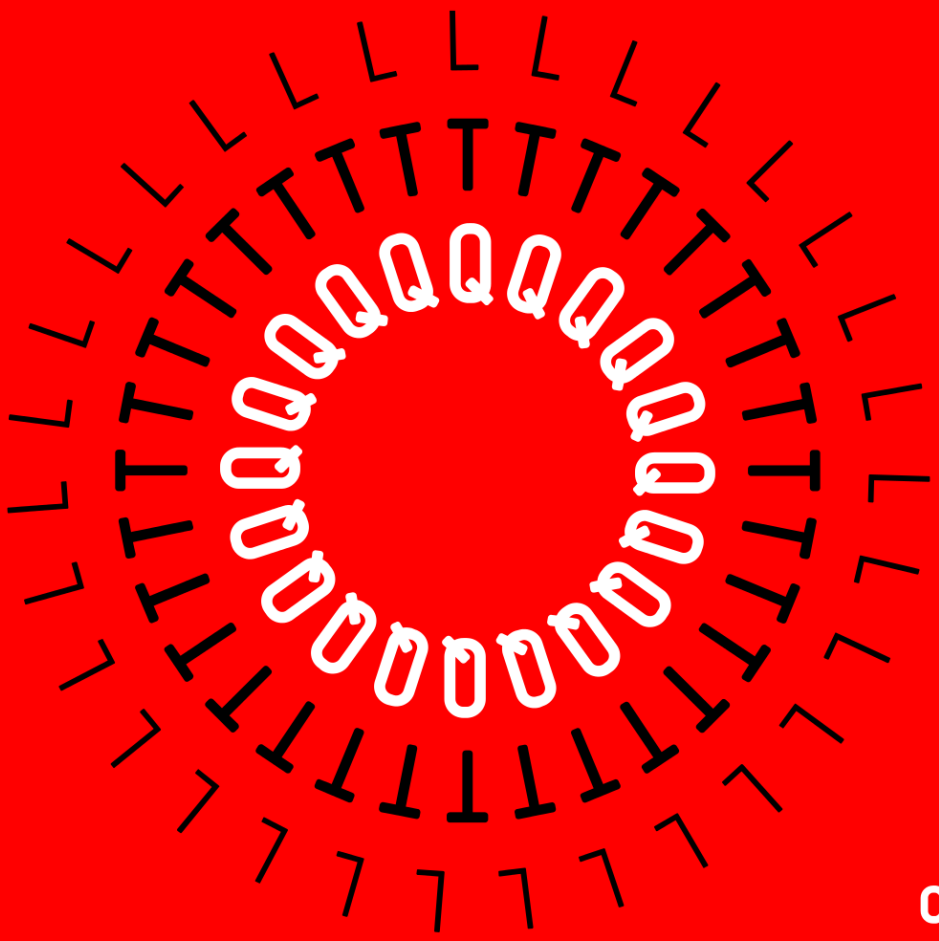


qtleap
quality translation by deep language engineering approaches

Report on second pilot version of LRTs enhanced to support deep processing

DELIVERABLE D4.10

VERSION 2.0 | 2015-11-12

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:

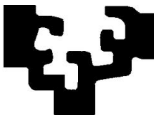
Faculty of Sciences, University of Lisbon

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	Mar 16, 2015	Petya Osenova	IICT-BAS	First draft
1.2	Mar 22, 2015	Aljoscha Burchardt	DFKI	Integrated text
1.1	Mar 27, 2015	João Silva	FCUL	Integrated text
1.4	Mar 27, 2015	Dieke Oele, Gertjan van Noord	UG	Integrated text
1.3	Mar 29, 2015	Gorka Labaka	UPV-EHU	Integrated text
1.5	Mar 30, 2015	Petya Osenova	IICT-BAS	Editing
1.6	Oct 16, 2015	Petya Osenova	IICT-BAS	Update of the text
1.7	Oct 16, 2015	Aljoscha Burchardt	DFKI	Update of the text
1.8	Oct 20, 2015	João Silva	FCUL	Update of the text
1.9	Oct 20, 2015	Dieke Oele, Gertjan van Noord	UG	Update of the text
1.10	Oct 22, 2015	Gorka Labaka	UPV-EHU	Update of the text
1.11	Oct 23, 2015	Petya Osenova	IICT-BAS	Editing
2.0	Oct 30, 2015	Markus Egg	UBER	Review comments incorporated

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 second pilot version of LRTs enhanced to support deep processing

DOCUMENT QTLEAP-2015-D4.10
EC FP7 PROJECT #610516

DELIVERABLE D4.10

completition
FINAL
status
SUBMITTED
dissemination level
PUBLIC

responsible
Kiril Simov (WP 4 Coordinator)
reviewer
Markus Egg
contributing partners
FCUL, DFKI, IICT-BAS, UBER, UPV-EHU, UG

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

Contents

1 Introduction **8**

2 Treebanks **8**
2.1 Basque . . . 9
2.2 Bulgarian . . . 10
2.2.1 BulTreeBank-DP . . . 10
2.2.2 BulEngTreebank . . . 10
2.2.3 ParDeepBankBG . . . 10
2.3 Dutch . . . 10
2.4 German . . . 11
2.4.1 German DeepBank . . . 11
2.4.2 The Berkeley Parser . . . 11
2.4.3 BitPar . . . 12
2.4.4 ParZu . . . 12
2.5 Portuguese . . . 13

3 Lexicons **13**
3.1 Bulgarian . . . 13
3.1.1 Bulgarian Ontology-Based Lexicon . . . 13
3.1.2 Bulgarian Valency Frame Lexicon . . . 13
3.2 Portuguese . . . 14

4 Conclusions **14**

5 Appendix A: Narrative Descriptions of D4.9 Language Resources **17**
5.1 Basque-English ParDeepBank part II . . . 18
5.2 Bulgarian Ontology-based Lexicon: BOL part II . . . 21
5.3 Bulgarian Valency Frame Lexicon: BVFL part II . . . 23
5.4 Dependency part of BulTreeBank: BulTreeBank-DP part II . . . 26
5.5 Bulgarian English Parallel Treebank: BulEngTreebank part II . . . 28
5.6 Bulgarian English Deep Bank: ParDeepBank part II . . . 30
5.7 Dutch Tree Bank . . . 33
5.8 Portuguese Deep Bank . . . 35
5.9 Portuguese Lexicon . . . 41

List of Abbreviations

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 interim report on the curation of language resources and tools that support deep MT associated do workpackage WP4, as planned in the DoW. This report is a follow-up of the deliverable D4.6 "First pilot version of language resources and tools (LRTs) enhanced to support robust deep processing" and D4.7 "Report on first pilot version of LRTs enhanced to support deep processing". Thus, it builds on the results of D4.6 and D4.7, specifying the additions since then. Here only novelties on resources and tools that have not been discussed there, will be considered. The resources are uploaded to the QTLeap repository.

The language resources are of two types:

- **Treebanks.** Syntactically annotated corpora to be used for deep (tree-based) machine translation models.
- **Lexicons.** Dictionaries that provide semantic and valency information to support MT transfer.

Both types of resources are either monolingual or multilingual. The novelties about supporting tools are introduced in the corresponding language and resource sections in this deliverable.

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. These parameters also take into account the need of creating more resources for the cases of languages with few resources (such as, Bulgarian). Table 1 summarizes the content of deliverable D1.3, which was a specification of what had been planned in the DoW. That planning indicates the resources and tools to be provided within deliverable D4.6 (already reported in companion report D4.7) and within deliverable D4.8, to be reported here.

All the data is equipped with metadata records, which are also uploaded on the project repository. In the repository, for each language in WP4 there is a separate directory — BG (for Bulgarian), DE (for German), EU (for Basque), NL (for Dutch) and PT (for Portuguese).

Since the language resources have been created in an incremental set-up, we will distribute them and QTLeap branded in the final stage, when the resources are considered complete and fully curated. In this way, the data will be ready to be used by the stakeholders and the community of interested users.

Spanish and Czech are not taken into account in the present series of WP4 deliverables, devoted to curation of language resources and tools, because at planning time, in the DoW, Czech was considered to have all the necessary instrumentarium prior to the project, and Spanish was added later during the negotiation phase and is not to be covered by this series of deliverables.

The rest of the present deliverable is organized as follows: first the extensions of the treebanks due and delivered in D4.9 are presented in Chapter 2; then the extensions of the lexicons due and delivered in D4.9 are described in Chapter 3.

2 Treebanks

Treebanks and deepbanks are presented below. The translation models are enhanced in both treebank types: monolingual (Bulgarian, German) and parallel (Basque, Bulgarian,

Language/Resources	M7	M13	M17	M21
Basque				
Basque-Eng ParDeepBank	2.5K	5.0K	6.0K	7.0K
Bulgarian				
BOL	600	1200	1600	2400
BVFL	900	1800	2400	3600
BulTreeBank-DP	15000	30000	40000	60000
BulEngTreebank	12000	24000	32000	48000
ParDeepBank	5000	10000	15000	25000
Dutch				
Lassy: manually	500/10K	1000/20K	1300/26K	3400/42K
Lassy: automatic	15K/300K	30K/600K	40K/800K	60K/1200K
German				
DeepBankDE	2.5K	5.0K	6.0K	8.0k
Portuguese				
Lexicon	15/300	30/600	40/750	60/1200
Corpora	1.5K/20K	3K/40K	4K/50K	6/70K

Table 1: Language Resources for Supporting deep processing, due at M23 (figures are cumulative).

Dutch, Portuguese).

2.1 Basque

Basque-Eng ParDeepBank is part of Deliverable 2.5. In its current development, it consists of 400 sentences and 7,868 tokens (4,146 English tokens and 3,722 Basque tokens). The sentences are excerpts from news text from the Wall Street Journal that have been manually translated into Basque to generate a parallel corpus.

The English sentences are part of the Penn Treebank corpus, and have been selected as they are already part of a English-Spanish parallel corpus[1]. In this way, we will additionally contribute to the development of 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, the results were corrected manually.

For English, Stanford dependency tags are used,[2] whereas Basque syntactic annotation follows the BDT guidelines [Aldezabal et al., 2009]. It is important to note 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

[1] http://repositori.upf.edu/handle/10230/20049

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

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 modules that will transform source language parses into target language ones.

2.2 Bulgarian

2.2.1 BulTreeBank-DP

The part of BulTreeBank-DP in D4.9 is an extension of D4.6 with newly processed sentences. It comprises 5049 sentences (78142 tokens). The sentences were manually annotated on the morphological level. Then they were automatically annotated syntactically. Part of the syntactic annotation was checked manually. Similarly to the previous data portion, the new part of the treebank includes the following levels of linguistic information: tokenization, POS, morphosyntactic features, dependency relations, and coreferences.

The data is represented in the CoNLL format. The metadata is given as an XML document and as a PDF document in the archive. The resource in its final version will be available for external use.

2.2.2 BulEngTreebank

The part of Bulgarian English Parallel Treebank in D4.9 is an extension of D4.6. It contains 3365 sentences (128 499 tokens). It includes English sentences from Wikipedia and their translation into Bulgarian. The English sentences were translated into Bulgarian by professional translators. The Bulgarian sentences are processed by the Bulgarian QTLeap pipeline and the English sentences are processed by IXA pipeline. The alignments on word level were done automatically.

The data is represented in the CoNLL format. The metadata is given as an XML document and as a PDF document in the archive. The resource in its final version will be available for external use.

2.2.3 ParDeepBankBG

The part of the Bulgarian English Parallel Deepbank in D4.9 is an extension of D4.6. It consists of 2133 sentences (51 730 tokens). It includes English sentences from the English Deepbank, whose domain is news and finance (Wall Street Journal). These sentences had already been analyzed using the ERG grammar (Copestake and Flickinger [2000]), and manually disambiguated. The sentences were translated into Bulgarian by professional translators. Bulgarian and English sentences were aligned on the word level. Then they were annotated morphologically and parsed by a dependency parser. Part of the result was corrected manually. The dependency analyses are in the CoNLL 2006 format.

The metadata is given as an XML document and as a PDF document in the archive. The resource in its final version will be available for external use.

2.3 Dutch

In addition to the data provided in D4.6, we have also provided the Alpino abstract dependency structures for the first **75 thousand sentences** (this corresponds to about

1.2 million words) from DPC[3], and about **50 thousand sentences** (which is 750 thousand words) from the manually annotated Lassy Small corpus. P11

In the latter case, we have used the gold standard annotations to guide the parser in finding the abstract dependency structure that is as close as possible to that gold standard.

The Alpino abstract dependency structures – together with the original sentences from the respective corpora – provide good examples for the mapping of an abstract meaning representation to an actual sentence. As such, this data provides learning and evaluation data for experiments in generation. The DPC data is somewhat further removed from actual gold data because we use the parser, which may make mistakes, to find the abstract dependency structure. The Lassy Small data is more reliable since we exploit the manual syntactic annotations to ensure that the abstract dependency structures are as close as possible to the correct analysis.

We chose DPC, rather than Europarl, because the data is much less noisy than KDE, and (to a lesser extent) Europarl. Please note, on the qtleap archive, the file train.tok.en and train.tok.nl as used in Pilot 0 and Pilot 1 contain *both* KDE and DPC. The DPC data starts at line 104916 in the train.tok.en and train.tok.nl files.

2.4 German

2.4.1 German DeepBank

The syntactic corpus contains 15 000 sentences taken from the German TIGER treebank[4] 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 130 000 tokens (file size 64 MB).

Experiments with an alternative German HPSG grammar (GG[5]) have shown that the grammar that has been used so far has the far better coverage. Manual editing and selection will only be performed if relevant for the MT development within the project.

The data is represented in ERG text format. The metadata is given as XML document in the archive as well. The resource can only be used if the requesting institution has a license for the original Tiger treebank.

The German Deep Bank is available on META-SHARE under a CC - BY - NC - SA license.[6]

Additional TreeBank resources for German are being used in several parts of the project, through the usage of specific parsing tools. A description of our efforts to adapt, evaluate and enhance LRPs and processing tools for each use case is following:

2.4.2 The Berkeley Parser

The Berkeley Parser is a state-of-the-art Probabilistic Context-Free Grammar (PCFG) parser that supports unlexicalized parsing with hierarchically state-split PCFGs, supporting optimal pruning via a coarse-to-fine method [Petrov and Klein, 2007]. It has the advantage that it is accurate and fast, by using multi-threading technology. Apart from

[3]Dutch Parallel Corpus — http://www.kuleuven-kulak.be/DPC
[4]http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html
[5]http://gg.dfki.de
[6]http://metashare.dfki.de/repository/browse/deepbankde/3dff1c96556511e5b94c003048d082a4bff589ce425

the best tree for each parse, it also provides the parsing log-likelihood and a number of k-best trees along with their parse probabilities. The English grammar has been trained on the Wall Street Journal. The German grammar, using Latent Variable Grammar [Petrov and Klein, 2008], has been trained on the TIGER [Brants et al., 2004] and TueBaD/Z [Telljohann et al., 2004] treebanks, as released by the ACL 2008 workshop on Parsing German [Kübler, 2008].

The use of the Berkeley parser has shown good results as a quality indicator for Quality Estimation. In this frame, our engineering efforts have focused on connecting the parser in the broader pipeline of sentence selection in Pilot 1, by providing a socket interface that exposes the Java library of the parser as a python object (see Py4J[7]). Additionally, in out effort to acquire features for qualitative translation, we use word and phrase alignment methods in order to map node labels between source and produced translations.

2.4.3 BitPar

BitPar is a parser for highly ambiguous probabilistic context-free grammars. It makes use of bit-vector operations that allow parallelizing and speeding up the the basic parsing operations [Schmid, 2006]. The English grammar is based on the PENN treebank [Marcus et al., 1993], whereas the German grammar is also based on the TIGER treebank. BitPar was also included on our annotation pipeline in order to provide additional evidence and allow comparisons to the observations on the Berkeley parses. It also provides sentence-level tree likelihood and k-best lists. Unfortunately, contrary to the Berkeley Parser, the k-best lists of BitPar were of limited usability due to small differences in their relative likelihood.

2.4.4 ParZu

The Zurich Dependency Parser for German (ParZu) follow a hybrid architecture including both a hand-written grammar and a statistics module that chooses the most-likely parse of each sentence [Sennrich et al., 2009]. As compared to many other German parsers, it integrates morphological information [Sennrich et al., 2013] and it does not use a chunker.

This parser has been employed for for the parsing needs of the German version of TectoMT. It has been chosen after an analysis of the capabilities of several parsers and their compiled grammars, including MDparser, Stanford Dependency Parser and MaltParser. ParZu was found optimal, as it provides the necessary morphological disambiguation upon parsing and connects well with relevant morphological analyzers and generators. Additionally, ParZu has shown to perform well in comparison to the other parsers in previous work [Williams et al., 2014].

In order to acquire morphological analysis for ParZu, we have been using the Zurich Morphological Analyzer for German (ZMORG), based on finite-state-transducers automatically extracted from Wiktionary [Sennrich and Kunz, 2014]. The tool can also function as a morphological generator and outputs the analysis in a modified SMOR format.

As part of our QTleap efforts on TectoMT, an "driver" between the SMOR format and the universal Lingua Interset [Zeman, 2008] was built and commited to the open repository. This was required as the Lingua Interset is needed by TectoMT. In a later stage, this conversion can allow interaction with the Universal Dependencies standard [de Marneffe et al., 2014].

[7] http://py4j.sourceforge.net/

2.5 Portuguese

Following the planning in D1.3, the extension of previous D4.6 Portuguese treebank amounts to 6,000 sentences. This extension of this resource is documented in the present section and it eventually consists of 6,356 sentences (64,624 tokens) taken from CETEM-Público, a corpus of news texts. These sentences have been annotated by LXGram, an HPSG computational grammar for deep linguistic processing of Portuguese, and then manually disambiguated under a method of double-blind annotation followed by adjudication by a separate annotator (see [Branco and Costa, 2008] for more details on this process). This method ensures a highly reliable data set, while the use of a deep computational grammar provides each sentence with an extremely consistent and rich linguistic annotation, which includes syntactic constituency, grammatical dependencies and different renderings of MRS-based representations of the meaning of the sentence [Copestake et al., 2005].

The QTLeap repository contains the associated XML metadata record, in META-SHARE format, and the narrative description (D4.9-Deepbank.pdf) of the resource.

Within the compressed D4.9_deepbank.zip file there is one file per sentence. Each file results from exporting the disambiguated LXGram analysis, and includes the full attribute-value matrix representation produced by the grammar. Due to the size of these representations, each file is compressed with GZip. If all individual sentence files are uncompressed, the corpus size reaches 1,470 MB.

Like its previous, shorter version, the full CINTIL-Deepbank v1.3 that integrates this 6,356 sentence extension keeps being distributed through the META-SHARE repository.

3 Lexicons

This section presents the planned and performed developments of Bulgarian and Portuguese lexicons, which include WordNet synsets, valence frames, ontological classes, etc.

3.1 Bulgarian

3.1.1 Bulgarian Ontology-Based Lexicon

In this part of the Bulgarian Ontology-based Lexicon (BOL) we provided additionally 1724 synsets. As it was described in D4.7, BOL is organized in synsets as in the WordNet, but the relations between the synsets are represented via a 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).

The data is represented in a table format. The metadata is given as an XML document and as a PDF document in the archive. The resource will be added to the data made available previously and extended to its final size.

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. The new part of the lexicon comprises 1368 new verb frames. For the format of the entries, please, consult D4.7. The metadata is given as an XML document and as a PDF document in the archive.

3.2 Portuguese

Following deliverable D1.3, the extension of previous D4.6 Portuguese lexicon amounts to to 1,200 lexical entries.

This extension of this resource is documented in the present section and it eventually consists of 1,200 entries for the lexicon of LXGram, the HPSG computational grammar for deep linguistic processing of Portuguese used [Costa and Branco, 2010], together with the definition of their corresponding deep lexical types. The entries in the lexicon originate from two sources: (i) those manually built by the grammar developers and (ii) those obtained from a lexicon external to the grammar that is developed and maintained by the annotators. Each entry is associated with a deep lexical type from the type hierarchy defined in the grammar. HPSG is a highly lexicalized grammar framework, which means that most linguistic information is encoded into the lexicon, viz., in the deep lexical type of an entry. This information includes part-of-speech, subcategorization (valence) frame, patterns for anticausative alternations, clitic behavior, etc.

The QTLeap repository contains the associated XML metadata record contains the associated XML metadata record (D4.9-Lexicon.xml), in META-SHARE format, and the narrative description (D4.9-Lexicon.pdf) of the resource. Within the compressed D4.9_lexicon.zip file there are two files: D4.9-lexicon.tdl (220 KB), with the lexical entries; and D4.9-types.tdl (159 KB), with the definitions of the corresponding deep lexical types. The lexicon and types are represented in the Type Definition Language (TDL) format used by LXGram.

The full lexicon and corresponding types, which includes this extension, are distributed together with the grammar, when a stable version is reached.

4 Conclusions

In this deliverable we describe the recently completed extensions of the language resources for supporting deep processing. These extensions amount to approx. 60% of the data development to be covered in WP4. The resources covered two types: treebanks (and deepbanks) as well as lexicons. The additional planned data was successfully reached for all the respective languages and resources. As mentioned in the introduction, for the moment, the resources are available only through the QTLeap intranet page, since they are not considered completed yet. As planned, when they will be completed, they will be made publicly available through the QTLeap branding.

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.

António 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.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation*, 2 (4):597–620, 2004. ISSN 1570-7075. doi: 10.1007/s11168-004-7431-3.

Ann Copestake and Dan Flickinger. An open source grammar development environment and broad-coverage english grammar using hpsg. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece, 2000.

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

Francisco Costa and António Branco. Lxgram: A deep linguistic processing grammar for Portuguese. In *Lecture Notes in Artificial Intelligence*, volume 6001, pages 86–89. Springer, Berlin, May 2010. URL http://nlx.di.fc.ul.pt/~fcosta/papers/propor2010.pdf.

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

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of LREC*, 2014.

Sandra Kübler. The PaGe 2008 Shared Task on Parsing German. In *Proceedings of the Workshop on Parsing German*, PaGe '08, pages 55–63, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. ISBN 978-1-932432-15-2. URL http://dl.acm.org/citation.cfm?id=1621401.1621409.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Comput. Linguist.*, 19(2):313–330, June 1993. ISSN 0891-2017. URL http://dl.acm.org/citation.cfm?id=972470.972475.

Slav Petrov and Dan Klein. Improved inference for unlexicalized parsing. In *Proceedings of the 2007 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York, 2007. Association for Computational Linguistics.

Slav Petrov and Dan Klein. Parsing German with Latent Variable Grammars. In *Proceedings of the Workshop on Parsing German*, PaGe '08, pages 33–39, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. ISBN 978-1-932432-15-2. URL http://dl.acm.org/citation.cfm?id=1621401.1621406.

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.

Helmut Schmid. Trace Prediction and Recovery with Unlexicalized PCFGs and Slash Features. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 177–184, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220198. URL http://dx.doi.org/10.3115/1220175.1220198.

Rico Sennrich and Beat Kunz. Zmorge: A German Morphological Lexicon Extracted from Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014. ISBN 978-2-9517408-8-4. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/116_Paper.pdf.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. A new hybrid dependency parser for German. *Proceedings of the German Society for Computational Linguistics and Language Technology*, pages 115–124, 2009.

Rico Sennrich, Martin Volk, and Gerold Schneider. Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis. In *RANLP*, pages 601–609, 2013.

Heike Telljohann, Erhard Hinrichs, Sandra Kübler, Ra Kübler, and Universität Tübingen. The Tüba-D/Z Treebank: Annotating German with a Context-Free Backbone. In *In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004*, pages 2229–2235, 2004.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, Eva Hasler, and Philipp Koehn. Edinburgh's Syntax-Based Systems at WMT 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 207–214, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W14/W14-3324.

Daniel Zeman. Reusable Tagset Conversion Using Tagset Drivers. In *LREC*, 2008.

5 Appendix A: Narrative Descriptions of D4.9 Language Resources

P17

In this appendix we present the narrative descriptions for each language resource uploaded in the repository for D4.9.

5.1 Basque-English ParDeepBank part II

P18

ENGLISH-BASQUE PARALLEL DEEPBANK

1 BASIC INFORMATION

1.1 Corpus composition

This resource is part of Deliverable 2.5 of the QTLeap FP7 project (Contract number 610516). In its current development, it is composed of 400 sentences (4,146 English tokens and 3,722 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

Creative Commons – Attributions (CC-BY)

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 400 parallel sentences, making up a total of 7,868 tokens (4,146 English tokens and 3,722 Basque tokens) and needs about 350 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

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*

5.2 Bulgarian Ontology-based Lexicon: BOL part II

D4.9: BULGARIAN ONTOLOGY-BASED LEXICON (BOL)

LEXICA DOCUMENTATION

1. BASIC INFORMATION

1.1 Lexicon type: **The BulTreeBank WordNet (BTB-WN) Part II**
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)
==========

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

Licence
---

3. TECHNICAL INFORMATION

3.1 Directories and files: ***in Table Text Style***
3.2 Data structure of an entry:
3.3 Lexicon size (nmb. of lexical items, KB occupied on disk): ***466 944 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:

- 1,724 ***synsets***
- 4,197 words
- Covers the most frequent word in BulTreeBank treebank

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
}

5.3 Bulgarian Valency Frame Lexicon: BVFL part II

D4.9: BULGARIAN VALENCY FRAME LEXICON (BVFL)

LEXICA DOCUMENTATION

1. BASIC INFORMATION

1.1 Lexicon type: **Bulgarian Valency Frame Lexicon (BVFL) part II**
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)

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***

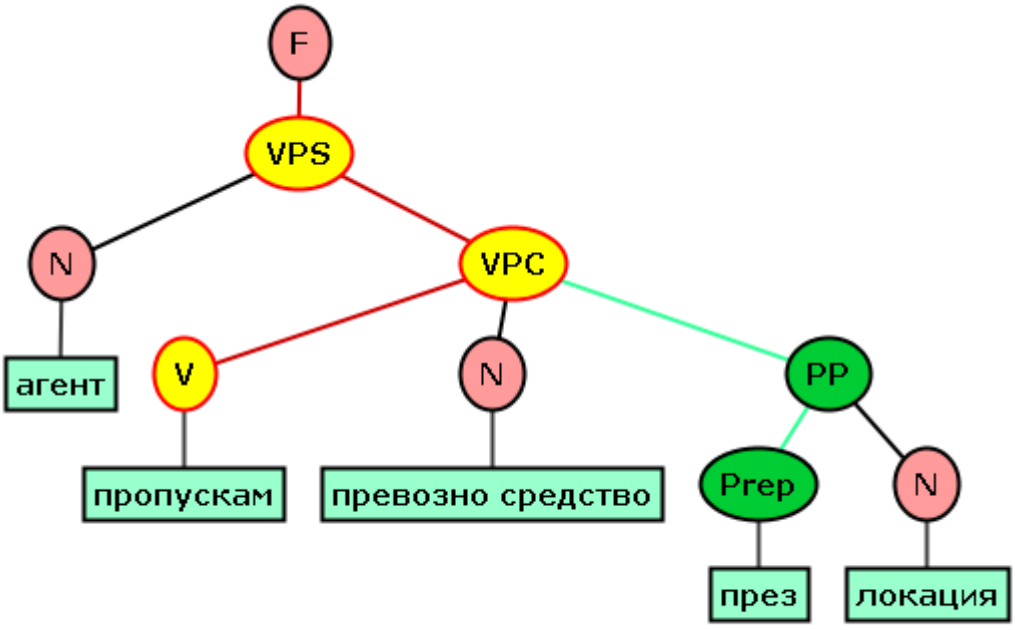
4. 2 Entry Type:

<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: ***3091 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.

5.4 Dependency part of BulTreeBank: BulTreeBank-DP part II

D4.9: BulTreeBank-DP M23

1 BASIC INFORMATION

1.1 Corpus composition: ***BulTreeBank-DP M17 is a morphologically and syntactically annotated sentences***

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 M17
==========

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

3 TECHNICAL INFORMATION

3.1 Directories and files: ***6 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

3.3 Corpora size (nmb. of tokens, MB occupied on disk): ***78142 tokens; 5049 sentences, 4,6 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***
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)
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) ***– morphological annotation is completely manual, syntactic annotation is automatic and manually checked***

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/

5.5 Bulgarian English Parallel Treebank: BulEngTreebank part II

P28

D4.9-BulEngTreebank

1 BASIC INFORMATION

1.1 Corpus composition:

This resource is part of Deliverable 4.9. It is composed of 3365 sentences (128 499 tokens) which are part of Bulgarian English Parallel Treebank part II. It includes English sentences from Wikipedia and their translation in Bulgarian.
Bulgarian sentences are aligned automatically to English on word level. Then they are morphologically annotated and parsed by a dependency parser. Then the result is manually corrected. English sentences are analysed by IXA pipeline.

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

D2.5-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): **3365 sentences; 128 499 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: ***Wikipedia***

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 IXA analysis.

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 automatically 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/

5.6 Bulgarian English Deep Bank: ParDeepBank part II

P30

D4.9-ParDeepBankBG part II

1 BASIC INFORMATION

1.1 Corpus composition:

This resource is part of Deliverable 4.9. It is composed of 2133 sentences (51 730 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

D2.5-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): 51 730 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/

DeepBankDE for QTLeapProject

1 BASIC INFORMATION

1.1 Corpus composition

The corpus contains 15.000 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.

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)

The corpus is available in the project intranet. It is externally available via META-SHARE.

2.3 Copyright statement and information on IPR

The corpus is available under a CC - BY - NC - SA license. The resource may only be used if the requester has a license for the original TIGER treebank.

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

5.7 Dutch Tree Bank

Description of Language Resources

1 Language Resources for Dutch

1.1 Basic Information:

We provide the Alpino abstract dependency structures for the first 75 thousand sentences (this corresponds to about 1.2 million words) from DPC (Dutch Parallel Corpus — http://www.kuleuven-kulak.be/DPC), and about 50 thousand sentences (about 750 thousand words) from the manually annotated Lassy Small corpus. In the latter case, we have used the gold standard annotations to guide the parser to find the abstract dependency structure that is as close as possible to that gold standard.

The Alpino abstract dependency structures - together with the original sentences from the respective corpora - provide good examples of the mapping of an abstract meaning representation to an actual sentence. As such, this data provides learning and evaluation data for experiments in generation. The DPC data is somewhat further removed from actual gold data because we use the parser, which may make mistakes, to find the abstract dependency structure. The Lassy Small data is more reliable since we exploit the manual syntactic annotations to ensure that the abstract dependency structures are as close as possible to the correct analysis.

We chose DPC, rather than Europarl, because the data is much less noisy than KDE, and (to a lesser extent) Europarl.

Representation of the resource:
Database

Character encoding:
UTF-8

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

1.3 Technical Information

In the qtleap archive, the file train.tok.en and train.tok.nl as used in Pilot 0 and Pilot 1 contain *both* KDE and DPC. The DPC data starts at line 104916 in the train.tok.en and train.tok.nl files.

Resource size:
125 thousand sentences (ca. 1.95 million words)

1.4 Content Information

Type of the corpus:
Alpino abstract dependency structures

The natural language(s) of the corpus:
Dutch

Domain of the corpus:
General

Intended application of the resource in the project:
This data provides learning and evaluation data for experiments in generation.

Reliability of the annotations:
Automatically (DPC)/ Manually (Lassy Small)

5.8 Portuguese Deep Bank

Deliverable 4.9 - DeepBank 1.3

I. Basic Information

1.1. Corpus information

This resource is part of Deliverable 4.9 of the QTLeap FP7 project (Contract number 610516). It is composed of 6,356 sentences (64,624 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 1,369 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 3,409 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 6,356 sentences with 134 MB compressed (1,470 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.

5.9 Portuguese Lexicon

Deliverable 4.9 - Lexicon

I. Basic Information

1.1. Lexicon information

This resource is part of Deliverable 4.9 of the QTLeap FP7 project (Contract number 610516). It comprises 1,291 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 ],
```

3.3. Lexicon size

The file with the lexicon has 1,291 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.