT-M11M12/WP4/NL as an archive file KDE.tar. The data is represented in XML format according to Treex Schema. The metadata is given as XML document and as PDF document in the directory. The resource is available for external use on demand.

2.4 German

The corpus currently contains the first batch of 4600 sentences taken from the German TIGER treebank[3] that have been parsed using the Cheetah grammar for German (Cramer [2011]) and the PET parser. The corpus consists of files containing Trees and MRSs. STTS tags from the original TIGER corpus are preserved in the Derivation Tree. The corpus contains 40.000 tokens (4.6K, size 17 MB).

It is planned to also experiment with an alternative German HPSG grammar and compare the analyses w.r.t. to the needs of the project. If the results are satisfactory, they will also be added to the project repository. Manual editing and selection will only be performed if relevant for the MT development within the project.

This resource is uploaded in the directory /Shared/Project/LRT-M11M12/WP4/DE as an archive file DeepBankDE-batch1.zip[4]. The data is represented in ERG text format. The metadata is given as XML document and as PDF document in the archive. The resource can only be accessed if the requesting institution has a license for the original Tiger treebank.

2.5 Portuguese

This resource is part of Deliverable 4.6. It is composed of 3,134 sentences (36,566 tokens) taken from CINTIL-DeepBank. The sentences are excerpts from journalistic text taken from CETEMPúblico.

Each sentence is annotated with several levels of linguistic information, including its derivation tree obtained during parsing, its syntactic constituency tree, different renderings of MRS based representations of its meaning (Copestake et al. [2005]), and its fully-fledged grammatical representation in AVM format.

This annotation is the result of a semi-automatic process, where an automatic analysis performed by a deep computational grammar is followed by a stage of double-blind annotation with adjudication (see Branco and Costa [2008], for a full description of the process) that prunes the parse forest produced by the grammar down to a single analysis.

[3] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html
[4] Batch1 will be included in future Batch2, etc.

The main motivation behind the creation of this resource was to build a high quality data set with rich grammatical information that could support the development of a large set of high level language resources and processing tools for Portuguese.

This resource is uploaded in the directory /Shared/Project/LRT-M11M12/WP4/PT as an archive file D4.6_deepbank.zip. The data is represented in ERG export format. The metadata is given as a XML document and as a PDF document in the archive.

This resource corresponds to folders 1721–2420 of the eCTMP sub-corpus of CINTIL-DeepBank, available through META-SHARE.[5]

3 Lexicons

Lexicons are rich lexical databases, which include various information, such as WordNet synsets, valence frames, ontological classes. etc.

3.1 Bulgarian

3.1.1 Bulgarian Ontology-Based Lexicon

The Bulgarian Ontology-based Lexicon is organized in synsets as WordNets, but the relations between the synsets are represented via mapping to different semantic resources. The goal is for them to be mapped to an appropriate ontology. In the version provided here the mapping is to the English Princeton WordNet 3.0 (WN3.0). Other mappings exist to DOLCE ontology done via OntoWordNet and via WN3.0 to SUMO ontology.

This resource is uploaded in directory /Shared/Project/LRT-M11M12/WP4/BG as an archive file D4.6-BOL.zip. The data is represented in table format, WordNet-LMF and LEMON encoding. The metadata is given as XML document and as PDF document in the archive. The resource is available for external use on demand. The resource is freely available via Open Multilingual Wordnet[6].

3.1.2 Bulgarian Valency Frame Lexicon

The Valency Lexicon is a treebank-driven resource of extracted valency frames from BulTreeBank. The frames were manually curated. At the moment it comprises more than 1000 verb frames. The frames follow the surface representation in the sentences. The frame roles (or participants in the corresponding event) were assigned ontological constraints from the SIMPLE ontology (translated into Bulgarian), such as ARTEFACT, COGNITIVE FACT, etc. The representation of an entry is as follows:

<FD>
<lemma></lemma>
<def></def>
<F><F>
</FD>

where FD = Frame Description, lemma = lemma, def = definition, and F = Frame.

This resource is uploaded in the directory /Shared/Project/LRT-M11M12/WP4/BG as an archive file D4.6-BVFL.zip. The data is represented in XML format as described

[5] http://metashare.metanet4u.eu/ (search for "CINTIL-DeepBank").
[6] http://compling.hss.ntu.edu.sg

above. The metadata is given as XML document and as PDF document in the archive. The resource is available for external use on demand.

3.2 Portuguese

This resource is part of Deliverable 4.6. It comprises 600 lexicon entries used in LXGram, an HPSG computational grammar for deep linguistic processing of Portuguese. The lexicon was built manually. Each lexical entry is associated with a deep lexical type which is part of the type hierarchy defined in the grammar (the types associated with the 600 lexicon entries are also included in the deliverable). The deep lexical type encodes a great deal of information about the grammatical behavior of the word, such as its part-of-speech, subcategorization (valence) frame, the pattern for forming anticausative alternations, whether a verb is a raising verb or not, etc.

This resource is uploaded in the directory /Shared/Project/LRT-M11M12/WP4/PT as an archive file D4.6_lexicon.zip. The data is represented in LXGram format. The metadata is given as a XML document and as a PDF document in the archive.

4 Conclusion

In this deliverable we describe two types of language resources to support the deep language processing in the project, viz., treebanks and lexicons. The resources for the different languages are provided within the deliverable D4.6.

References

Izaskun Aldezabal, Maria Jesus Aranzabe, Jose Mari Arriola, and Arantza Diaz de Ilarraza. Syntactic annotation in the reference corpus for the processing of basque (epec): Theoretical and practical issues. *Corpus Linguistics and Linguistic Theory*, 5(2):241–269, 2009.

Antonia Branco and Francisco Costa. LXGram in the Shared Task "Comparing Semantic Representations" of STEP 2008. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 299–314. College Publications, 2008. URL http://www.aclweb.org/anthology/W08-2224.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan Sag. Minimal recursion semantics: An introduction. *Research on Language & Computation*, 3(4):281–332, 2005.

Bart Cramer. *Improving the feasibility of precision-oriented HPSG parsing*. PhD thesis, Universit, 2011.

Rudolf Rosa, Jan Mašek, David Mareček, Martin Popel, Daniel Zeman, and Zdeněk Žabokrtský. HamleDT 2.0: Thirty dependency treebanks stanfordized. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, and Joseph Mariani, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2334–2341, Reykjavík, Iceland, 2014. European Language Resources Association. ISBN 978-2-9517408-8-4.

A Summary of availability

Name of LRT	language	QTLeap	License	URL
Basque-English ParDeepBank	EU, EN	Yes	CC BY 3.0	http://metashare.metanet4u.eu/go2/Basque-English-ParDeepBank
Bulgarian Ontology-based Lexicon	BG	Yes	CC BY 3.0	The final corpus will be available from http://www.bultreebank.org/QTLeap/
Bulgarian Valency Frame Lexicon	BG	Yes	CC-BY-NC-SA v4.0	The final corpus will be available from http://www.bultreebank.org/QTLeap/
BulTreeBank-DP	BG	Yes	CC-BY-NC-SA v4.0	The final corpus will be available from http://www.bultreebank.org/QTLeap/
Bulgarian English Parallel Treebank	BG, EN	Yes	CC-BY-NC-SA v4.0	The final corpus will be available from http://www.bultreebank.org/QTLeap/
Bulgarian English Deep Bank	BG, EN	Yes	CC-BY-NC-SA v4.0	The final corpus will be available from http://www.bultreebank.org/QTLeap/
Dutch Tree Bank	NL	Yes	Free on demand	g.j.m.van.noord@rug.nl
German Deep Bank	DE	Yes	CC BY NC SA[a]	The final corpus will be available from http://metashare.dfki.de/.
Portuguese Deep Bank	PT	Yes	MS-NC	http://metashare.metanet4u.eu/
Portuguese Lexicon	PT	Yes	ELDA	The final lexicon will be available from http://metashare.metanet4u.eu/

Table 2: Summary of publicly available LRTs mentioned in this deliverable. QTLeap column indicates with "yes" those LRTs which have been (partially) funded by QTLeap.

[a]Usage of the corpus requires that the user has the TIGER corpus license that can be acquired online: http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/TIGERCorpus/license/htmlicense.html

B Appendix B: Narrative Description of D4.6 Language Resources

In this appendix we present the narrative descriptions for each language resource uploaded in the repository for D4.6.

B.1 Basque-English ParDeepBank

ENGLISH-BASQUE PARALLEL DEEPBANK

1 BASIC INFORMATION

1.1 Corpus composition

This resource is part of Deliverable 4.6 of the QTLeap FP7 project (Contract number 610516). In its current development (15% of the intended goal of the project), it is composed of 150 sentences (1,416 English tokens and 1,275 Basque tokens). The sentences are excerpts from journalistic text from the Wall Street Journal that have been manually translated into Basque to generate a parallel corpus.

It includes several levels of linguisic information for each sentence, including lemmatization and mophological analysis as well as dependency parsing trees. This is the result of a semi-automatic annotation process by means of automatic analysis followed by a human correction phase.

The main motivation behind the creation of this resource was to build a high quality data set with rich grammatical information that could support the development of linguistically-informed translation tools.

1.2 Representation of the corpora (flat files, database, markup)

The corpus is stored in 4 separate files, two per language. All of these are plain text files.

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Eneko Agirre
Affiliation: University of the Basque Country
Telephone: +34 943 015019
e-mail: e.agirre@ehu.es

Name: Gorka Labaka
Affiliation: University of the Basque Country
Telephone: +34 943 018307
e-mail: gorka.labaka@ehu.es

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be stored in the QTLeap repository. Available for external use on demand.

2.3 Copyright statement and information on IPR

The resource is free, license-based, for research purposes and fee license-based for commercial purposes.

3 TECHNICAL INFORMATION

3.1 Directories and files

Two files for each language (English and Basque): one file contains the original text (.text extension) and the second file contains the linguistic analyses (.conll extension).

3.2 Data structure of an entry

Dependency parsing annotated on CoNLL format.

1	He	he	PRP	PRP	_	2	nsubj		
2	earned	earn	VBD	VBD	_	0	root		
3	his	his	PRP$	PRP$	_	4	poss		
4	doctorate		doctorate		NN	NN	_	2	dobj
5	in	in	IN	IN	_	4	prep		
6	nuclear	nuclear	JJ	JJ	_	7	amod		
7	physics	physics	NN	NN	_	5	pobj		
8	from	from	IN	IN	syn=CLR		2	prep	
9	the	the	DT	DT	_	11	det		
10	Massachusetts		massachusetts		NNP	NNP	_	11	nn
11	Institute	institute	NNP	NNP	_	8	pobj		
12	of	of	IN	IN	_	11	prep		
13	Technology		technology		NNP	NNP	_	12	pobj
14	.	.	.	.	_	2	punct		

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

The corpus contains 150 parallel sentences, making up a total of 2691 tokens (1416 English tokens and 1275 Basque tokens) and needs about 150 KB of disk storage.

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is multilingual and parallel. Annotated at dependency parsing level.

4.2 The natural language(s) of the corpus

English and Basque.

4. 3 Domain(s)/register(s) of the corpus

English sentences extracted from Wall Street Journal (WSJ) corpus and translated into Basque.

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)
Text annotated up to dependency parsing, including sentence splitting, tokenization, morphological analysis and dependency parsing.

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),
Stanford dependency tags used for English. For a detailed description see http://nlp.stanford.edu/software/dependencies_manual.pdf

Basque annotation follows BDT guidelines (Aranzabal et al., 2009). BDT is already included in HamleDT. Therefore, harmonization rules are already developed and can be used to convert the current resource's analyses into harmonized parses, if needed.

4.4.3 Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)
Sentence level alignment ensured at translation. Basque corpus was created by means of human translation of English sentences.

4.4.4 Attributes and their values (if annotated)
Not applicable

4.5 Intended application of the corpus
Training data for Machine Translation applications

4.6 Reliability of the annotations (automatically/manually assigned) – if any
Automatically assigned and manually revised annotations.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Aldezabal I., Aranzabe M., Arriola J., Díaz de Ilarraza A. 2009. ”Syntactic annotation in the Reference Corpus for the Processing of Basque (EPEC): Theoretical and practical issues”. *Corpus Linguistics and Linguistic Theory 5-2 (2009), 241-269. Mouton de Gruyter. Berlin-New York. Print ISSN: 1613-7027 Online ISSN: 1613-7035*

B.2 Bulgarian Ontology-based Lexicon: BOL

P20

D4.6: BULGARIAN ONTOLOGY-BASED LEXICON (BOL)

LEXICA DOCUMENTATION

1. BASIC INFORMATION

   1.1 Lexicon type: **The BulTreeBank WordNet (BTB-WN)**
   1.2 Representation of the lexicon: ***markup***
   1.3 Character encoding: ***UTF-8***

2. ADMINISTRATIVE INFORMATION

   2.1 Contact person:

   Name: ***Kiril Simov***
   Address: ***25A Acad. Bonchev Str., Sofia, 1113, Bulgaria***
   Affiliation: ***IICT-BAS, Linguistic Modeling Department***
   Position: ***associate professor, PhD***
   Mobile: ***(00359) 888 473 413***
   Email: ***kivs@bultreebank.org***

   2.2 Delivery medium (if relevant; description of the content of each piece of medium)
   2.3 Copyright statement and information on IPR:

The BulTreeBank WordNet (BTB-WN)
==========

Copyright (c) 2014 BulTreeBank Group

Contacts:
Kiril Simov (kivs@bultreebank.org)
Petya Osenova (petya@bultreebank.org)

Licence

The The BulTreeBank WordNet (BTB-WN) is distributed under the following licence: CC BY 3.0

For informal information please see here:
http://creativecommons.org/licenses/by/3.0/
For formal representation please see here:
http://creativecommons.org/licenses/by/3.0/legalcode

3. TECHNICAL INFORMATION

   3.1 Directories and files: ***in WordLMF XML, Lemon XML, and Table Text Style***
   3.2 Data structure of an entry:
   3.3 Lexicon size (nmb. of lexical items, KB occupied on disk): ***2 091 765 bytes in table format***

4. CONTENT INFORMATION

The Bulgarian Ontology-based Lexicon is organized in synsets as WordNets, but the relations between the synsets are represented via mapping to different resources. In the version provided here the mapping is to English Princeton WordNet 3.0 (WN3.0). Other mappings exist to DOLCE ontology done via OntoWordNet and via WN3.0 to SUMO ontology and other semantic resources to be presented in D5.4.

This version is made freely available via Open Multilingual Wordnet http://compling.hss.ntu.edu.sg.

4.1 The natural language(s) of the lexicon: ***Bulgarian***
4. 2 Entry Type:

In text table view the first column contains the WN3.0 id, the next column represents the type of information: lemma, definition, and example and the last column represents the actual value. For definitions and examples there are numbers because there could be more than one of them.

00007846-n	***bul:lemma***	***индивид***
00007846-n	***bul:lemma***	***лице***
00007846-n	***bul:lemma***	***личност***
00007846-n	***bul:lemma***	***особа***
00007846-n	***bul:def 0***	***Отделен човек, който със своите неповторими качества се отличава, различава от другите хора.***
00007846-n	***bul:exe 0***	***високопоставена особа***
00007846-n	***bul:exe 1***	***ерудирана личност***
00007846-n	***bul:exe 2***	***запомняща се личност***
00007846-n	***bul:exe 3***	***известна личност***
00007846-n	***bul:exe 4***	***индивиди от различни поколения.***

4.3 Attributes and their values: ***definition, lemma, example***
4.4 Coverage of the lexicon:

- 4,999 ***synsets***
- 6,783 words
- 9,056 senses
- Covers the Core WordNet in 100 %

4.5 Intended application of the lexicon: ***Sense Annotation and Machine Translation applications***
4.6 POS assignment: ***yes***
4.7 Reliability (automatically/manually constructed): ***automatically mapped and manually curated***

5. RELEVANT REFERENCES AND OTHER INFORMATION

@InProceedings{Simov:Osenova:2010,
author = {Kiril Simov and Petya Osenova},
title = {Constructing of an Ontology-based Lexicon for Bulgarian},
booktitle = {Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)},
year = {2010},
month = {may},
date = {19-21},
address = {Valletta, Malta},
editor = {Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Bente Maegaard and Joseph Mariani and Jan Odijk and Stelios Piperidis and Mike Rosner and Daniel Tapias},
publisher = {European Language Resources Association (ELRA)},
isbn = {2-9517408-6-7},
language = {english}
url = http://www.lrec-conf.org/proceedings/lrec2010/summaries/848.html
}

B.3 Bulgarian Valency Frame Lexicon: BVFL

D4.6: BULGARIAN VALENCY FRAME LEXICON (BVFL)

LEXICA DOCUMENTATION

1. BASIC INFORMATION

1.1 Lexicon type: **Bulgarian Valency Frame Lexicon (BVFL)**
1.2 Representation of the lexicon: ***markup***
1.3 Character encoding: ***UTF-8***

2. ADMINISTRATIVE INFORMATION

2.1 Contact person:

Name: ***Kiril Simov***
Address: ***25A Acad. Bonchev Str., Sofia, 1113, Bulgaria***
Affiliation: ***IICT-BAS, Linguistic Modeling Department***
Position: ***associate professor, PhD***
Mobile: ***(00359) 888 473 413***
Email: ***kivs@bultreebank.org***

2.2 Delivery medium (if relevant; description of the content of each piece of medium)
2.3 Copyright statement and information on IPR:

Bulgarian Valency Frame Lexicon (BVFL)
==========

Contacts:
Kiril Simov (kivs@bultreebank.org)
Petya Osenova (petya@bultreebank.org)

License
---

3. TECHNICAL INFORMATION

3.1 Directories and files: ***in XML***
3.2 Data structure of an entry:
3.3 Lexicon size (nmb. of lexical items, KB occupied on disk):

4. CONTENT INFORMATION

The Valency Lexicon is a treebank-driven resource of extracted valency frames from BulTreeBank. The frames were manually curated. The frames followed the surface representation in the sentences. The frame participants were assigned ontological constraints from SIMPLE ontology (translated into Bulgarian), such as ARTEFACT, COGNITIVE FACT, etc.

4.1 The natural language(s) of the lexicon: ***Bulgarian***

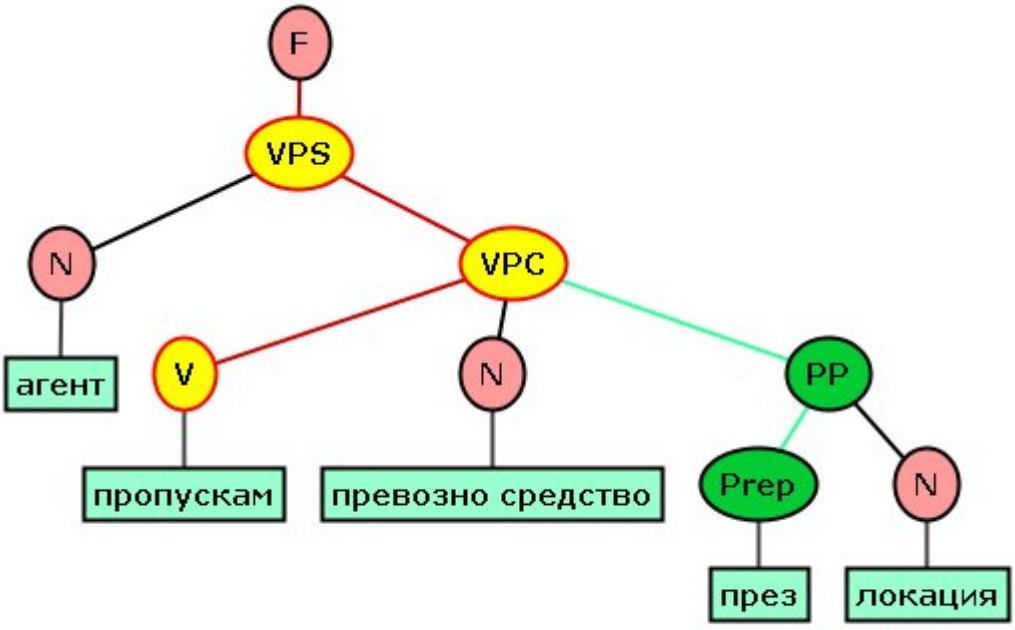
4. 2 Entry Type:

<FD>
<lemma></lemma>
<def></def>
<F><F>

</FD>

FD = Frame Description
lemma = lemma
def = definition
F = Frame

Example: AGENT let pass VEHICLE through LOCATION



The structure of the frame follows the BulTreeBank syntactic structure (more in the BulTreeBank Stylebook - http://www.bultreebank.org/TechRep/BTB-TR04.pdf). Please note that the leaves here do not encode words but concepts from SIMPLE Core Ontology. In the phase of the project these concepts will be aligned to the concepts in other ontologies, such as DOLCE.

4.3 Attributes and their values:

4.4 Coverage of the lexicon: ***1047 lexical entries***

4.5 Intended application of the lexicon: ***Sense Annotation and Machine Translation applications***

4.6 POS assignment: ***yes***

4.7 Reliability (automatically/manually constructed): ***automatically extracted verb frames and manually curated***

5. RELEVANT REFERENCES AND OTHER INFORMATION

Osenova et. al 2012: Petya Osenova, Kiril Simov, Laska Laskova and Stanislava Kancheva. A *Treebank-driven Creation of an OntoValence Verb lexicon for Bulgarian*. In: Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk and Stelios Piperidis (eds.) Proceedings of LREC'12, Istanbul, Turkey. ELRA. 978-2-9517408-7-7, pp. 2636-2640.

B.4 Dependency part of BulTreeBank: BulTreeBank-DP

D4.6: BulTreeBank-DP

1 BASIC INFORMATION

1.1 Corpus composition: ***BulTreeBank-DP is a dependency CoNLL-based conversion of the original HPSG-based treebank***.

1.2 Representation of the corpora (flat files, database, markup): ***CoNLL 2006 table text format***

1.3 Character encoding: ***UTF-8***

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: ***Kiril Simov***
Address: ***25A Acad. Bonchev Str., Sofia, 1113, Bulgaria***
Affiliation: ***IICT-BAS, Linguistic Modeling Department***
Position: ***associate professor, PhD***
Mobile: ***(00359) 888 473 413***
Email: ***kivs@bultreebank.org***

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

2.3 Copyright statement and information on IPR

BulTreeBank-DP
==========

Contacts:
Kiril Simov (kivs@bultreebank.org)
Petya Osenova (petya@bultreebank.org)

License: freely available for research purposes

3 TECHNICAL INFORMATION

3.1 Directories and files: ***38 files corresponding to their text sources***

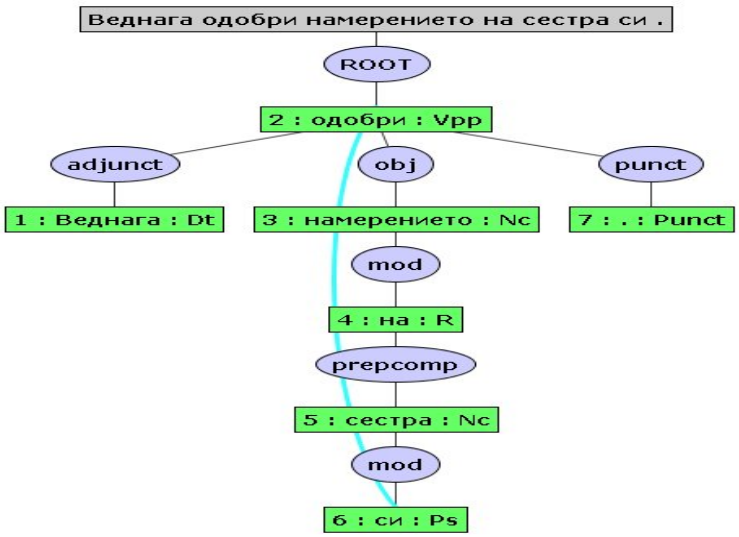
3.2 Data structure of an entry:

- ***CoNLL 2006 data format***

Here is the information:
Column 1: wordform number in sentences
Column 2: wordform
Column 3: lemma
Column 4: only POS
Column 5: coarse POS tag
Column 6: morphosyntactic characteristics
Columns 7, 9: information about the head
Columns 8, 10: grammatical relations
Column 11: complete POS tag
Column 12: information about coreferences

1	Като	като	R	R	_	7	adjunct	7	adjunct	R	-
2	всички	всеки	P	Pc	ref=e|case=n|num=p	4	mod	4	mod	Pce-op	-
3	умни	умен	A	A	num=p|def=i	4	mod	4	mod	A-pi	-
4	хора	хора	N	Nc	num=pl_tantum|def=i	1	prepcomp	1	prepcomp	Nc-li	-
5	аз	аз	P	Pp	ref=e|case=n|num=s|pers=1	7	subj	7	subj	Ppe-os1	5
6	не	не	T	Tn	type=neg	7	mod	7	mod	Tn	-
7	ритам	ритам	V	Vpi	trans=t|mood=i|tense=r|pers=1|num=s	0	ROOT	0	ROOT	Vpitf-r1s	-
8	тръна	трън	N	Nc	gen=m|num=s|def=h	7	obj	7	obj	Ncmsh	-
9	бос	бос	A	Am	gen=m|num=s|def=i	7	adjunct	7	adjunct	Amsi	5
10	.	.	Punct	Punct	_	7	punct	7	punct	punct	-

- ***Graphical view, generated from the CoNLL presentation***



3.3 Corpora size (nmb. of tokens, MB occupied on disk): ***196 000 tokens; 12,3 Mb***

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated): ***monolingual and syntactically annotated with dependency structures***

4.2 The natural language(s) of the corpus: ***Bulgarian***

4. 3 Domain(s)/register(s) of the corpus: ***news media(70%), literature(25%), administrative documents(5%).***

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up): ***morphological-mark-up; syntactic mark-up; sentence-internal coreferences mark-up***

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed): ***POS, grammatical features, grammatical roles; tagged and parsed.***

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

4.4.4 Attributes and their values (if annotated):

The morphological mark-up follows the BulTreeBank positional Tagset (see the reference below). Here is an example: Npmsi- N-noun, p-proper, m-masculine, s-singular, i-indefinite.

The syntactic mark-up follows the CoNLL dependency relations tagset.

4.5 Intended application of the corpus: ***as gold standard for training parsers for Bulgarian; for typological comparison of syntactic structures with other languages***

4.6 Reliability of the annotations (automatically/manually assigned) ***– the conversion to dependency format was done automatically, but then the sentences were manually curated***

5 RELEVANT REFERENCES AND OTHER INFORMATION

1.BulTreeBank Morphosyntactic Tagset: http://www.bultreebank.org/TechRep/BTB-TR03.pdf
2.A short description of the Dependency Part of BulTreeBank
(BulTreeBank-DP): http://www.bultreebank.org/dpbtb/

B.5 Bulgarian English Parallel Treebank: BulEngTreebank

P28

D4.6-BulEngTreebank

1 BASIC INFORMATION

1.1 Corpus composition:

This resource is part of Deliverable 4.6. It is composed of 920 sentences (9308 tokens) which are part of Bulgarian English Parallel Treebank. It includes English sentences from datasets distributed with English Resource Grammar (ERG). These sentences are already analysed using ERG and manually disambiguated. The sentences are translated into Bulgarian by professional translators.
Bulgarian sentences are aligned manually to English on word level. Then they are morphologically annotated and parsed by a dependency parser. Then the result is manually corrected.
Within the project this treebank will be extended with more pairs of sentences. Also the Bulgarian part will be represented in other formats: Universal Dependency Tagset, Minimal Recursion Semantics.

Character encoding: ***UTF-8***

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: ***Kiril Simov***
Address: ***25A Acad. Bonchev Str., Sofia, 1113, Bulgaria***
Affiliation: ***IICT-BAS, Linguistic Modeling Department***
Position: ***associate professor, PhD***
Mobile: ***(00359) 888 473 413***
Email: ***kivs@bultreebank.org***

2.2 Delivery medium (if relevant; description of the content of each piece of medium)
2.3 Copyright statement and information on IPR

D4.6-BulEngTreebank

==========

Contacts:
Kiril Simov (kivs@bultreebank.org)
Petya Osenova (petya@bultreebank.org)

3 TECHNICAL INFORMATION

3.1 Directories and files: ***4 files***
3.2 Data structure of an entry:
- ***CoNLL 2006 data format***

Here is the information:
Column 1: wordform number in sentence

Column 2: wordform
Column 3: lemma
Column 4: only POS
Column 5: coarse POS tag
Column 6: morphosyntactic characteristics
Columns 7: information about the head
Columns 8: grammatical relations

3.3 Corpora size (nmb. of tokens, MB occupied on disk): **920 sentences; 9308 tokens**.

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated): ***bilingual BG-EN syntactically annotated corpus***

4.2 The natural language(s) of the corpus: ***English***

4. 3 Domain(s)/register(s) of the corpus: ***Tourism and Linguistic examples in the CSLI dataset***

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up): ***morphological-mark-up; syntactic mark-up***

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed): ***POS, grammatical features, grammatical roles; tagged and parsed.***

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved): ***The alignment was done manually on word level.***

4.4.4 Attributes and their values (if annotated):
For the Bulgarian part:

The morphological mark-up follows the BulTreeBank positional Tagset (see the reference below). Here is an example: Npmsi- N-noun, p-proper, m-masculine, s-singular, i-indefinite.
The syntactic mark-up follows the CoNLL dependency relations tagset.

For the English part: the ERG resource grammar was used for tagging, parsing and semantics.

4.5 Intended application of the corpus: ***as gold standard for the purposes of Machine Translation from Bulgarian to English and backwards;***

4.6 Reliability of the annotations (automatically/manually assigned) ***– The alignments were done manually on word level. The processing of both corpora was done automatically.***

5 RELEVANT REFERENCES AND OTHER INFORMATION

1.BulTreeBank Morphosyntactic Tagset: http://www.bultreebank.org/TechRep/BTB-TR03.pdf
2.A short description of the Dependency Part of BulTreeBank (BulTreeBank-DP): http://www.bultreebank.org/dpbtb/
3.English Resource Grammar: http://www.delph-in.net/erg/

B.6 Bulgarian English Deep Bank: ParDeepBank

P30

D4.6-ParDeepBankBG

1 BASIC INFORMATION

1.1 Corpus composition:

This resource is part of Deliverable 4.6. It is composed of 838 sentences (21 949 tokens) which are part of Bulgarian English Parallel Deepbank. It includes English sentences from English Deepbank. These sentences are already analysed using ERG and manually disambiguated. The sentences are translated into Bulgarian by professional translators. Bulgarian sentences are aligned manually to English on word level. Then they are morphologically annotated and parsed by a dependency parser. Then the result is partially manually corrected. The dependency analyses are represented in CoNLL 2006 format. Within the project this treebank will be extended with more pairs of sentences. Also the Bulgarian part will be represented in other formats: Universal Dependency Tagset, Minimal Recursion Semantics.

Character encoding: ***UTF-8***

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: ***Kiril Simov***
Address: ***25A Acad. Bonchev Str., Sofia, 1113, Bulgaria***
Affiliation: ***IICT-BAS, Linguistic Modeling Department***
Position: ***associate professor, PhD***
Mobile: ***(00359) 888 473 413***
Email: ***kivs@bultreebank.org***

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

2.3 Copyright statement and information on IPR

D4.6-ParDeepBankBG

==========

Contacts:
Kiril Simov (kivs@bultreebank.org)
Petya Osenova (petya@bultreebank.org)

3 TECHNICAL INFORMATION

3.1 Directories and files: ***838 sentences***

3.2 Data structure of an entry:

- ***CoNLL 2006 data format***

Here is the information:
Column 1: wordform number in sentence
Column 2: wordform
Column 3: lemma

Column 4: only POS
Column 5: coarse POS tag
Column 6: morphosyntactic characteristics
Columns 7: information about the head
Columns 8: grammatical relations

3.3 Corpora size (nmb. of tokens, MB occupied on disk): 21 949 tokens

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated): ***bilingual BG-EN syntactically annotated corpus***

4.2 The natural language(s) of the corpus: ***English***

4. 3 Domain(s)/register(s) of the corpus: ***Wall Street Journal***

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up): ***morphological-mark-up; syntactic mark-up***

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed): ***POS, grammatical features, grammatical roles; tagged and parsed.***

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved): ***The alignment was done manually on word level.***

4.4.4 Attributes and their values (if annotated):
For the Bulgarian part:

The morphological mark-up follows the BulTreeBank positional Tagset (see the reference below). Here is an example: Npmsi- N-noun, p-proper, m-masculine, s-singular, i-indefinite.
The syntactic mark-up follows the CoNLL dependency relations tagset.

For the English part: the ERG resource grammar was used for tagging, parsing and semantics.

4.5 Intended application of the corpus: ***as gold standard for the purposes of Machine Translation from Bulgarian to English and backwards;***

4.6 Reliability of the annotations (automatically/manually assigned) ***– The alignments were done manually on word level. The processing of both corpora was done automatically.***

5 RELEVANT REFERENCES AND OTHER INFORMATION

1.BulTreeBank Morphosyntactic Tagset: http://www.bultreebank.org/TechRep/BTB-TR03.pdf
2.A short description of the Dependency Part of BulTreeBank (BulTreeBank-DP): http://www.bultreebank.org/dpbtb/
3.English Resource Grammar: http://www.delph-in.net/erg/

B.7 Dutch Tree Bank

Template for Description of Language Resources in QTLeap Project

1 Language Resources for Language

1.1 Alpino2Treex Dutch

1.1.1 Basic Information:

This database contains parses created with Alpino [3] and their conversions to Treex [1] a-trees.

The parsed sentences where taken from the Dutch part of the paralell OPUS-KDE4corpus [2], the manual of KDE which is a Windowing Manager and Graphical User Interface for the UNIX operating system.

Representation of the resource:
Database

Character encoding:
UTF-8

1.1.2 Administrative Information

Contact person:
prof.dr. Gertjan van Noord:
g.j.m.van.noord@rug.nl
0503637811

Dieke Oele:
d.oele@rug.nl
0503635858

1.1.3 Technical Information

Directories and files:
Alpino parses: .xml
Treex: .treex

Resource size:
Sentences: 3500
Words: 56270
162MB in .tar

1.1.4 Content Information

Type of the corpus:
Monolingual (extracted from the Dutch part of a parallel corpus

The natural language(s) of the corpus:
Dutch

Domain of the corpus:
IT-domain

Intended application of the resource in the project:
Training data for Machine Translation applications

Reliability of the annotations:
Automatically assigned

References

[1] Martin Popel and Zdeněk Žabokrtský. Tectomt: Modular nlp framework. In *Proceedings of the 7th International Conference on Advances in Natural Language Processing*, IceTAL'10, pages 293–304, Berlin, Heidelberg, 2010. Springer-Verlag.

[2] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[3] Gertjan van Noord. **A**t **L**ast **P**arsing **I**s **N**ow **O**perational. In *TALN 2006 Verbum Ex Machina, Actes De La 13e Conference sur Le Traitement Automatique des Langues naturelles*, pages 20–42, Leuven, 2006.

B.8 German Deep Bank

DeepBankDE for QTLeapProject – batch 1

1 BASIC INFORMATION

1.1 Corpus composition

The corpus currently contains the first batch of 4600 sentences taken from the German TIGER treebank[1] that have been parsed using the Cheetah grammar for German (Cramer 2011) using the PET parser. As requested by the EC, existing resources have been re-used this to produce the first pilot data. Manual editing and selection will only be performed if it will become relevant for MT development within the project.

The second batch is in preparation.

We also plan to do use an alternative German grammar and compare the analyses w.r.t. to the need of the project. If the results are suitable, we will add them to the repository in addition.

On this basis, the documentation will be extended with technical and content information.

1.2 Representation of the corpora (flat files, database, markup)

The corpus consist of files containing Trees and MRS.

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Aljoscha Burchardt,
Address: Alt-Moabit 91c, 10559 Berlin
Affiliation: German Research Center for Artificial Intelligence
Position: Director
Telephone: +49 30 23895 1800
Fax: +49 30 23895 1810
e-mail: Aljoscha.Burchardt@dfki.de

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

[1] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html

The first batch will be available in the project intranet.

2.3 Copyright statement and information on IPR

The resource can only be accessed if the requesting institution has a license for the original Tiger treebank.

3 TECHNICAL INFORMATION

3.1 Directories and files

The directory contains one file with all information.

3.2 Data structure of an entry

TAB-delimited fields: ID, Derivation Tree, Tree, MRS.

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

The corpus contains 40.000 tokens (4.6K, size 17 MB) (130.000 token for 15k corpus, size 60 MB).

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

Monolingual/annotated

4.2 The natural language(s) of the corpus

German

4. 3 Domain(s)/register(s) of the corpus

News

4.4 Annotations in the corpus (if an annotated corpus) [2]

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

Syntactic and semantic representations of sentences and phrases.

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed)

[2] More information will be provided together with batch 2.

STTS tags from the original TIGER corpus are preserved in the Derivation Tree.

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

Not relevant

4.4.4 Attributes and their values (if annotated)

4.5 Intended application of the corpus

Research and Development

4.6 Reliability of the annotations (automatically/manually assigned) – if any

The annotations are highly reliable.

5 RELEVANT REFERENCES AND OTHER INFORMATION

Ann Copestake, Dan Flickinger, Carl Pollard and Ivan A. Sag. 2005. *Minimal Recursion Semantics: An Introduction.* Research on Language and Computation, Springer*, 3*, 281–332

Bart Cramer. 2011. *Improving the feasibility of precision-oriented HPSG parsing*. PhD thesis, Universität des Saarlandes.

Bart Cramer and Yi Zhang. 2009. *Constructon of a German HPSG grammar from a detailed treebank*. In: Proceedings of the ACL 2009 Grammar Engineering across Frameworks workshop, pages 37-45, Singapore, Singapore.

Bart Cramer and Yi Zhang. 2010. *Constraining robust constructions for broad-coverage parsing with precision grammars*. In: Proceedings of COLING-2010.

B.9 Portuguese Deep Bank

Deliverable 4.6-DeepBank 1.1

I. Basic Information

1.1. Corpus information

This resource is part of Deliverable 4.6. It is composed of 3,134 sentences (36,566 tokens) which are part of CINTIL-DeepBank (available in the META-SHARE repository). The sentences are excerpts from journalistic text from CETEMPúblico.

It includes several levels of information for each sentence, including its derivation tree originated during parsing, its syntactic constituency tree, different renderings of MRS based representations of its meaning (Copestake, 2006), and its fully-fledged grammatical representation in AVM format. This is the result of a semi-automatic annotation process by means of automatic analysis by the grammar followed by a double-blind annotation followed by adjudication (see (Branco and Costa, 2008), for a full description of the process).

The main motivation behind the creation of this resource was to build a high quality data set with rich grammatical information that could support the development of a large set of high level language resources and processing tools for Portuguese.

1.2. Representation of the corpora (flat files, database, markup)

The corpus is stored in an archive composed by 694 folders. Each folder contains several files, one per sentence. These are plain text files, compressed with gzip.

1.3. Character encoding

The files are encoded in UTF-8.

II. Administrative Information

2.1. Contact person

Name: António Branco
Address: Departamento de Informática NLX - Grupo de Fala e Linguagem Natural, Faculdade de Ciências da Universidade de Lisboa, Edifício C6, Campo Grande 1749-016 Lisboa
Position: Associate Professor
Affiliation: Faculty of Sciences, University of Lisbon
Telephone: +351 217 500 087
Fax: +351 217 500 084
E-mail: antonio.branco@di.fc.ul.pt

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

This resource is available through META-SHARE.

2.3. Copyright statement and information on IPR

This resource is available for both research and commercial purposes, with attribution required, and no redistribution nor derivatives allowed. It is available through META-SHARE.

3.1. Directories and files

The archive that can be downloaded on the META-SHARE site is a gzip file with 2711 folders. Each file contains one gzip file per sentence.

3.2. Data structure of an entry

There is a file for each sentence. The file starts with a line at the top with the sentence id (between square brackets), followed by the sentence between quote marks in raw text. Under this there are a variety of analysis of the sentence, separated by a blank line, as illustrated by the example below:

```
[11] (1 of 1) {1} `a criança obedece apenas a a mãe.' []
```

Derivation:

```
(469 ROOT 4.95827e+17 0 7
 (468 SUBJECT-HEAD 3.84742e+17 0 7
  (461 FUNCTOR-HEAD-HCOMPS-SCOPAL -2.2851e+16 0 2
   (65 SG-NOMINAL 4.1552e+15 0 1
    (63 FEM-NOMINAL 4.1552e+15 0 1
     (8 O_DEFINITE-ARTICLE 2.0776e+15 0 1 ("a" 0 1))))
   (145 SG-NOMINAL 0 1 2
    (140 FEM-NOMINAL 0 1 2 (15 CRIANÇA 0 1 2 ("criança" 1 2)))))
  (358 HEAD-COMP_NOTCLITIC 2.34842e+17 2 7
   (98 3SG-VERB 0 2 3
    (97 PRES-IND-VERB 0 2 3 (16 OBEDECER 0 2 3 ("obedece" 2 3))))
   (357 FUNCTOR-HEAD-HCOMPS-SCOPAL 4.5623e+16 3 7
    (17 APENAS_NP-ADJUNCT 3.9016e+15 3 4 ("apenas" 3 4))
    (356 HEAD-COMP_NOTCLITIC 3.76979e+16 4 7
     (29 A_NONPREDICATIONAL-NP_OR_VP-PREPOSITION 1.03477e+16 4 5 ("a" 4 5))
     (355 FUNCTOR-HEAD-HCOMPS-SCOPAL 6.65476e+15 5 7
      (66 SG-NOMINAL 4.1552e+15 5 6
       (64 FEM-NOMINAL 4.1552e+15 5 6
        (38 O_DEFINITE-ARTICLE 2.0776e+15 5 6 ("a" 5 6))))
      (47 SG-NOMINAL 2.95058e+16 6 7
       (46 FEM-NOMINAL 2.95058e+16 6 7
        (45 MÃE_1_NOUN 0 6 7 ("mãe." 6 7)))))))))))
```

Syntactic constituency tree:

```
(CP
 (S (NP-SJ-ARG1 (ART-SP (ART-SP (ART-SP (a)))) (N (N (N (criança)))))
  (VP (V (V (V (obedece))))
   (PP-IO-ARG2 (ADV-M-M (apenas))
    (PP (P (a)) (NP-C (ART-SP (ART-SP (ART-SP (a)))) (N (N (N (mãe.)))))))))))
```

AVM: Due to its large size, this representation is left out of this document. You may find it in the sample document that is provided in the META-SHARE site.

MRS:

```
 [ LTOP: h1
   INDEX: e2 [ e ELLIPTICAL-PUNCT: BOOL SF: PROPOSITION-OR-QUESTION E.TENSE:
PRESENTE E.ASPECT.PERF: - E.MOOD: INDICATIVO ]
   RELS: <
          [ _o_q_rel
            LBL: h3
            ARG0: x6 [ x GENDER: FEMININE NUMBER: SINGULAR PERSON: 3RD ]
            RSTR: h4 [ h SCOPE: NARROW ]
```

```
          BODY: h5 [ h SCOPE: NARROW ] ]
        [ "_criança_n_rel"
          LBL: h7
          ARG0: x6 ]
        [ "_obedecer_v_-a-_rel"
          LBL: h8
          ARG0: e2
          ARG1: x6
          ARG2: x9 [ x PERSON: 3RD NUMBER: SINGULAR GENDER: FEMININE ] ]
        [ "_apenas_q_rel"
          LBL: h10 [ h SCOPE: SCOPE ]
          ARG0: e12
          ARG1: h11 [ h SCOPE: SCOPE ] ]
        [ _o_q_rel
          LBL: h11
          ARG0: x9
          RSTR: h13 [ h SCOPE: NARROW ]
          BODY: h14 [ h SCOPE: NARROW ] ]
        [ "_mãe_n_1-de-_rel"
          LBL: h15
          ARG0: x9
          ARG1: y16 ] >
  HCONS: < h1 qeq h8 h4 qeq h7 h13 qeq h15 > ]
```

Indexed MRS:

```
<h1,e2:BOOL:PROPOSITION-OR-QUESTION:PRESENTE:-:INDICATIVO,
{h3:_o_q(x6:FEMININE:SINGULAR:3RD, h4:NARROW, h5:NARROW),
h7:_criança_n(x6),
h8:_obedecer_v_-a-(e2, x6, x9:3RD:SINGULAR:FEMININE),
h10:_apenas_q(:SCOPEe12, h11:SCOPE),
h11:_o_q(x9, h13:NARROW, h14:NARROW),
h15:_mãe_n_1-de-(x9, y16)},
{h1 qeq h8,
h4 qeq h7,
h13 qeq h15}>
```

Prolog MRS:

```
psoa(h1,e2,[rel('_o_q',h3,
[attrval('ARG0',x6),attrval('RSTR',h4),attrval('BODY',h5)]),rel('_criança_n',h7,
[attrval('ARG0',x6)]),rel('_obedecer_v_-a-',h8,
[attrval('ARG0',e2),attrval('ARG1',x6),attrval('ARG2',x9)]),rel('_apenas_q',h10,
[attrval('ARG0',e12),attrval('ARG1',h11)]),rel('_o_q',h11,
[attrval('ARG0',x9),attrval('RSTR',h13),attrval('BODY',h14)]),rel('_mãe_n_1-
de-',h15,
[attrval('ARG0',x9),attrval('ARG1',y16)])],hcons([qeq(h1,h8),qeq(h4,h7),qeq(h13,
h15)]))
```

RMRS (Robust MRS):

```
 h1
 _o_q(h3,x6:)
 _criança_n(h7,x6:)
 _obedecer_v_-a-(h8,e2:)
 _apenas_q(h10,e12:)
 _o_q(h11,x9:)
 _mãe_n_1-de-(h15,x9:)
 RSTR(h3,h4:)
 BODY(h3,h5:)
 ARG1(h8,x6:)
 ARG2(h8,x9:)
```

```
 ARG1(h10,h11:)
 RSTR(h11,h13:)
 BODY(h11,h14:)
 ARG1(h15,u16:)
 qeq(h1:,h8)
 qeq(h4:NARROW:,h7)
 qeq(h13:NARROW:,h15)
```

XML MRS:

```
<rmrs cfrom='-1' cto='-1'a criança obedece apenas a a mãe.'11 @ 0 @ '>
<label vid='1'/>
<ep cfrom='-1' cto='-1'><realpred lemma='o' pos='q'/><label vid='3'/><var
sort='x' vid='6'/></ep>
<ep cfrom='-1' cto='-1'><realpred lemma='criança' pos='n'/><label vid='7'/><var
sort='x' vid='6'/></ep>
<ep cfrom='-1' cto='-1'><realpred lemma='obedecer' pos='v' sense='-a-'/><label
vid='8'/><var sort='e' vid='2'/></ep>
<ep cfrom='-1' cto='-1'><realpred lemma='apenas' pos='q'/><label vid='10'/><var
sort='e' vid='12'/></ep>
<ep cfrom='-1' cto='-1'><realpred lemma='o' pos='q'/><label vid='11'/><var
sort='x' vid='9'/></ep>
<ep cfrom='-1' cto='-1'><realpred lemma='mãe' pos='n' sense='1-de-'/><label
vid='15'/><var sort='x' vid='9'/></ep>
<rarg><rargname>RSTR</rargname><label vid='3'/><var sort='h' vid='4'/></rarg>
<rarg><rargname>BODY</rargname><label vid='3'/><var sort='h' vid='5'/></rarg>
<rarg><rargname>ARG1</rargname><label vid='8'/><var sort='x' vid='6'/></rarg>
<rarg><rargname>ARG2</rargname><label vid='8'/><var sort='x' vid='9'/></rarg>
<rarg><rargname>ARG1</rargname><label vid='10'/><var sort='h' vid='11'/></rarg>
<rarg><rargname>RSTR</rargname><label vid='11'/><var sort='h' vid='13'/></rarg>
<rarg><rargname>BODY</rargname><label vid='11'/><var sort='h' vid='14'/></rarg>
<rarg><rargname>ARG1</rargname><label vid='15'/><var sort='u' vid='16'/></rarg>
<hcons hreln='qeq'><hi><var sort='h' vid='1'/></hi><lo><label
vid='8'/></lo></hcons>
<hcons hreln='qeq'><hi><var sort='h' vid='4' SCOPE='NARROW'/></hi><lo><label
vid='7'/></lo></hcons>
<hcons hreln='qeq'><hi><var sort='h' vid='13' SCOPE='NARROW'/></hi><lo><label
vid='15'/></lo></hcons>
</rmrs>
```

Elementary dependencies:

```
{e2:
 x6:_o_q[]
 e2:_obedecer_v_-a-[ARG1 x6:_criança_n, ARG2 x9:_mãe_n_1-de-]
 e12:_apenas_q[ARG1 x9:_o_q]
 x9:_o_q[]
}
```

Discriminants:

```
{
  _o_q ARG0 _criança_n
  _obedecer_v_-a- ARG1 _criança_n
  _obedecer_v_-a- ARG2 _mãe_n_1-de-
  _apenas_q ARG1 _o_q
  _o_q ARG0 _mãe_n_1-de-
  _criança_n GENDER feminine
  _criança_n NUMBER singular
  _criança_n PERSON 3rd
  _obedecer_v_-a- ELLIPTICAL-PUNCT bool
  _obedecer_v_-a- SF proposition-or-question
```

```
  _obedecer_v_-a- E.TENSE presente
  _obedecer_v_-a- E.ASPECT.PERF -
  _obedecer_v_-a- E.MOOD indicativo
  _mãe_n_1-de- PERSON 3rd
  _mãe_n_1-de- NUMBER singular
  _mãe_n_1-de- GENDER feminine
  _apenas_q _criança_n
  _apenas_q _mãe_n_1-de-
  _apenas_q _o_q
  _apenas_q _obedecer_v_-a-
  _criança_n _mãe_n_1-de-
  _criança_n _o_q
  _criança_n _obedecer_v_-a-
  _mãe_n_1-de- _o_q
  _mãe_n_1-de- _obedecer_v_-a-
  _o_q _obedecer_v_-a-
}
```

3.3. Corpus size (nmb. of tokens, NB occupied in disk)

The corpus is composed by 3,134 sentences with 68 MB compressed (105 MB uncompressed).

IV. Content Information

4.1. Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This is a monolingual annotated corpus.

4.2. The natural language(s) of the corpus

The language of the corpus is Portuguese in the orthographic norm pre-dating the orthographic norm of 1990[1].

4.3. Domain(s)/register(s) of the corpus

Excerpts from newspapers articles.

4.4. Annotation in the corpus (if an annotated corpus)

4.4.1. Types of annotation (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

Deep grammatical representations.

4.4.2. Tags (if POS/WSD/TIME/discourse/etc – tagged or parsed)

Not applicable.

4.4.3. Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

Not applicable.

4.4.4. Attributes and their values (if annotated)

Not applicable.

4.5. Intended application of the corpus

The corpus can be used in linguistic research and in the development and testing of language

1 This means that the orthography rules used are those that are described by the Orthography Reform of 1945. The orthographic agreement of 1990 was adopted in May of 2009.

processing tools.

4.6. Reliability of the annotations (automatically/manually assigned) – if any

CINTIL-DeepBank is developed along a semi-automatic process, where an automatic annotation output by the grammar is manually revised by language experts with post-graduate degrees in Linguistics. In the first stage, a deep computational grammar (Branco and Costa, 2008) is used to generate all the possible parses for a given sentence (the parse forest). This is followed by a manual disambiguation stage where the correct parse is chosen from among those in the parse forest. This second stage is performed along the double-blind annotation method followed by adjudication: two annotators work independently and, for those cases where their decisions differ, a third annotator (the adjudicator) makes the final decision. For this corpus, the level of inter-annotator agreement (ITA) is 0.83 in terms of the specific inter-annotator metric developed for this kind of corpora and annotation (Castro, 2011).

V. Relevant References and Other Information

Branco, António, Francisco Costa, João Silva, Sara Silveira, Sérgio Castro, Mariana Avelãs, Clara Pinto, and João Graça, 2010. “Developing a Deep Linguistic Databank Supporting a Collection of Treebanks: The CINTIL DeepGramBank”. In Proceedings of the Seventh International Conference on Language Resources and evaluation (LREC'10) May 19-21, Valetta, Malta pp. 1810-1815.

Branco, António and Francisco Costa, 2008, “A computational grammar for deep linguistic processing of portuguese: LXGram”. In Technical Reports Series. University of Lisbon, Department of Informatics, 2008.

Castro, Sérgio, 2011, Developing Reliability Metrics and Validation Tools for datasets with deep linguistic Information, MA Dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.

Copestake, Ann, 2006, “Minimal Recursion Semantics: An Introduction”. In Research on Language and Computation, 3.4, pp. 281-332.

B.10 Portuguese Lexicon

Deliverable 4.6- Lexicon

I. Basic Information

1.1. Lexicon information

This resource is part of Deliverable 4.6. It comprises 600 lexicon entries used in LXGram, an HPSG computational grammar for deep linguistic processing of Portuguese.

1.2. Representation of the lexicon

The lexicon has two files, one with the lexicon and the other with the types.

1.3. Character encoding

The files are encoded in UTF-8.

II. Administrative Information

2.1. Contact person

Name: António Branco
Address: Departamento de Informática NLX - Grupo de Fala e Linguagem Natural, Faculdade de Ciências da Universidade de Lisboa, Edifício C6, Campo Grande 1749-016 Lisboa
Position: Associate professor
Affiliation: Faculty of Sciences, University of Lisbon
Telephone: +351 217 500 087
Fax: +351 217 500 084
E-mail: antonio.branco@di.fc.ul.pt

III. Technical Information

3.1. Directories and files

The archive is a .zip file containing two plain text files, one with the lexicon and the other with the lexical types.

3.2. Data structure of an entry

Each entry, be it a lexical entry or a lexical type, is defined through an AVM in TDL format.

Lexical entry:

```
afirmar :=
verb-comp_np_inf_cp+ind_declarative-lex &
[ STEM < "afirmar" >,
  SYNSEM.LOCAL.CONT.KEYS.KEY.PRED "_afirmar_v_rel" ].
```

Lexical type:

```
noun-or-pronoun-item :=
  nominal-elem &  synsat-no_subj-plus-elem &
  no-ctxt-lex-item &
  non-negative-polarity-non-clause-introducing-non-verb-premodifier-item &
  [ SYNSEM.LOCAL [ CAT [ HEAD noun,
                         VAL.HCOMPS.COMPS-POSITION adjacent ],
                   CONT.HOOK [ SARG #n-index,
                               MOTHER-SARG #n-index ] ] ].
```

3.3. Lexicon size

The file with the lexicon has 600 entries.

IV. Content Information

4.1. Type of the lexicon (monolingual/multilingual, parallel/comparable, raw/annotated)

This is a monolingual lexicon.

4.2. The natural language(s) of the lexicon

The language of the lexicon is Portuguese with pre-spelling reform of 1990[1].

4.3. Domain(s)/register(s) of the lexicon

Not applicable.

V. Relevant References and Other Information

Branco, António and Francisco Costa, 2008, "A computational grammar for deep linguistic processing of portuguese: LXGram". In Technical Reports Series. University of Lisbon, Department of Informatics, 2008.

1 This means that the orthography rules used are those that are described by the Orthography Reform of 1945. The orthographic agreement of 1990 was adopted just in may of 2009 and is being implemented until 2012.