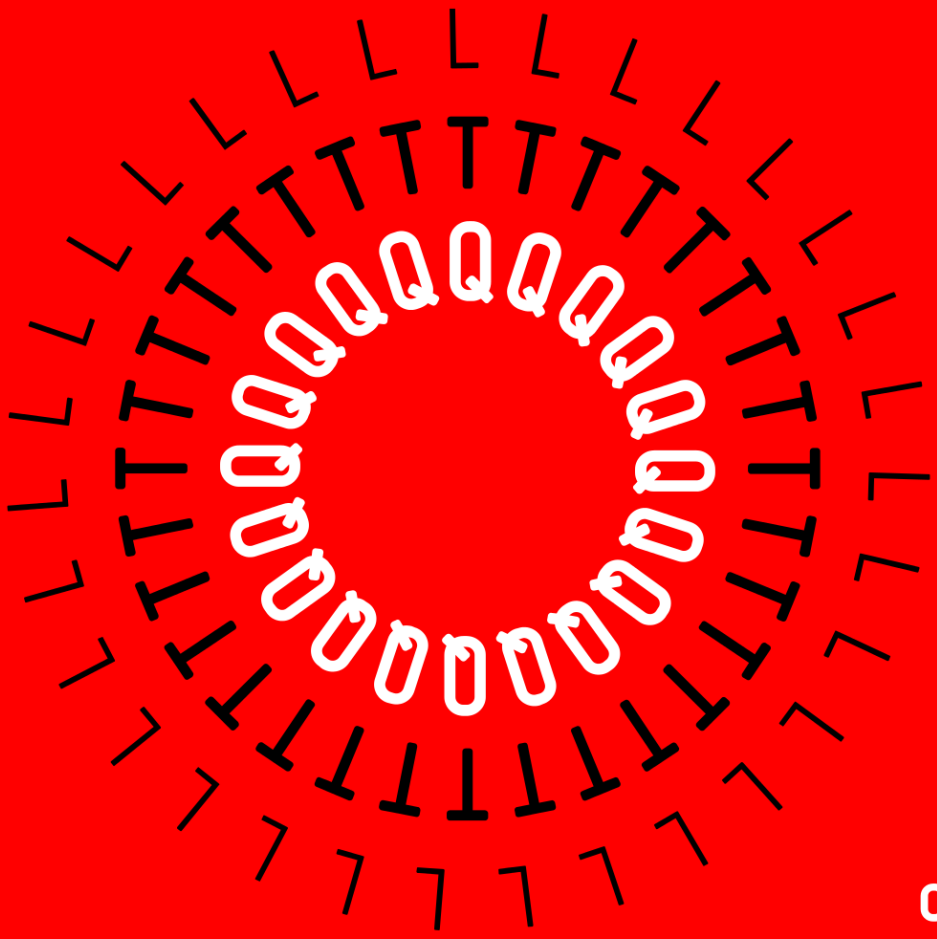


qtleap
quality translation by deep language engineering approaches

Report on the Final Version of LRTs Enhanced to support Advanced Crosslingual Lexical Ambiguity Resolution

DELIVERABLE D5.9

VERSION 1.2 | 2016-07-26

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:

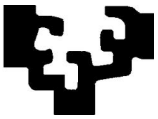
Faculty of Sciences, University of Lisbon

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
0.1	November 30, 2015	Eneko Agirre, Arantxa Otegi	UPV/EHU	First draft of tools sections
0.2	December 11, 2015	Eneko Agirre	UPV/EHU	First draft of the improvement section
0.3	January 15, 2016	Eneko Agirre, Arantxa Otegi	UPV/EHU	First complete structure
0.4	March 17, 2016	Arantza Otegi, Kiril Simov	UPV/EHU, IICT-BAS	Move sections from D5.4 and D5.6
0.5	March 21, 2016	Steven Neale, Roman Sudarikov	FCUL, CUNI	Update tools and resources sections
0.6	April 14, 2016	Eneko Agirre, Oier Lopez de Lacalle, João Silva, Chakaveh Saedi	UPV/EHU, FCUL	Update tools and improvements sections
0.7	April 19, 2016	Roman Sudarikov	CUNI	Update WSD and NED sections
0.8	April 28, 2016	Michal Novák	CUNI	Update the section on Czech coreference
0.9	April 28, 2016	Eneko Agirre, Arantxa Otegi	UPV/EHU	Finalize deliverable
1.0	May 16, 2016	Kiril Simov, Eneko Agirre, Roman Sudarikov, João Silva, Arantxa Otegi	IICT-BAS, FCUL, CUNI, UPV/EHU	Reflected the comments of the internal reviewer
1.1	May 17, 2016	Kiril Simov	IICT-BAS	Availability tables updated
1.2	Jul 26, 2016	Eneko Agirre, Arantxa Otegi	UPV/EHU	Corrected some errata

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 the Final Version of LRTs Enhanced to support Advanced Crosslingual Lexical Ambiguity Resolution

DOCUMENT QTLEAP-2016-D5.9
EC FP7 PROJECT #610516

DELIVERABLE D5.9

completition
FINAL
status
SUBMITTED
dissemination level
PUBLIC

responsible
Eneko Agirre (WP5 Coordinator)
reviewer
Gertjan van Noord
contributing partners
UPV/EHU, CUNI, FCUL, IICT-BAS

authors
E. Agirre, A. Otegi, N. Aranberri, G. Labaka, O. Lopez de Lacalle, R. Sudarikov, M. Popel, M. Novàk, O. Dušek, O. Bojar, S. Neale, J. Silva, C. Saedi, A. Branco, P. Osenova, K. Simov

Contents

P8

Executive summary

The goal of the QTLeap project is to develop Machine Translation (MT) technology that goes beyond the state of the art in terms of "depth" of the methods and knowledge used. The goal of WP5 is to enhance MT with advanced crosslingual methods for the resolution of referential and lexical ambiguity by pursuing the following objectives:

1. to provide for the assembling and curation of the language resources and tools (LRTs) available to support the resolution of referential and lexical ambiguity (Task 5.1, starting M1);

2. to leverage the resolution of referential and lexical ambiguity by means of advanced crosslingual named entity and word sense resolution methods (Task 5.2, starting M1);

3. to proceed with the intrinsic evaluation of the solutions found in the previous task (Task 5.3, starting M10);

4. to contribute for high quality machine translation by using semantic linking and resolving to improve MT (Task 5.4, starting M17). In particular Pilot 2 (M24) will be devoted to check the contribution of the tools in this WP to MT.

The work reported on this document has been carried out along the plans and is based on the project Description of Work, Deliverable 1.3 ("Management plan for language resources and tools"), Deliverable 1.7 ("LRTs Interim Report and Plan Update"), Deliverable 1.10 ("LRTs second interim management report and plan update") and Deliverable 5.1 ("State of the art").

The present deliverable documents the language resources and tools that compose deliverable D5.8 "Final version of language resources and tools (LRTs) enhanced to support advanced crosslingual ambiguity resolution".

Deliverables D1.3, D1.7 and D1.10 describe the new resources and tools in deliverable D5.8, as follows:

- Sense annotated corpora, for all languages in WP5 (BG Bulgarian, CS Czech, EN English, ES Spanish, EU Basque, PT Portuguese): 1M tokens aligned, 10M tokens comparable.

For easier use in the project, D5.8 actually includes the new LRTs, plus all LRTs already in D5.3 (which are described in the accompanying report, D5.4), D5.5 (which are described in D5.6). As this is the final report, it includes material from reports D5.4 and D5.6. It is thus not necessary to read those reports.

A few of the LRTs in D5.8 may have less wide distribution, but the large majority are publicly available, as described in detail in each Section below and summarized in Appendix B. For project internal purposes and the sake of replicability, all LRTs, private and public, are also stored in our internal repository.

WP5 comprises other activities in the third year of the project:

- Task 5.2: Advanced tools for NED, WSD, CR, due in M36.

- Task 5.3: Intrinsic evaluation of NERC/NED, WSD, CR, due in M34.

- Task 5.4: MT experiments, due in M33.

The experiments in Task 5.4 will be reported in D5.11, due in M36 ("Report on MT improved with semantic linking and resolving"). The evaluation of advanced tools (Tasks 5.2 and 5.3) as used in D5.8 are reported here. Some improvements are ongoing, and will be reported in D5.11. P11

1 Introduction

The goal of the QTLeap project is to develop Machine Translation (MT) technology that goes beyond the state of the art in terms of "depth" of the methods and knowledge used. The goal of WP5 is to enhance MT with advanced crosslingual methods for the resolution of referential and lexical ambiguity by pursuing the following objectives:

1. to provide for the assembling and curation of the language resources and tools (LRTs) available to support the resolution of referential and lexical ambiguity (Task 5.1);
2. to leverage the resolution of referential and lexical ambiguity by means of advanced crosslingual named entity and word sense resolution methods (Task 5.2);
3. to proceed with the intrinsic evaluation of the solutions found in the previous task (Task 5.3);
4. to contribute for high quality machine translation by using semantic linking and resolving to improve MT (Task 5.4).

This deliverable reports the LRTs which have been developed to contribute to high quality machine translation (Task 5.4), enabling the exploration of advanced semantic processing for machine translation. In particular, this deliverable documents the LRTs for 6 languages (BG Bulgarian, CS Czech, EN English, ES Spanish, EU Basque, PT Portuguese) that compose deliverable D5.8 "Final version of LRTs enhanced to support crosslingual ambiguity resolution". These LRTs are described in Appendix B, which summarizes all resources in D5.8, including previous releases under deliverable D5.3 and D5.5.

For easier reading, we include all relevant materials from D5.4 and D5.6, which means that this deliverable describes and documents all resources and tools included in WP5 throughout the project. The contents are organized as follows:

- Section 2: methodology to align ontologies.
- Section 3: the basic processing tools (PoS tagger, lemmatizer, Named Entity Recognition and Classification – NERC) used to annotate the corpora.
- Section 4: Word Sense Disambiguation (WSD), Named Entity Disambiguation (NED) and Correference (CR) tools used to annotate the corpora.
- Section 5: the annotated corpora released in D5.8.
- Section 6: evaluation of the basic tools.
- Section 7: evaluation of WSD.
- Section 8: evaluation of NED.
- Section 9: evaluation of CR.
- Section 10: further improvements of advanced tools.
- Section 11: final remarks.
- Appendix A: Sample annotations for basic tools.

- Appendix B: Summary of LRTs described in this deliverable, alongside availability information. P13

The workplan for the WP5 working package includes research on advanced processors until the end of the project. The results reported in Sections 7, 8 and 9 refer to the status of advanced processors used in D5.8. Section 10 reports further improvements on some of the advanced processors, including WSD, NED and NERC for several of the project languages. QTLeap will continue to develop some advanced processors until the end of the project, and the final results for those improvements will be reported in D5.11.

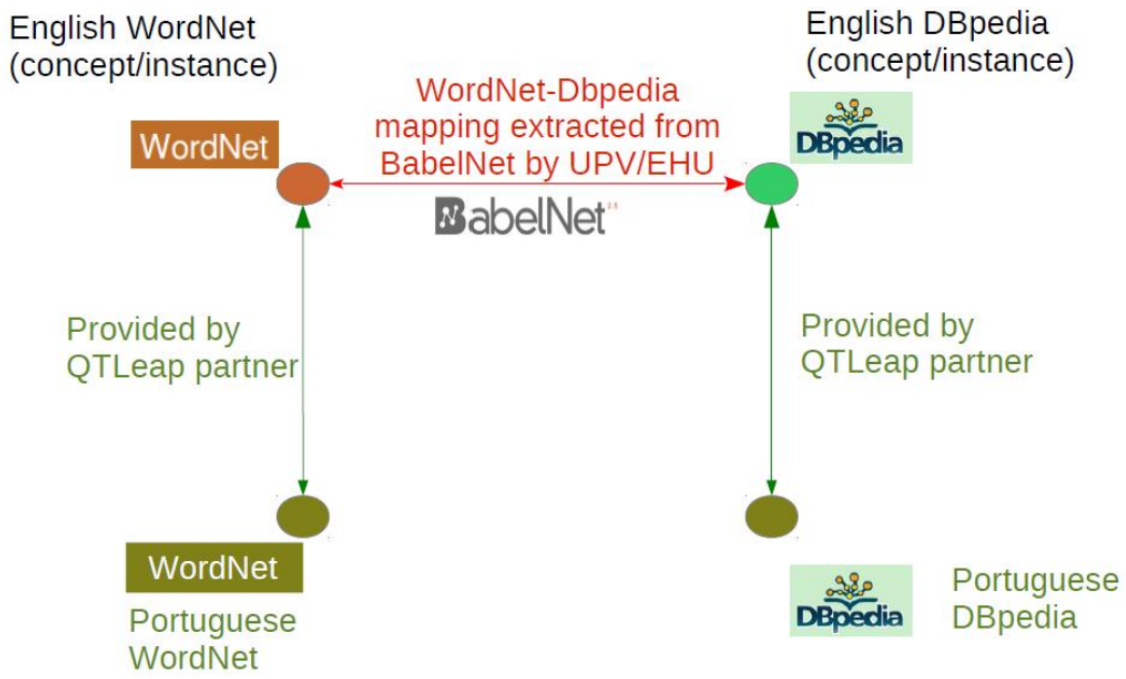


Figure 1: Example of figure of the ontology alignment procedure for a sample QTLeap language (Portuguese shown for illustration). The design for the other languages is analogous

2 Aligned Ontologies

This section describes the methodology to align the ontologies for all languages (T5.1).

2.1 Methodology to build the alignment

Our strategy is one of loose coupling, where each partner is responsible for its ontologies, and where QTLeap keeps a central inventory of concepts/senses based on English WordNet and DBpedia. Each partner needs to maintain the alignment of his resources to the English WordNet or DBpedia. In addition, UPV/EHU will provide an alignment between English WordNet URIs and DBpedia URIs (extracted from BabelNet, [Navigli and Ponzetto, 2012]).

Figure 2.1 shows the design, illustrated by the links from the Portuguese WordNet and the Portuguese DBpedia. The design for the rest of languages is analogous. In the figure, the Portuguese WordNet is aligned to the English WordNet using the alignments between both wordnets. The Portuguese DBpedia concepts and instances are mapped to the English DBpedia using the cross-lingual alignments provided by DBpedia. Finally, the English WordNet is aligned to the English DBpedia using the alignments provided by BabelNet.

The QTLeap list of interlingual concepts and instances will be composed of the union of the following:

- DBpedia v3.9 URI, based on the March-June 2013 dump `http://wiki.dbpedia.org/Downloads39`. This DBpedia release was the latest as of May 23rd, 2014. An example URI for an instance: `http://dbpedia.org/resource/Barack_Obama`

- English WordNet v3.0 URI, based on the Lemon model[1]. An example URI for a concept: http://lemon-model.net/lexica/pwn/wn30-09213565-n

These resources will be frozen, to allow for comparability alongside project development. Note that the Statistical Machine Translation (SMT) pilots developed throughout the project also use frozen datasets, which reduces the need to use newer versions of WordNet or DBpedia.

Each language will provide a mapping between their specific concept and entity ids (or URIs) to one of the following:

- DBpedia v3.9 URI
- English WordNet v3.0 URI

We discarded other alternatives like using Freebase URIs, but note that DBpedia provides a *sameAs* property which also includes Freebase URIs, allowing for interoperability with Freebase-based ontologies.

Note that there is no requirement for a common format for the local ontologies.

All the aligned ontologies are listed in the Appendix B.

2.2 Basque

WordNet and DBpedia are the ontologies used for Basque. The statistics for the versions which were current when they were used for the project are the following:

- WordNet 3.0 contains 30,615 synsets and 50,691 variants [Gonzalez-Agirre et al., 2012].
- DBpedia 3.9 contains 148,260 instances on the Basque localized data set and 118,662 on canonicalized data set.

2.3 Bulgarian

WordNet and DBpedia are the ontologies used for Bulgarian. The statistics for the versions which were current when they were used for the project are the following:

- BTB WordNet 3.1 contains 9,628 synsets, 15,704 words. It covers 100% of the Core WordNet[2] and the open class words in BulTreeBank.
- DBpedia 3.9 contains 71,117 instances on the Bulgarian localized data set. The main problem with Bulgarian data set of DBpedia is that important named entities are missing. For example, one of the recent presidents - Petar Stoyanov - is not presented there, while five other people with the same name are included. For that reason we have manually added some instances from Wikipedia using the appropriate classification of the DBpedia ontology. At the same time, semi-automatic transfer of such classifications from English DBpedia to Bulgarian Wikipedia missing URIs is in progress.

[1] http://lemon-model.net/

[2] http://compling.hss.ntu.edu.sg/omw

2.4 Czech

The statistics for the versions which were current when they were used for the project are the following:

- Czech BabelNet contains 646,000 lemmas, 410,000 synsets, 897 word senses[3].
- Czech DBpedia contains 225,000 localized data sets[4].
- Czech WordNet 1.9 captures nouns, verbs, adjectives, and partly adverbs, and contains 23,094 word senses (synsets). 203 of these were created or modified by UFAL CUNI during correction of annotations (http://hdl.handle.net/11858/00-097C-0000-0001-4880-3). This version of WordNet was used to annotate word senses in the Prague Dependency Treebank.

2.5 English

WordNet and DBpedia are the ontologies used for English. The statistics for the versions which were current when they were used for the project are the following:

- WordNet 3.0 contains 118,431 synsets and 207,995 variants [Gonzalez-Agirre et al., 2012].
- DBpedia 3.9 contains 4,004,478 instances[5].

BabelNet [Navigli and Ponzetto, 2012] was used to extract the mapping between WordNet and DBpedia. BabelNet contains 4,107,138 BabelNet synsets, 8,374,951 lemmas and 11,056,960 word senses[6], 206,941 WordNet variants and 10,719,133 DBpedia articles (including 4,854,205 redirects, 2,035,867 Wikidata articles). In addition BabelNet also includes 58,971 OmegaWiki and 71,915 Wiktionary entries. BabelNet combines WordNet and Wikipedia by automatically acquiring a mapping between WordNet senses and Wikipedia pages, avoiding duplicate concepts and allowing their inventories of concepts to complement each other.

We extracted the mapping between WordNet and DBpedia from BabelNet 2.5, obtaining the following statistics:

- 44,328 WordNet synsets
- 46,699 DBpedia instances
- 47,956 synset-instance pairs

The mapping is available in a text file in the QTLeap repository with the following format:

- WordNet 3.0 URI

[3] http://babelnet.org/stats.jsp

[4] http://wiki.dbpedia.org/Datasets/DatasetStatistics

[5] http://wiki.dbpedia.org/Datasets39/DatasetStatistics?v=dqp (accessed Sept. 2014). Note that DBpedia instances in this context might refer to concepts (e.g. http://dbpedia.org/resource/President) or actual instances in the ontological sense (e.g. http://dbpedia.org/resource/Barack_Obama).

[6] http://babelnet.org/stats version 2.5 (accessed Sept. 2014)

- tab
- DBpedia 3.9 URI

We also considered using the mappings provided[7] by Fernando and Stevenson [2012], but the quality reported in Navigli and Ponzetto [2012] compares favorably.

2.6 Portuguese

The wordnet MWN.PT - MultiWordNet of Portuguese is used for the work on the Portuguese language in WP5. The synsets in this wordnet have been manually aligned with the translationally equivalent concepts of the English Princeton WordNet (and, transitively, with the equivalent concepts in the MultiWordNets of Italian, Spanish, Hebrew, Romanian and Latin). As such, the alignment with the English WordNet arises naturally from the way MWN.PT is built.

MWN.PT - MultiWordnet of Portuguese (version 1) spans over 17,200 manually validated concepts/synsets, linked under the semantic relations of hyponymy and hypernymy. These concepts are made of over 21,000 word senses/word forms and 16,000 lemmas from both European and American variants of Portuguese. MWN.PT includes the subontologies under the concepts of Person, Organization, Event, Location, and Art works, which are covered by the top ontology made of the Portuguese equivalents to all concepts in the 4 top layers of the English Princeton WordNet and to the 98 Base Concepts suggested by the Global WordNet Association, and the 164 Core Base Concepts indicated by the EuroWordNet project. It is available at http://catalog.elra.info/product_info.php?products_id=1101.

DBpedia 3.9 for Portuguese contains 736,443 instances on the localized data set and 493,944 on the canonicalized data set.

2.7 Spanish

WordNet and DBpedia are the ontologies used for Spanish. The statistics for the versions which were current when they were used for the project are the following:

- WordNet 3.0 contains 59,227 synsets and 59,227 variants [Gonzalez-Agirre et al., 2012].
- DBpedia 3.9 contains 964,838 instances on the Spanish localized data set and 601,258 on canonicalized data set.

[7] http://staffwww.dcs.shef.ac.uk/people/S.Fernando/resources.shtml

3 Basic Processing Tools

This section describes the state-of-the-art basic processing tools for all languages (T5.1), as follows:

- PoS Tagger
- Lemmatizer
- NERC module

Basic tools for English are provided by UPV/EHU and by CUNI as the processing of language pairs X↔EN may be carried out by different partners. The partners can use either set of tools, and note that the NED, WSD and CR tools in Section 5 are interoperable with the tools provided by UPV/EHU.

The evaluation section will show that our basic processing tools are state-of-the-art when compared to freely available Natural Language Processing (NLP) pipelines.

All the basic processing tools are listed in the Appendix B.

3.1 Basque

3.1.1 PoS tagger and lemmatizer

ixa-pipe-pos-eu [Alegria et al., 2002] is a robust and wide-coverage morphological analyser and a Part-of-Speech tagger for Basque. The analyser is based on the two-level formalism and has been designed in an incremental way with three main modules: the standard analyser, the analyser of linguistic variants, and the analyser without lexicon which can recognize word-forms without having their lemmas in the lexicon. ixa-pipe-pos-eu provides the lemma, PoS and morphological information for each token. It also recognizes date/time expressions, numbers. In the tagger, combination of stochastic and rule-based disambiguation methods is applied to Basque. The methods we have used in disambiguation are Constraint Grammar formalism and an HMM based tagger.

The module reads raw text and outputs a file in Natural Language Processing Annotation Format (NAF) [Fokkens et al., 2014].

The tool is released under license GPLv3.0[8]. The tool is partly funded by QTLeap, as the wrapper to produce NAF has been developed in this project.

3.1.2 NERC

The module ixa-pipe-nerc is multilingual Named Entity Recognition and Classification tagger, and is part of IXA pipes tool (see Section 3.4.1). The named entity types are based on: a) the CONLL 2002[9] and 2003[10] tasks which were focused on language-independent supervised named entity recognition for four types of named entities: persons, locations, organizations and names of miscellaneous entities that do not belong to the previous three groups. We provide very fast models trained on local features only, similar to those of Zhang and Johnson [2003] with several differences: We do not use PoS tags, chunking or gazetteers in our baseline models but we do use bigrams, trigrams and character n-grams.

[8] http://ixa2.si.ehu.es/ixa-pipes/eu/ixa-pipe-pos-eu.tar.gz
[9] http://www.clips.ua.ac.be/conll2002/ner/
[10] http://www.clips.ua.ac.be/conll2003/ner/

The module reads lemmatized and PoS tagged text in NAF format. The module allows to format its output in NAF and CoNLL style tabulated BIO format as specified in the CoNLL 2003 shared evaluation task.

The tool is released under the Apache License 2.0 (APL 2.0)[11]. The tool has been developed independently from QTLeap.

3.2 Bulgarian

These two components of the Bulgarian pipeline existed before the start of the QTLeap project. They were minimally extended with domain specific lexica.

The Bulgarian pipeline is distributed as a program with all modules. Thus it has a license that covers the whole architecture: GPL v3.0.

3.2.1 PoS tagger and lemmatizer

The Bulgarian PoS tagger is hybrid. It uses a rich morphological dictionary, a set of linguistic rules and a statistical component. It assigns tags from a rich tagset, which encodes detailed information about the morphosyntactic properties of each word [Simov et al., 2004]. The task of choosing the correct tag is carried out by the guided learning system described in Georgiev et al. [2012] - GTagger, and by a rule-based module which utilizes a large morphological lexicon and disambiguation rules [Simov and Osenova, 2001]. It performs with 97% accuracy on news data.

Lemmatization module is based on rules, generated using this morphological lexicon. It performs with 95% accuracy.

3.2.2 NERC

The Bulgarian NERC is a rule-based module. It uses a gazetteer with names categorized in four types: Person, Location, Organization, Other. The identification of new names is based on two factors - sure positions in the text and classifying contextual information, such as, titles for persons, types of geographical objects or organizations, etc.

The disambiguation module uses simple unigram-based statistics.

3.3 Czech

3.3.1 PoS tagger and lemmatizer

MorphoDiTa[12] is an open-source tool for morphological analysis of natural language texts. It performs morphological analysis, morphological generation, tagging and tokenization and is distributed as a standalone tool or a library, along with trained linguistic models. For the Czech language, MorphoDiTa achieves state-of-the-art results while reaching a throughput of around 10-200K words per second.

The tool is released under the CC BY-NC-SA 3.0. The tool has been developed independently from QTLeap.

[11] https://github.com/ixa-ehu/ixa-pipe-nerc/

[12] http://ufal.mff.cuni.cz/morphodita

3.3.2 NERC

NameTag[13] is an open-source tool for NER. NameTag identifies proper names in text and classifies them into predefined categories, such as names of persons, locations, organizations, etc. For Czech, entities are classified into two-level hierarchy of categories consisting of 42 fine-grained categories merged into 7 super-classes. NameTag is distributed as a standalone tool or a library, along with trained linguistic models. In the Czech language, NameTag achieves state-of-the-art performance [Straková et al., 2013].

The tool is released under the CC BY-NC-SA 3.0. The tool has been developed independently from QTLeap.

3.4 English and Spanish

3.4.1 IXA pipes tool

IXA pipes is a modular set of Natural Language Processing tools (or pipes) which provide easy access to NLP technology for English and Spanish[14]. It provides ready to use modules to perform efficient and accurate linguistic annotation (PoS tagger, lemmatizer and NERC among others). The data format in which both the input and output of the modules needs to be formatted to represent and pipe linguistic annotations is NAF[15]. Our Java modules all use the kaflib[16] library for easy NAF integration. It has an active mailing-list for users.

The NLP processing for English and Spanish is the same as they both share the modules to perform the processing.

PoS tagger and lemmatizer The module ixa-pipe-pos provides PoS tagging and lemmatization for English and Spanish. We have obtained the best results so far with Perceptron models and the same feature set as in Collins [2002].

Lemmatization is currently performed via 3 different dictionary lookup methods: a) Simple Lemmatizer: It is based on HashMap lookups on a plain text dictionary. Currently we use dictionaries from the LanguageTool project[17] under their distribution licenses; b) Morfologik-stemming:[18] The Morfologik library provides routines to produce binary dictionaries, from dictionaries such as the one used by the Simple Lemmatizer above, as finite state automata. This method is convenient whenever look-ups on very large dictionaries are required because it reduces the memory foot-print to 10% of the memory required for the equivalent plain text dictionary; and c) We also provide lemmatization by look-up in WordNet-3.0 [Fellbaum, 1998] via the JWNL API[19].

Regarding to Spanish, lemmatization is performed via 2 different dictionary look-up methods (methods a and b described above).

By default, the module accepts tokenized text in NAF format as standard input and outputs NAF or CoNLL formats, with lemmas and PoS-tags.

The tool is released under the Apache License 2.0 (APL 2.0)[20]. The tool has been developed independently from QTLeap.

[13] http://ufal.mff.cuni.cz/nametag
[14] http://ixa2.si.ehu.es/ixa-pipes/
[15] http://wordpress.let.vupr.nl/naf/
[16] https://github.com/ixa-ehu/kaflib
[17] http://languagetool.org/
[18] https://github.com/morfologik/morfologik-stemming
[19] http://jwordnet.sourceforge.net/
[20] https://github.com/ixa-ehu/ixa-pipe-pos

NERC The module ixa-pipe-nerc is multilingual Named Entity Recognition and Classification tagger. ixa-pipe-nerc is part of IXA pipes. The named entity types are based on: a) the CONLL 2002[21] and 2003[22] tasks which were focused on language-independent supervised named entity recognition for four types of named entities: persons, locations, organizations and names of miscellaneous entities that do not belong to the previous three groups. We provide very fast models trained on local features only, similar to those of Zhang and Johnson [2003] with several differences: We do not use PoS tags, chunking or gazetteers in our baseline models but we do use bigrams, trigrams and character n-grams.

For English, we also provide some models with external knowledge; b) the Ontonotes 4.0 dataset. We have trained our system on the full corpus with the 18 named entity types, suitable for production use. We have also used 5K sentences at random for testset from the corpus and leaving the rest (90K approx) for training.

The module reads lemmatized and PoS tagged text in NAF format. The module allows to format its output in NAF and CoNLL style tabulated BIO format as specified in the CoNLL 2003 shared evaluation task.

The tool is released under the Apache License 2.0 (APL 2.0)[23]. The tool has been developed independently from QTLeap.

3.4.2 Treex

The Treex framework provides a full pipeline for English analysis. This pipeline integrates inter alia MorphoDiTa and NameTag tools.

PoS tagger and lemmatizer MorphoDiTa[24] (Morphological Dictionary and Tagger) is an open-source tool for morphological analysis of natural language texts. It performs morphological analysis, morphological generation, tagging and tokenization and is distributed as a standalone tool or a library, along with trained linguistic models.

NERC NameTag[25] is an open-source tool for named entity recognition (NER). NameTag identifies proper names in text and classifies them into predefined categories, such as names of persons, locations, organizations, etc. NameTag is distributed as a standalone tool or a library, along with trained linguistic models.

3.5 Portuguese

3.5.1 PoS tagger and lemmatizer

LX-Suite [Branco and Silva, 2006a] is composed by the set of shallow processing tools briefly described below.

LX-Chunker: Marks sentence boundaries with <s>...</s>, and paragraph boundaries with ... . Unwraps sentences split over different lines.

LX-Tokenizer: Besides the separation of words, this tools expands contractions.: do → |de_|o| It detaches clitic pronouns from the verb and the detached pronoun is marked with a - (hyphen) symbol.

[21] http://www.clips.ua.ac.be/conll2002/ner/
[22] http://www.clips.ua.ac.be/conll2003/ner/
[23] https://github.com/ixa-ehu/ixa-pipe-nerc/
[24] http://ufal.mff.cuni.cz/morphodita
[25] http://ufal.mff.cuni.cz/nametag

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

This tool also handles ambiguous strings. These are words that, depending on their particular occurrence, can be tokenized in different ways. For instance:

deste → |deste| when occurring as a Verb
deste → |de|este| when occurring as a contraction (Preposition + Demonstrative)

LX-Tagger: Assigns a single morpho-syntactic tag to every token:

um exemplo → um/IA exemplo/CN

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

de maneira a que → de/LCJ1 maneira/LCJ2 a/LCJ3 que/LCJ4

This tagger was developed over Hidden Markov Models technology.

LX-Featurizer (nominal): Assigns inflection feature values to words from the nominal categories, namely Gender (masculine or feminine), Number (singular or plural) and, when applicable, Person (1st, 2nd and 3rd):

os/DA gatos/CN → os/DA#mp gatos/CN#mp

It also assigns degree feature values (diminutive, superlative and comparative) to words from the nominal categories:

os/DA gatinhos/CN → os/DA#mp gatinhos/CN#mp-dim

LX-Lemmatizer (nominal): Assigns a lemma to words from the nominal categories (Adjectives, Common Nouns and Past Participles):

gatas/CN#fp → gatas/GATO/CN#fp
normalíssimo/ADJ#ms-sup → normalíssimo/NORMAL/ADJ#ms-sup

LX-Lemmatizer and Featurizer (verbal): Assigns a lemma and inflection feature values to verbs.

escrevi/V → escrevi/ESCREVER/V#ppi-1s

This tool disambiguates among the various lemma-inflection pairs that can be assigned to a verb form.

LX-Suite has been developed independently from QTLeap and is not available.

3.5.2 NERC

LX-NER is a NERC tools that identifies, circumscribes and classifies the expressions for named entities. It handles the following types of expressions: Numbers (Arabic, Decimal, Non-compliant, Roman, Cardinal, Fraction, Magnitude class, Measures (Currency, Time, Scientific units), Time (Date, Time periods, Time of the day) and Addresses) and name-based expressions (Persons, Organizations, Locations, Events, Works, Miscellaneous). The number-based component is built upon handcrafted regular expressions. It was developed and evaluated against a manually constructed test-suite including over 300 examples. The name-based component is based on Hidden Markov Models technology and was trained over a manually annotated corpus of approximately 208,000 words.

LX-NER has been developed independently from QTLeap and is not available.

4 NED, WSD and Coreference tools

This section describes the advanced tools, namely NED, WSD and CR. The availability information is summarized in the Appendix B.

4.1 Basque

4.1.1 NED

The ixa-pipe-ned-ukb module performs the Named Entity Disambiguation (NED) task based on UKB, a graph-based Word Sense Disambiguation (WSD) tool (see next section). In this case, the Wikipedia graph built from the hyperlinks between Wikipedia articles is used for the processing. This tool was successfully used for English NED [Agirre et al., 2015].

The input of the module is text where named entity mentions have been recognized and represented using the Natural Language Processing Annotation Format (NAF)[Fokkens et al., 2014][26][27]. In the output it returns the corresponding Basque Wikipedia in NAF format.

The tool is released under license GPLv3.0[28]. The tool is partly funded by QTLeap, as the wrapper to read and produce NAF has been developed in this project.

4.1.2 WSD

UKB is a collection of programs for performing graph-based Word Sense Disambiguation[29]. It applies the so-called Personalized PageRank on a Knowledge Base (KB) to rank the vertices of the KB and thus perform disambiguation. We used WordNet 3.0 as the KB for performing WSD. We run the tool with too variants: using a uniform distribution across senses, and using the distribution of senses attested in training corpora.

ixa-pipe-wsd-ukb takes lemmatized and PoS tagged text in NAF format as standard input and outputs NAF. The tool is released under license GPLv3.0, packaged with the resources to run it on Basque[30]. The tool has been developed independently from QTLeap.

4.1.3 Coreference

The module of Basque coreference resolution (ixa-pipe-coref-eu) is an adaptation of the Stanford Deterministic Coreference Resolution [Lee et al., 2013], which gives state-of-the-art performance for English. The original system applies a succession of ten independent deterministic coreference models or sieves. During the adaptation process, firstly, a baseline system has been created which receives as input texts processed by Basque analysis tools and uses specifically adapted static lists to identify language dependent features like gender, animacy or number. Afterwards, improvements over the baseline system have been applied, adapting and replacing some of the original sieves, taking into account that morpho-syntactic features are crucial in the design of the sieves for agglutinative languages like Basque.

[26] http://wordpress.let.vupr.nl/naf/
[27] https://github.com/newsreader/NAF/blob/master/naf.pdf?raw=true
[28] http://ixa2.si.ehu.eus/ixa-pipes/eu/ixa-pipe-ned-ukb.tar.gz
[29] https://github.com/asoroa/naf_ukb
[30] http://ixa2.si.ehu.eus/ixa-pipes/eu/ixa-pipe-wsd-ukb.tar.gz

The module needs a NAF document annotated with lemmas, entities and constituents, and outputs a NAF document.

The tool is released under license GPLv3.0[31]. The tool is partly funded by QTLeap, as the wrapper to read and produce NAF has been developed in this project.

4.2 Bulgarian

The three tasks were new for Bulgarian, thus, we have adopted two approaches to implement them. First, training of existing tools by third parties on Bulgarian data, and second, implementation of rule-based components over the output of the Bulgarian pipeline. The first approach gave very poor results. The reason for this, in our view, is the lack of enough annotated data for Bulgarian. For that reason we proceeded with rule-based modules. For NED and WSD we are exploiting the annotation of BulTreeBank treebank of Bulgarian with instances from DBpedia and senses from BTB Bulgarian WordNet. These annotations are done within the project.

All the modules were developed within the project. They are distributed as part of the Bulgarian pipeline under license GPL v3.0.

4.2.1 NED

It is an unfortunate fact that DBpedia Spotlight[32] does not support Bulgarian. Thus, the module for NED for Bulgarian is implemented by ourselves via ranking DBpedia instances with respect to their frequency in the manually annotated corpus within the project. After the application of the Bulgarian pipeline each instance in DBpedia was classified with respect to several classes in DBpedia's ontology (such as City, Politician, etc.). For the classification in the text we use the most general category in the ontology. These classes were manually mapped to WordNet, thus, they comply with the WSD module in the pipeline. In cases of several candidates for a given name, we selected the most frequent one.

The input is the result from the POS tagger and the lemmatizer for Bulgarian as well as the classification of Named Entities with respect to the DBpedia categories.

The output is converted to NAF format[33].

4.2.2 WSD

The first version of WSD was implemented on the basis of the frequency of the synsets in an annotated corpus with senses from BTB Bulgarian WordNet. We assumed that the realization of senses in text follows the assumption of one sense per discourse. Additionally, we performed statistics on bigrams. This approach gives relatively high accuracy, but the coverage is low since the annotated corpus is relatively small (BulTreeBank). At a later stage of the project, when the BTB WordNet showed better coverage (more than 12000 synsets) we relied on the UKB system for knowledge-based WSD. Thus, better coverage of the annotation was achieved, but the accuracy dropped.

[31] http://ixa2.si.ehu.eus/ixa-pipes/eu/ixa-pipe-coref-eu.tar.gz
[32] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki
[33] http://www.newsreader-project.eu/files/2013/01/techreport.pdf

4.2.3 Coreference

We have implemented a version of a coreference resolution module, using paths in the dependency tree of each sentence. By using the path patterns, we mainly focused on anaphora resolution. When dealing with the rest of the word forms, we consider the open class words that belong to the same synsets in WordNet grouping them together. Similarly for the named entities.

Additionally, we tried to train the RelaxCor system[34], but there was no successful result. Thus, we abandoned this approach.

4.3 Czech

All the tools mentioned below produce results in Treex and NAF formats.

4.3.1 NED

At the beginning of the project, there was no publicly available implementation for NED of Czech. During the preparation phase for Named Entities Disambiguation task we created the Named Entities Linking table. Each row of that table consisted of the lemmatized Czech Wikipedia article's title, Czech Wikipedia URL and English DBpedia URL. In order to make this table we downloaded 2 dumps: Czech Wikipedia dump (containing Czech titles and corresponding Wikipedia URLs), English-Czech DBpedia dump (containing Czech labels and English DBpedia URLs). Czech labels from the DBpedia were mapped to the titles of corresponding Czech Wikipedia articles thus creating the resulting table. We additionally applied lemmatization and tagging for each title using MorphoDiTa [Straková et al., 2014].

Named Entity Disambiguation was done in two steps. During the first step we used the Treex block for the NameTag tool [Straková et al., 2014] to detect named entities in the corpus. During the second step we used the previously created Named Entities Linking table in the following way: for each entity that was detected by NameTag we lemmatized its form and then searched the table for the occurrences of this lemmatized form. We used lemmatization to resolve the problem of forms' inflection. If the search returned results, we looked for the DBpedia URL and labeled the entity if we could find one. In case of ambiguity the algorithm picked up "most popular" article. The popularity of the article was computed using Wikipedia page-to-page link records, so the article with the highest number of reference links was preferred. We are looking forward to further improving of the algorithm by adding the context from the Wikipedia articles.

The development of Treex wrappers for MorphoDiTa and NameTag is partly funded by the project. They are available under open-source license (Perl Artistic + GPL) at GitHub repository.[35] The tools (MorphoDiTa and NameTag) themselves are also open source (LGPL) and available from GitHub or `http://www.lindat.cz/`.

4.3.2 WSD

Experiments in Czech WSD [Honetschläger, 2003, Semecký, 2007, Hajič et al., 2009] typically use the Prague Dependency Treebank (PDT) [Hajič et al., 2006, Bejček et al., 2012], which provides valency frame reference annotation, i.e., word sense labeling for all

[34] `http://nlp.lsi.upc.edu/relaxcor/`
[35] `https://github.com/ufal/treex/`

verbs and many other content words. PDT word senses are based on the PDT-Vallex Czech valency lexicon [Hajič et al., 2003, Urešová, 2011]. A mapping [Urešová et al., 2015][36] connects it to the EngVallex valency lexicon [Cinková, 2006], which itself contains links to PropBank and can thus be mapped to English WordNet [Pazienza et al., 2006]. Note that PDT word senses do not form a hierarchy, which makes it incompatible with graph-based WSD (see Sections 4.1.2 and 4.5.2).

Although there is a WordNet for Czech [Pala et al., 2011], it is typically not used for WSD tasks. It is based on an outdated version of the Princeton WordNet (2.0) and it has been further modified, and so its mapping onto current English WordNet is not trivial.

The Czech WSD annotation developed herein uses two approaches: First, a tool based on [Dušek et al., 2014], which uses the VowpalWabbit linear classifier on top of automatic deep syntactic analysis and achieves high performance for verbal WSD on PDT data. This is used for real user scenarios.[37] Second, for the WSD-annotated parallel corpus, we opted for a more straightforward way of achieving compatibility with English WordNet IDs: since the corpus contains the same sentences as the EN-ES parallel corpus provided in D5.3, we could use the English WordNet ID annotation from this corpus and project it onto Czech words using GIZA++ word alignment. This method was manually evaluated in 7.3.

4.3.3 Coreference

The coreference resolution system for Czech consists of multiple modules, each of them aiming at a specific type of coreference: coreference of reflexive pronouns, relative pronouns, zeros, personal and possessive pronouns in 3rd person and coreference of noun phrases. Coreference relations are annotated between the nodes of dependency trees that serve as a deep syntax representation of sentences. This enables the system to take advantage of rich linguistic annotations available in the trees as well as to resolve coreference even for subject pronouns dropped from the surface representation (zeros), which is a common practice in Czech.

Due to the pro-drop nature of Czech, the places where a subject is unexpressed have to be identified before proceeding to coreference resolution of zeros. This is performed based on the syntactic information and a special node representing the zero is added to the deep syntax tree. The grammatical categories of the newly added zero are then inferred from the grammatical categories of its governing verb. Reconstruction of zeros is implemented in the Treex block `A2T::CS::AddPersPron`.

The modules targeting coreference of *relative and reflexive pronouns* are based on the rules presented in Nguy [2006]. The rules exploit morphological information together with syntactic structure and stick to the principle that the antecedent of a reflexive pronoun is the sentence's subject whereas the antecedent of the relative pronoun usually directly governs the relative clause introduced by the pronoun. These resolvers are implemented in Treex blocks `A2T::CS::MarkRelClauseCoref` and `A2T::CS::MarkReflpronCoref`.

Unlike the previous cases, resolution of *personal and possessive pronouns and zeros in 3rd person* is treated by a machine learning approach. It adheres to a so-called mention-ranking model [Denis and Baldridge, 2007] with features capturing the distance between the mentions (in words, clauses and sentences), grammatical information (e.g., agree-

[36] http://ufal.mff.cuni.cz/czengvallex

[37] The tool is implemented as a Treex block `A2T::SetValencyFrameRefVW` and is available under Perl Artistic and GPL license in the Treex Git repository: https://github.com/ufal/treex/.

ment in their numbers and genders) as well as semantic information (semantic roles, classes in the Czech part of EuroWordNet [Vossen, 1998]). The currently used model is built using logistic regression in Vowpal Wabbit[38] and is available in the Treex block `A2T::CS::MarkTextPronCoref`. A more detailed description and evaluation can be found in [Nguy et al., 2009, Bojar et al., 2012].

Coreference of *noun phrases* is modeled in the same way as the pronouns and zeros in the previous case. However, the feature set is more oriented on lexical features (equality of head lemmas), semantic features (synonymy approximation extracted from the English-Czech parallel corpus CzEng 0.9 [Bojar et al., 2009], EuroWordNet classes) and the information about named entities. Unlike the previous modules, the module for noun phrases does not have a Treex binding, yet. A more detailed description and evaluation can be found in Novák and Žabokrtský [2011].

All modules are available under open-source license (Perl Artistic + GPL) and the Treex blocks can be downloaded from its Git repository.[39]

4.4 English and Spanish

The NLP processing for English and Spanish is the same as they both share the modules to perform the processing.

4.4.1 NED

The ixa-pipe-ned module performs the Named Entity Disambiguation task based on DBpedia Spotlight[40]. Assuming that a DBpedia Spotlight Rest server for a given language is locally running, the module will take NAF as input (containing elements) and perform Named Entity Disambiguation. The module offers the "disambiguate" and "candidates" service endpoints. The former takes the spotted text input and it returns the identifier for each entity. The later is similar to disambiguate, but returns a ranked list of candidates.

The module accepts text with named entities in NAF format as standard input, it disambiguates them and outputs them in NAF.

The tool is released under license GPLv3.0[41]. The tool has been developed independently from QTLeap.

In addition, we also tested the ixa-pipe-ned-ukb module which we introduced for Basque NED (see Section 4.1.1).

4.4.2 WSD

UKB is a collection of programs for performing graph-based Word Sense Disambiguation[42]. UKB applies the so-called Personalized PageRank on a Lexical Knowledge Base (LKB) to rank the vertices of the LKB and thus perform disambiguation. WordNet will be the LKB used for this processing. For English, we checked the tool using either a uniform distribution of senses, or the distribution of senses learned from Semcor Miller et al. [1993]. For Spanish, We run the tool using either the distribution of senses learned from training data, or a uniform distribution of senses.

[38] https://github.com/JohnLangford/vowpal_wabbit
[39] https://github.com/ufal/treex/
[40] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki
[41] https://github.com/ixa-ehu/ixa-pipe-ned
[42] https://github.com/asoroa/naf_ukb

ixa-pipe-wsd-ukb accepts lemmatized and PoS tagged text in NAF format as standard input and outputs NAF.

The tool is released under license GPLv3.0, packaged with the resources to run it on English and Spanish[43]. The tool has been developed independently from QTLeap.

4.4.3 Coreference

The module of coreference resolution (ixa-pipe-coref) included in the IXA pipeline is loosely based on the Stanford Multi Sieve Pass system [Lee et al., 2013]. The system consists of a number of rule-based sieves. Each sieve pass is applied in a deterministic manner, reusing the information generated by the previous sieve and the mention processing. The order in which the sieves are applied favors a highest precision approach and aims at improving the recall with the subsequent application of each of the sieve passes. This is illustrated by the evaluation results of the CoNLL 2011 Coreference Evaluation task [Lee et al., 2013, 2011], in which the Stanford's system obtained the best results. The results show a pattern which has also been shown in other results reported with other evaluation sets [Raghunathan et al., 2010], namely, the fact that a large part of the performance of the multi pass sieve system is based on a set of significant sieves. Thus, this module focuses for the time being, on a subset of sieves only, namely, Speaker Match, Exact Match, Precise Constructs, Strict Head Match and Pronoun Match [Lee et al., 2013].

The module needs a NAF document annotated with lemmas, entities and constituents, and outputs a NAF document.

The tool is released under the Apache License 2.0 (APL 2.0)[44]. The tool has been developed independently from QTLeap.

4.5 Portuguese

The following modules for Portuguese were developed within the QTLeap project. They are distributed through META-SHARE under the Apache License 2.0.

4.5.1 NED

The named entity disambiguation pipeline for Portuguese uses DBpedia Spotlight [Daiber et al., 2013] to find links to resources about entities identified in pre-processed input text. It creates a process to run a Portuguese extraction of DBpedia Spotlight on a local server, then takes an input text pre-processed with lemmas, Part of Speech tags and named entities using the LX-Suite [Branco and Silva, 2006b] and converts it to the 'spotted' format understood by Spotlight. This spotted input text is then disambiguated using DBpedia Spotlight, returning among other information links to existing Portuguese DBpedia resource pages for each named entity discovered.

For each Portuguese DBpedia resource page link found, the tool performs a DBpedia sparql query to find any English words that the link in question relates back to. These results can then be used to determine the corresponding English DBpedia resource page link, for example: `http://pt.dbpedia.org/resource/Paquistao` relates to 'Pakistan', thus the equivalent link in English must be `http://dbpedia.org/resource/Pakistan`. This process has been found to return working English resource links in almost all cases,

[43] http://ixa2.si.ehu.eus/ixa-pipes/eu/ixa-pipe-wsd-ukb.tar.gz
[44] https://bitbucket.org/Josu/corefgraph

with the exception of Portuguese resource links that despite existing contain no actual information (having perhaps been corrupted, or created and then for some reason deleted later).

The output displays each potential named entity found in the input text with: its positional offsets (sentence and position within sentence); the disambiguated Portuguese DBpedia resource link (if found); and the corresponding English DBpedia resource link (if found).

4.5.2 WSD

For WSD, a pipeline was used that takes pre-processed input text and runs it through the UKB word-sense disambiguation algorithm [Agirre and Soroa, 2009]. The pre-processed texts, .txt files lemmatized and PoS-tagged using the LX-Suite [Branco and Silva, 2006b], are passed as an argument to the pipeline, which converts the text to the context format recognized by UKB. The Lexical Knowledge Base (LKB) from which UKB returns word senses within the pipeline has been generated from an extraction of the Portuguese MultiWordNet [MultiWordNet, n.d.].

The output displays each potentially ambiguous word (noun, verb, adjective or adverb) found in the input text with: incrementing ID numbers; its UKB context (sentence number); its UKB word id (position within sentence); its part-of-speech; its lemma; whether or not it was tagged by UKB; the Portuguese MultiWordNet sense returned by UKB; and the ILI code.

4.5.3 Coreference

As an initial study for the coreference pipeline, a decision tree classifier was experimented with. Given a pair of expressions (markables), the classifier returns a true or false value that indicates whether those expressions are coreferent. The classifier uses the J48 algorithm in the Weka machine-learning toolkit [Hall et al., 2009].

For training and evaluating the classifier we used a small portion of CINTIL corpus [Barreto et al., 2006] that was manually annotated with coreference chains using the MMAX2 tool [Müller and Strube, 2006].

The most relevant features for coreference resolution, according to the work of de Souza et al. [2008], are the following:

- cores-match, which indicates whether the "cores" of the two expressions (i.e. their syntactic heads) have the same form;
- gender-agreement, which indicates whether the cores of the two expressions have matching gender;
- number-agreement, which is similar to the previous one, but for number;
- distance, which indicates the distance, in sentences, between the two expressions (if the two expressions occur in the same sentence, their distance is zero);
- antecedent-is-pron, which indicates whether the antecedent (the first markable) is a pronoun;
- anaphora-is-pron, which indicates whether the anaphora (the second markable) is a pronoun;

- both-proper-names, which indicates whether both markables are proper names; P30

These features were extracted from the result of an automatic parsing of the corpus, cross-referenced with the manual coreference annotation done in MMAX2, and then used to train the J48 algorithm, using the default parameters in Weka.

4.6 Harmonisation and crosslingual ambiguity resolution

The basic tools in each language use a different set of labels, following different linguistic principles, creating inter-operability issues. Fortunately, there has been previous work on harmonizing the output of linguistic tools, which is reused in this project, as follows:

- PoS tags and syntactic tags: HamleDT[45] provides harmonized treebanks for all project languages.
- NERC tags: all languages and annotations schemes provide three common tags, person, location and organization.

Regarding WSD and NED, the alignment of the ontologies described in Section 2 allows crosslingual ambiguity resolution. Assuming that concepts and instances are shared across languages and cultures, at least to a great extent, it is in theory possible to construct a common repository of concepts and instances. Following our design for aligned ontologies, WSD and NED tools return, respectively interlingual concept identifiers and instance identifiers, as follows:

- The interlingual concept id's are inspired in EuroWordNet [Vossen, 1998], which presented the design of a multilingual database with wordnets for several languages. The design was based on the ILI, based on the English wordnet. Via this index, the languages are interconnected at the senses level, so that it is possible to go from the words in one language to similar words in any other language via equivalent senses. Current ILIs are based on the English WordNet 3.0 synset numbers, and are strings like the following: `ili-30-05799212-n`, where `30` stands for the 3.0 WordNet version, `05799212` corresponds to the English WordNet 3.0 synset number, and `-n` to the PoS of the synset. Note that the synset number in the ILI has 9 digits, which are obtained appending 0 to the 8 digits of the WordNet 3.0 synsets. This allows some room to incorporate concepts which are not found in the English WordNet, although, given the fact that all translations involve English, this possibility is not needed in QTLeap.
- Instance id's are based on English DBpedia v3.9 URIs.

Producing ILI's is straightforward in the cases where the wordnets are aligned to the English version. This is possible for all the languages covered in our work.

Producing English DBpedia v3.9 URIs is also straightforward for NED tools, as DBpedia maintains interlingual links between articles in different languages.

[45] https://ufal.mff.cuni.cz/hamledt

5 Annotated corpora

This section describes the corpora which have been annotated with the WP5 tools mentioned above. All the corpora below have been packaged in two multilingual corpora released through meta-share[46] and CLARIN Lindat[47]

- Europarl-QTLeap WDS/NED corpus. In addition to BG, CS, EN, ES, PT subsets of the Europarl parallel corpus[48] [Koehn, 2005], it contains an EN-EU parallel corpus from non-Europarl sources.
- QTLeap WSD/NED corpus. It contains Batches 1 and 2 of the QTLeap corpus[49] annotated.
- News-QTLeap WSD/NED corpus. It contains news texts from WMT translated to project languages and used for out-of-domain evaluation in the project.

Note that, due to licensing restrictions, we are only allowed to redistribute part of the EN-EU dataset. The rest of the EN-EU dataset is available to project partners in the project internal repository.

The Europarl-QTLeap WSD/NED corpus is distributed under the license CC BY 4.0.

The QTLeap WSD/NED corpus is distributed under the license CC BY-NC-SA 4.0.

The News-QTLeap WSD/NED corpus is distributed under the license CC BY 4.0.

For all language pairs, parallel corpora were annotated instead of comparable corpora (as initially planned in the DoW), as this provides better quality for training machine translation systems.

5.1 Basque-English

Given that Europarl does not include Basque, we gathered publicly available and private corpora. Regarding publicly available corpora, we focus on the GNOME corpus [Tiedemann, 2012]. Regarding private corpora, we have access to the translation memories of Elhuyar Foundation (obtained via Eleka, member of the Advisory Board of Potential Users), which we cannot redistribute.

Prior to being annotated with ixa-pipe-wsd-ukb, ixa-pipe-ned-ukb and ixa-pipe-coref-eu, we preprocessed the corpus with the tools described in D5.4. Thus, the annotation in each document is an output of the following annotation process:

- Tokenization
- Part of Speech tagger and lemmatization
- Named Entity Recognition and Classification
- Named Entity Disambiguation
- Word Sense Disambiguation

[46] http://metashare.metanet4u.eu/go2/europarl-qtleap-wsdned-corpus
http://metashare.metanet4u.eu/go2/qtleap-wsdned-corpus
http://metashare.metanet4u.eu/go2/news-qtleap-wsdned-corpus

[47] https://lindat.mff.cuni.cz/

[48] http://www.statmt.org/europarl/

[49] The QTLeap corpus is described in deliverable D2.5

- Dependency parser
- Coreference

These are the annotated corpora:

- Parallel corpus
  - Elhuyar-QTLeap WSD/NED corpus (private). Table 1.

Corpus	EU
tokens	10,639,863
terms / linked to WordNet	10,639,863 / 4,725,833 (44.42%)
entities / linked to DBpedia	130,119 / 50,345 (38.69%)
coreference chains	1,551,340

Table 1: Statistics on Elhuyar-QTLeap WSD/NED corpus (Basque)

  - GNOME section of the Europarl-QTLeap WSD/NED corpus (public). Table 2.

Corpus	EU
tokens	4,194,823
terms / linked to WordNet	4,194,823 / 1,940,424 (46.26%)
entities / linked to DBpedia	45,801 / 21,118 (46.11%)
coreference chains	563,570

Table 2: Statistics on the GNOME section of the Europarl-QTLeap WSD/NED corpus (Basque)

  - QTLeap WSD/NED corpus: batch 1 and 2. Table 3

Corpus	EU
tokens	53,239
terms / linked to WordNet	53,239 / 24,691 (46.38%)
entities / linked to DBpedia	869 / 252 (29.00%)
coreference chains	5,542

Table 3: Statistics on QTLeap WSD/NED corpus (Basque)

  - News-QTLeap WSD/NED corpus. The Basque translation of the QTLeap news corpus used in Pilot 2. Table 4.

5.2 Bulgarian-English

The Bulgarian corpora were processed by BTB-Pipeline (more details in deliverable D5.4). The annotation in each document is an output of the following annotation process:

- Tokenization
- Part of Speech tagging and lemmatization

Corpus	EU
tokens	20,869
terms / linked to WordNet	20,869 / 9,492 (45,48%)
entities / linked to DBpedia	790 / 406 (51,39%)
coreference chains	2,278

Table 4: Statistics on News QTLeap WSD/NED corpus (Basque)

- Named Entity Recognition and Classification
- Named Entity Disambiguation
- Word Sense Disambiguation
- Dependency parsing
- Coreference

The resulting annotations are represented in NAF. The annotated corpora are the following:

- Parallel corpus
  - Europarl-QTLeap WSD/NED corpus: We used part of Bulgarian-English Europarl corpus v7.0, which intersects with Spanish-English corpus, provided in D5.3. We have got over 4M tokens out of total 14M. Table 5.

Corpus	BG
tokens	4,835,287
terms / linked to WordNet	4,835,287 / 1,666,359 (34.46%)
entities / linked to DBpedia	61,195 / 61,195 (100%)
coreference chains	30,922

Table 5: Statistics on annotated Europarl parallel corpus (Bulgarian)

  - QTLeap WSD/NED corpus: batch 1 and 2: The Bulgarian translation of QTLeap batch 1 and 2 questions ans answers. Table 6.

Corpus	BG
tokens	67,591
terms / linked to WordNet	67,591 / 12,627 (18.7%)
entities / linked to DBpedia	180 / 180 (100%)
coreference chains	306

Table 6: Statistics on QTLeap WSD/NED corpus (Bulgarian)

  - News-QTLeap WSD/NED corpus. The Bulgarian translation of the QTLeap news corpus used in Pilot 2. Table 7.
  - SETIMES QTLeap WSD/NED corpus. This is the Bulgarian part of SETIMES corpus cleaned and checked manually within EuroMatrixPlus Project. Within QTLeap project it is annotated with WSD/NED. Table 8.

Corpus	BG
tokens	23,549
terms / linked to WordNet	23,549 / 7,238 (30.74%)
entities / linked to DBpedia	509 / 509 (100%)
coreference chains	106

Table 7: Statistics on News QTLeap WSD/NED corpus (Bulgarian)

Corpus	BG
tokens	582,376
terms / linked to WordNet	582,376 / 142,638 (24.5%)
entities / linked to DBpedia	57,585 / 22,935 (39.8%)
coreference chains	39,637

Table 8: Statistics on SETIMES QTLeap WSD/NED corpus (Bulgarian)

- Comparable corpus
  - Wikipedia-QTLeap WSD/NED corpus: In addition to parallel Bulgarian-English corpora presented above we have annotated comparable corpus based on articles from Bulgarian and English wikipedia. Table 9.

Corpus	BG
tokens	5,533,725
terms / linked to WordNet	5,533,725 / 1,382,014 (24.97%)
entities / linked to DBpedia	155,133 / 155,133 (100%)
coreference chains	95,118

Table 9: Statistics on annotated Wikipedia comparable corpus (Bulgarian)

  - Bulgarian Radio-QTLeap WSD/NED corpus: We used comparable documents provided to QTLeap project by Bulgarian National Radio. Some of the corresponding documents are exact translations of the Bulgarian original texts, but there are some translations of English text into Bulgarian which are not complete. This is why we consider the corpus comparable. Table 10.

5.3 Czech-English

The whole annotation process is run in Treex scenario. All processes are implemented as Treex blocks. Word sense disambiguation was based on the Valency lexicon disambiguation. PoS tagger MorphoDiTa and NERC tool NameTag annotation processes were described in deliverable D5.4. The annotation in each document is an output of the following annotation process:

- Tokenization
- Part of Speech tagger and lemmatization
- Named Entity Recognition and Classification
- Named Entity Disambiguation

Corpus	BG
tokens	656,853
terms / linked to WordNet	656,853 / 169,593 (25.82%)
entities / linked to DBpedia	18,223 / 18,223 (100%)
coreference chains	13,230

Table 10: Statistics on annotated comparable corpora from Bulgarian National Radio (Bulgarian)

- Word Sense Disambiguation
- Dependency parser
- Coreference

These are the annotated corpora:

- Parallel corpus
  - Europarl-QTLeap WSD/NED corpus: We used part of Czech-English Europarl corpus v7.0, which intersects with Spanish-English corpus, provided in D5.3. We have got 9M tokens out of total 14M. Table 11.

Corpus	CZ
tokens	17,839,988
terms / linked to WordNet	17,839,988 / 8,402,381 (47%)
entities / linked to DBpedia	630,360 / 239,163(37.9%)
coreference chains	377,083

Table 11: Statistics on annotated Europarl parallel corpora (Czech)

  - QTLeap WSD/NED corpus: batch 1 and 2. Table 5.3.

Corpus	CZ
tokens	71,061
terms / linked to WordNet/Vallex	71,061 / 11,060(15.5%)
entities / linked to DBpedia	1715 / 572 (33.3%)
coreference chains	1027

Table 12: Statistics on QTLeap WSD/NED corpus (Czech)

  - News-QTLeap WSD/NED corpus. The Czech translation of the QTLeap news corpus used in Pilot 2. Table 13.

5.4 Portuguese-English

We have annotated the freely available Europarl v7.0 parallel corpus (5M tokens, 160K sentences). Prior to being annotated with the NED, WSD and coreference pipelines, the corpus was pre-annotated using LX-Suite (tokenization, PoS, lemmatization and morphological information) and LX-NER (named entity recognition), as well as being constituency and dependency parsed for the coreference task. Summarizing, the Portuguese annotated corpus contain:

Corpus	CZ
tokens	24,553
terms / linked to WordNet	24,553 / 6,488 (26%)
entities / linked to DBpedia	1,470 / 599 (40%)
coreference chains	657

Table 13: Statistics on News QTLeap WSD/NED corpus (Czech)

- Lemmas, part-of-speech tags and morphological information as part of the pre-processing provided by the LX-Suite on the raw corpus.
- Word senses, provided from the output of the WSD pipeline.
- URIs for Portuguese DBpedia links, provided by the output of the NED pipeline.
- Coreference information, provided by the output of the coreference pipeline.

The annotated corpora are the following:

- Parallel corpus
  - Europarl-QTLeap WSD/NED corpus: We annotated a 160 Kline subset of the Portuguese-English Europarl corpus v7.0. The intersection of this subset with the English side of the Spanish-English corpus is 91%. Table 14.

Corpus	PT
tokens	10,028,728
terms / linked to WordNet	4,030,944 / 1,322,136 (32.80%)
entities / linked to DBpedia	345,178 / 223,066 (64.62%)
coreferent pairs	347,620

Table 14: Statistics on Europarl-QTLeap WSD/NED corpus (Portuguese)

  - QTLeap WSD/NED corpus: batch 1 and 2. Table 15.

Corpus	PT
tokens	72,018
terms / linked to WordNet	29,985 / 6,116 (20.40%)
entities / linked to DBpedia	3,799 / 1,868 (49.17%)
coreferent pairs	183

Table 15: Statistics on QTLeap WSD/NED corpus (Portuguese)

  - News-QTLeap WSD/NED corpus. The Portuguese translation of the QTLeap news corpus used in Pilot 2. Table 16.

5.5 Spanish-English

The corpora reported here extends the ones released previously in D5.3. We used the same ixa-pipe tools [Agerri et al., 2014] described in D5.4 on this larger corpora. Thus, the annotation in each document is an output of the following annotation process:

Corpus	PT
tokens	30,117
terms / linked to WordNet	12,251 / 3,847 (31%)
entities / linked to DBpedia	1,184 / 622 (53%)
coreference markables	7,432

Table 16: Statistics on News QTLeap WSD/NED corpus (Portuguese)

- Tokenization
- Part of Speech tagger and lemmatization
- Named Entity Recognition and Classification
- Named Entity Disambiguation
- Word Sense Disambiguation
- Constituent parser
- Coreference

The annotated corpora are the following:

- Parallel corpus
  - Europarl-QTLeap WSD/NED corpus: Comprises the whole Spanish-English Europarl v7.0 corpus. Note that the English side of the Spanish-English Europarl corpus was the one chosen as the main English annotated corpus. Table 17.

Corpus	ES
tokens	57,053,435
terms / linked to WordNet	57,053,435 / 21,766,296 (38.15%)
entities / linked to DBpedia	2,153,689 / 1,763,055 (81.86%)
coreference chains	756,732

Table 17: Statistics on Europarl-QTLeap WSD/NED corpus (Spanish)

  - QTLeap WSD/NED corpus: batch 1 and 2. Table 18.

Corpus	ES
tokens	71,989
terms / linked to WordNet	71,989 / 22,704 (31.54%)
entities / linked to DBpedia	4,313 / 3,175 (73.61%)
coreference chains	705

Table 18: Statistics on QTLeap WSD/NED corpus (Spanish)

  - News-QTLeap WSD/NED corpus. The Spanish translation of the QTLeap news corpus used in Pilot 2. Table 19.

Corpus	ES
tokens	28,781
terms / linked to WordNet	28,781 / 10,000 (34.75%)
entities / linked to DBpedia	1,168 / 915 (78.34%)
coreference chains	393

Table 19: Statistics on News QTLeap WSD/NED corpus (Spanish)

5.6 English side of parallel and comparable corpora

We used the same ixa-pipe tools [Agerri et al., 2014] described in D5.4 on the annotation of these corpora. Thus, the annotation in each document is an output of the following annotation process:

- Tokenization
- Part of Speech tagger and lemmatization
- Named Entity Recognition and Classification
- Named Entity Disambiguation
- Word Sense Disambiguation
- Constituent parser
- Coreference

The annotated corpora are the following:

- Parallel corpus
  - Europarl-QTLeap WSD/NED corpus: The English side of the Europarl-QTLeap WSD/NED corpus contains two corpora. One is the English side of the EN-ES Europarl corpus v7.0, which aligns with the Bulgarian, Czech, Spanish and Portuguese sides of the respective Europarl corpus. Table 20. The second is the English side of the publicly available EN-EU corpus, which is not related to Europarl. Table 21.

Corpus	EN
tokens	54,407,887
terms / linked to WordNet	54,407,887 / 25,168,791 (46.26%)
entities / linked to DBpedia	1,985,109 / 1,559,783 (78.57%)
coreference chains	1,530,519

Table 20: Statistics on Europarl-QTLeap WSD/NED corpus, English side of the Spanish-English Europarl

  - QTLeap WSD/NED corpus: batch 1 and 2. Table 22.
  - News-QTLeap WSD/NED corpus. English side of the QTLeap news corpus used in Pilot 2. Table 23.

Corpus	EN
tokens	5,411,834
terms / linked to WordNet	5,411,834 / 2,051,214 (37.90%)
entities / linked to DBpedia	235,314 / 122,989 (52.27%)
coreference chains	54,857

Table 21: Statistics on the GNOME section of the Europarl-QTLeap WSD/NED corpus, English side of the Basque-English corpus

Corpus	EN
tokens	68,913
terms / linked to WordNet	68,913 / 25,807 (37.45%)
entities / linked to DBpedia	2,999 / 1,950 (65.02%)
coreference chains	1,199

Table 22: Statistics on QTLeap WSD/NED corpus, English side

  - SETIMES QTLeap WSD/NED corpus. This is the English side of SETIMES corpus cleaned and checked manually within EuroMatrixPlus Project. Within QTLeap project it is annotated with WSD/NED. Table 24.

- Comparable corpus
  - Wikipedia-QTLeap WSD/NED corpus: In addition to parallel Bulgarian-English corpora presented above we have annotated comparable corpus based on articles from Bulgarian and English Wikipedia. Table 25.
  - English part of Bulgarian Radio-QTLeap WSD/NED corpus. We used comparable documents provided to QTLeap project by Bulgarian National Radio. Some of the corresponding documents are exact translations of the Bulgarian original texts, but there are some translations of English text into Bulgarian which are not complete. This is why we consider the corpus comparable. Table 26.

Corpus	EN
tokens	25,432
terms / linked to WordNet	25,432 / 10,807 (42.49%)
entities / linked to DBpedia	1,288 / 1,056 (81.99%)
coreference chains	707

Table 23: Statistics on News QTLeap WSD/NED corpus. English side

Corpus	BG
tokens	578,405
terms / linked to WordNet	578,405 / 227,370 (39.04%)
entities / linked to DBpedia	43,077 / 36,379 (84.45%)
coreference chains	26,039

Table 24: Statistics on SETIMES QTLeap WSD/NED corpus. English side of Bulgarian-English SETIMES corpus

Corpus	EN
tokens	29,675,466
terms / linked to WordNet	29,675,466 / 8,568,430 (28.87%)
entities / linked to DBpedia	1,703,238 / 1,703,238 (100%)
coreference chains	178,119

Table 25: Statistics on annotated Wikipedia comparable corpus. English side of Bulgarian-English Wikipedia comparable corpus

Corpus	EN
tokens	1,603,598
terms / linked to WordNet	1,603,598 / 438,302 (27.33%)
entities / linked to DBpedia	63,961 / 63,961 (100%)
coreference chains	66,707

Table 26: Statistics on annotated comparable corpora from Bulgarian National Radio. English side of Bulgarian-English Bulgarian National Radio comparable corpus

6 Evaluation of basic processing tools

In this section, we report on the evaluation of the basic tools and the alignments mentioned in this deliverable. The next sections report the evaluation of the advanced tools. We report the quality of the tools and resources using standard metrics like precision, recall and F1 on publicly available datasets whenever possible (all the datasets used are listed in the Appendix B). In the case of aligned resources, we provide a qualitative statement. In the domain evaluation subsections we report on the quality of the output of the tools when run on the user scenario texts (batches one and two).

6.1 Basque

6.1.1 Aligned resources

The Basque WordNet is aligned to the English WordNet by design [Pociello et al., 2011, Gonzalez-Agirre et al., 2012], so there is no need for further evaluation. In the case of DBpedia for Basque, the alignment is also native. We did not see any issues in any of those mappings.

6.1.2 Lemmatization and PoS tagging

The EPEC corpus (the Reference Corpus for the Processing of Basque) is aimed to be a 'reference' corpus for the development and improvement of several NLP tools for Basque [Aduriz et al., 2006]. It is a 300,000-word sample collection of news published in Euskaldunon Egunkaria, a Basque language newspaper. This corpus has been manually tagged at different levels (morphology, syntax, phrases...). PoS tagging accuracy of ixa-pipe-pos-eu on its test set reaches 95.17%, when considering all morphological information accuracy obtained reaches 91.89%.

6.1.3 NERC

A fraction of the EPEC corpus, consisting in 60.000 tokens, was manually annotated with 4748 named entities. When evaluated over a subset of ca. 15,000 tokens, ixa-pipe-nerc's F1 measure is 76.72% on 3 class evaluation and 75.40 on 4 classes.

6.1.4 Domain evaluation

Lemmatizer For the Basque lemmatizer we have seen no difference in performance due to the change in domain. As we can see on the example in Appendix A.1, the lemmatizer correctly strips the morphological suffixes for all grammatical categories, in particular, nouns and verbs e.g., "dakit" has been lemmatized as "jakin", tuning the conjugated verb form I know into the verb lemma know and the possessive case noun "sarearen" of the net has been lemmatized as "sare" net. We see that the lemmatization of entities is generally correct, e.g. "Wi-fi" has been lemmatized as "Wi-fi" and "iPhone-an" as "iphone", but specialized terminology does show some occasional error, as is the case of Facebook, which was incorrectly lemmatized as "Faceboo". This is due to the final -k being a suffix marker for the ergative case in Basque.

PoS tagger The PoS tagger for Basque maintains its high accuracy levels for the domain of the use scenario. As an example (cf. Appendix A.1), we see how two regular sentences are correctly tagged, including domain-specific terminology such as "sarearen", "pasahitza" or "aplikazioa" which have been tagged as common names. (Notice that "Facebook" has been assigned a correct proper noun PoS tag despite the incorrect lemmatization). We see the occasional mistake in the tagging of iPhone-an, which has been tagged as a common noun, instead of a proper noun.

NERC In Batch 1 and Batch 2 of the HF user scenario corpus (4,002 sentences), the NERC module detected 3,885 entity mentions, which were aggregated into 1,672 unique entities (counts over lemmatized entities). After inspecting the recognized entities, we see that the performance of the tool remains at high accuracy level. We observed that the tool correctly recognizes domain-specific entities (see Table 27). We also noticed that it often recognizes user interface (UI) strings and some internet addresses as entities (although not paths, as happened with English and Spanish). This is the case of "kontrol panel" with 18 instances and "hasiera" with 11. We have found a number of general words, mainly verbs, that have been recognized as entities. These are most often imperative forms that appear at the beginning of sentence, as is the case of "Joan" Go with 23 instances.

Number of occurrences	Entities
171	Windows
119	Wi-Fi
98	Google
78	Skype
45	Gmail
42	internet
39	Facebook
39	Android
38	Dropbox
34	Word

Table 27: 10 most frequent entities for Basque

Most of the entities recognized by the NER tool fall out of the three classification categories. It mostly recognizes IT-related terminology, brand and product names. We believe that none of them can be classified as Person, Location or Organization. Therefore, the classification might not be appropriate for the user domain. For example, USB, Wi-Fi and Internet are all classified as Organization (cf. Table 28). We see that Windows, Google and Skype have instances classified in all three categories, which shows the difficulty the NERC tool has with these entities. It seems necessary to either set a fourth category to gather terminology and products or define which of the three categories will be accepted as valid. Additionally, given the instructive nature of the texts in our use scenario, imperatives are very frequent. We see that the NER tool incorrectly identifies them as entities and the NERC tool then incorrectly classifies them as Organization (Egin) and Person (Joan).

What this analysis shows is that the classification module is not tuned to deal with terminology, product names or highly instructive text, which is a known weakness of NERC tools trained on general corpora. We will have to see whether the disambiguation of entities by the NED tool is badly affected by this or whether the tool still manages

to select the appropriate sense. Should this be the case, we could choose to overlook the NERC classification, and perhaps try to use the NED output to recognize the correct class. Another alternative would be to apply domain adaptation techniques to improve NERC performance on product names.

Number of occurrences	Entity	Class
111	Wi-Fi	ORGANIZATION
80	Windows	PERSON
53	Windows	LOCATION
43	Google	PERSON
38	Windows	ORGANIZATION
32	Facebook	PERSON
31	Skype	PERSON
31	Google	ORGANIZATION
26	Egin	ORGANIZATION
25	Skype	LOCATION
24	ZON	PERSON
24	Google	LOCATION
23	Skype	ORGANIZATION
22	Word	PERSON
21	USB	ORGANIZATION
21	Saioa	PERSON
21	Joan	PERSON
21	joan_ezarpen	ORGANIZATION
21	internet	ORGANIZATION
20	IP	ORGANIZATION

Table 28: 20 most frequent entities with class for Basque

6.2 Bulgarian

6.2.1 Aligned resources

The Bulgarian WordNet is aligned manually to English Wordnet by one person and the alignment is checked manually by a second person. Each new sense is added to the Bulgarian WordNet as a new synset and then the new synset is aligned to the English WordNet. The alignment between Bulgarian DBpedia and English DBpedia is provided within DBpedia itself. The missing entities in DBpedia that were created on the basis of Wikipedia were also checked by two people.

The parallel corpus extracted from SETIMES has been aligned manually on sentence level within the European project EuroMatrixPlus. It is partially aligned on word level.

6.2.2 Lemmatization and PoS tagging

PoS tagging and lemmatization are evaluated on the basis of the annotation within Bulgarian Treebank - BulTreeBank. The best result over data from BulTreeBank is 97.98% [Georgiev et al., 2012]. The evaluation over out-of-the-treebank data (SETIMES corpus) showed around 97% accuracy. Lemmatization achieved 95% accuracy on new data - mainly because of errors in PoS tagger or new words.

6.2.3 NERC

For the evaluation we manually checked the performance on new text (12223 tokens). The gold standard annotation contains 810 named entities. The automatic procedure recognized 688 entities, the intersection annotations with the gold standard were 593. The precision of the tool is 86.1% and the recall is 73.2% (79.1 F1). During the rest of the project we will be improving the tool by adding more names to the gazetteers in use and by creating better rules for multiword names.

6.2.4 Domain evaluation

Lemmatizer In Bulgarian pipeline the lemmatization is rule-based and depends on disambiguation resolver. For example, “beli” (white.PL) as an adjective received the adjectival lemma “byal” (white.SG) and as a noun — a lemma for the noun “belya” (mischief). During the addition of domain specific lexical items we added the corresponding rules. Named Entities in Bulgarian have not received inflected forms. In cases when such inflected forms are possible, they have a rare usage. Thus, the performance of the lemmatizer is similar to the test for general texts.

PoS tagger The evaluation of the PoS tagger and lemmatizer on the user scenario texts shows considerable drop of performance. The accuracy of PoS tagging is 86.56%. The main type of errors is the proper treatment of menu items like Insert, Move, etc., and product names like Google Calendar, because they were not translated into Bulgarian. The other type of errors is related to new frequent words like "KLIKAM" (to click). Such words have predominantly wrong annotation. Other typical errors are related to grammatical features like imperative forms of verbs, differences in tenses and persons. The evaluation of the lemmatizer is more complicated, because in the cases of wrong part of speech even the correct lemma has to be considered as erroneous. The evaluation is done on the basis of 100 sentences (1273 tokens).

NERC For the domain names we created an extension of WordNet for Bulgarian and English. Thus the NE classification is performed during the WSD task. Within the pipeline they are classified just as named entities. We have processed manually Batch 1 of the QTLeap user scenario corpus. In the translation to Bulgarian most of the domain named entities are left as they where in the English text. Thus the recognition is relatively easy. Problematic cases are some multiword elements like “Network Settings”, “System Tools”, etc. Thus, the performance on Batch 1 is from 2493 manually annotated domain NEs. From them 768 terms are multiword expressions. The pipeline identified 2576 Named Entities while 152 were not recognized. From the multiword expressions only 47 were recognized. Thus, the recall is 64.98% and the precision is 62.89.

Table 29 shows the 10 most frequent named entities as returned by the Bulgarian pipeline for Batch 1.

From the table we can see that the elements of multiword expressions are very frequent. For example, “Settings”, “Account” are predominately elements of multiword expressions.

Number of occurrences	Entities
60	Settings
55	Windows
36	Options
35	Facebook
25	Control
25	Start
25	USB
24	File
23	Account
23	Tools

Table 29: 10 most frequent named entities for Bulgarian

6.3 Czech

6.3.1 Aligned resources

The link between the Czech and English Wikipedias is straightforward using the information in DBpedia and Wikipedia. CUNI will also evaluate the coverage of Czech Wikipedia by Babelnet, i.e. the amount of entries that exist in Czech Wikipedia but are missing in Babelnet.

6.3.2 Lemmatization and PoS tagging

Czech has standard resources with manual morphological annotation, i.a. the Prague Dependency Treebank[50]. Its part-of-speech tagset for includes also all morphological categories of Czech and contains several thousands of possible tags. Tagging plus lemmatization accuracy of MorphoDiTa on its test set reaches 95.03% [Straková et al., 2014], which is the state of the art for Czech.

6.3.3 NERC

NameTag is the state-of-the-art NERC tool for Czech. Its F1 measure on the test portion of Czech Named Entity Corpus 2.0[51] is 80.30% for the coarse-grained 7-classes classification and 77.22% for the fine-grained 42-classes classification [Straková et al., 2014].

6.3.4 Domain evaluation

Lemmatizer As the majority of words in the HF user scenario corpus come from a general domain, a difference in performance due to the change in domain is marginal. The lemmatizer works well for common dictionary words, e.g. "mohu" and "nabídce" have been lemmatized as "moci" and "nabídka", respectively. Whereas we spot no errors in lemmatization of terminology expressed by a common name, problems occur with some proper names not included in the dictionary, for which the lemma is guessed based on its affixes and context, e.g. "LibreOffice" turns into "LibreOffika". On the other hand, for names with a Czech morphological suffix guesser produces correct lemmas, e.g. "Notepadu" has been lemmatized to "notepad". The most obvious issue is varying

[50] http://ufal.mff.cuni.cz/pdt3.0
[51] http://ufal.mff.cuni.cz/cnec

tokenization of URLs and their subsequent lemmatization, e.g. "drive.google.com" is assigned the lemma "drive.google.co" (see an example on Appendix A.3).

PoS tagger On HF data, the Czech tagger shows good performance on both general and domain-specific words, especially if they are inflected for number and/or case. On the other hand, domain-specific words that do not inflect are often misanalyzed in terms of morphological features, as these are not marked on the words; still, we believe that since these words typically do not inflect in any of the focus languages, incorrect assignment of morphological categories is not a grave issue. See the example in Appendix A.3, where the inflected word "Photoshopu" is correctly analyzed for singular number (S) and locative case (6), and even the uninflected word "jpeg" is correctly analyzed for singular number (S) and accusative case (4), probably thanks to the preceding conjunction which requires accusative case; on the other hand, the number and case for "png" is has not been identified by the tagger (X), even though the preceding preposition is known to require genitive case (2).

NERC The Czech named entity recognizer identified only 819 mentions of 389 entities in the 2000 sentences of HF user scenario corpus batch 1.

Both comparisons of these numbers with other languages and manual inspection of the results show that the recall of the recognizer is unpleasantly low. This is undoubtedly due to the fact that the training corpus contains close to no occurrences of many of the domain-specific named entities that occur in the HF corpus, and was not created with this specific domain in mind. For example, on the HF corpus, NameTag tagged the word "Skype" 15 times as a named entity, although it occurs 82 times in the dataset; analysis of the NameTag training corpus revealed that "Skype" occurs only 4 times in it, and is never tagged as a named entity. Similarly, whereas the word "Windows" is among the most frequent entities in the other languages, it does not even reach the top 20 in Czech. Out of 98 occurrences of "Windows" in the dataset, NameTag tagged only 4 of them as a named entity and 16 occurrences as a part of a multiword entity, e.g. "Windows 7"; again, its frequency in the training corpus is very low, only 6 occurrences.

Table 30 shows the 10 most frequent named entities as returned by NameTag. While the absolute numbers are low, the precision of the named entity recognizer is rather good – in the top 20 named entities, there is only one non-entity word ("Mohu", which means "Can I"); this has been confirmed by a manual inspection of the whole set of found named entities, which showed a very small number of false positives.

The table also shows that NameTag is quite successful at detecting multiword entities, such as "MEO Cloud" or "Samsung TV", although this is not always true - e.g. "Zon HUB" was marked as two separate entities more often than as one multiword entity.

As for the class identification, NameTag performance is quite reasonable; it labels most named entities correctly, although mislabelings are frequent. Moreover, as already noted for other languages, there is a strong inherent ambiguity between "company" and "product" class for many of the named entities, such as "Google" or "YouTube". NameTag usually prefers the former, while the latter is usually much more reasonable in the domain.

The 20 most frequent entity-class pairs found are shown in Table 31. As mentioned in Section 3.3.2, NameTag for Czech works with 42 fine-grained classes merged into 7 super-classes. For convenience, the table also contains a mapping of these classes to the 4 standard classes used for other languages. We found domain-specific named entities are rare in the training corpus. Moreover, the hierarchy of named entities defined by the

Number of occurrences	Entities
27	2014
16	HUB
15	Skype
15	Google+
14	LibreOffice
12	MEO Cloud
12	Google
11	Samsung TV
10	Zon
10	Apple ID

Table 30: 10 most frequent entities for Czech

corpus, although quite detailed, is not well suited for our domain – in most cases, the best category found is "company" or "product", although the hierarchy defines other 40 named entity classes, which probably confuses the recognizer.

6.4 English

6.4.1 Aligned resources

BabelNet combines WordNet and Wikipedia by automatically acquiring a mapping between WordNet senses and Wikipedia pages, avoiding duplicate concepts and allowing their inventories of concepts to complement each other. The mapping algorithm [Navigli and Ponzetto, 2012] leverages resource-specific properties (monosemous senses and redirections) and, given a Wikipedia article, finds the WordNet sense that fits best the article. The accuracy reported by the authors is 82.7, as measured on a random sample of 1000 Wikipedia articles.

Note that in this project we also align between Wikipedia versions, and between Wikipedia and DBpedia. The mapping between Wikipedia versions is possible thanks to the fact that the Wikipedia team maintains redirects from older articles to new articles. The mapping between Wikipedia and DBpedia is straightforward: it suffices to ensure that the Wikipedia and DBpedia versions match (i.e. each DBpedia version is linked to a specific Wikipedia dump) and then match the names of the articles, as the automatic construction of DBpedia ensures a one-to-one mapping.

Although the quality of the mappings between Wikipedia versions has not been reported anywhere, in our experience as a top ranking team in Entity Linking competitions [Barrena et al., 2013], we have seen that in some cases the mapping is not 100% accurate and complete, but even if we have not quantified this exactly, the information loss is marginal. The Wikipedia to DBpedia mapping is 100% accurate.

6.4.2 Lemmatization and PoS tagging

The ixa-pipe-pos module for lemmatization and PoS tagging obtained the best results so far with Perceptron models and the same featureset as in Collins [2002]. The models have been trained and evaluated on the WSJ treebank using the usual partitions (e.g., as explained in Toutanova et al. [2003]. We currently obtain a performance of 96.88% vs 97.24% in word accuracy obtained by Toutanova et al. [2003].

Number of occurrences	Entity	Class	NameTag class
22	2014	MISC	number - sport score
16	HUB	PERSON	person - surname
13	Google+	MISC	artifact - product
13	LibreOffice	PERSON	person - surname
11	Samsung TV	ORGANIZATION	media - TV station
10	Zon	PERSON	person - first name
10	Google	ORGANIZATION	institution - company
9	Cloud	LOCATION	geography - castle/ chateau
9	7	MISC	number - sport score
8	McAfee	ORGANIZATION	institution - company
8	Mohu	PERSON	person - surname
8	Skype	MISC	artifact - product
8	Apple	ORGANIZATION	institution - company
7	Bitdefender	ORGANIZATION	institution - conference/ contest
7	Norton	PERSON	person - surname
7	Apple ID	ORGANIZATION	institution - company
7	YouTube	ORGANIZATION	institution - company
7	Google Drive	ORGANIZATION	institution - company
7	GB	MISC	artifact - measure unit
6	MEO Cloud	ORGANIZATION	institution - company

Table 31: 20 most frequent entities with class for Czech

MorphoDiTa reaches accuracy 97.27% on the same dataset [Straková et al., 2014], which is near state of the art.

6.4.3 NERC

The ixa-pipe-nerc module based on the CONLL 2002[52] and 2003[53], trained on local features only obtains F1 84.53, and the models with external knowledge F1 87.11. The Ontonotes CoNLL 4 NE types with local features model obtains F1 86.21. The Ontonotes 3 NE types with local features configuration obtains F1 89.41.

6.4.4 Domain evaluation

Lemmatizer As we mentioned for Basque, the lemmatizer for English performs almost perfectly. We have seen no difference in performance due to the change in domain. As we can see on the example in Appendix A.4, the lemmatizer performs well for the main linguistic changes that occur in English, namely, verbs e.g. “disappeared” has been lemmatized as “disappear”, and number e.g. “speakers” has been lemmatized as “speaker”. Also, we see no errors regarding the lemmatization of terminology and entities, e.g. “Gmail” has been lemmatized as “Gmail” and specialized terms such as “desktop” or “icon” have also been properly lemmatized as “desktop” and “icon”.

[52] http://www.clips.ua.ac.be/conll2002/ner/
[53] http://www.clips.ua.ac.be/conll2003/ner/

PoS tagger As already noted for Basque, the PoS tagger for English maintains its high accuracy levels for the domain of the use scenario. As an example (cf. Appendix A.4), we see how a regular sentence is correctly tagged, including the domain-specific product name such as Gmail, which has been properly tagged as a proper singular noun.

NERC In Batch 1 and Batch 2 of the HF user scenario corpus (4,002 sentences), the NERC module detected 1,893 entity mentions, which were aggregated into 749 unique entities. After inspecting the recognized entities, we see that the performance of the tool remains at high accuracy levels. We observed that the tool correctly recognizes domain-specific entities (see Table 32 below). We also noticed that it often recognizes user interface (UI) paths as entities. This is the case of "Menu > Settings" or "Menu Screen > Network > Network Connections", for example.

Number of occurrences	Entities
90	Windows
84	Facebook
65	Google
54	PC
31	USB
30	Google Chrome
29	Google Drive
24	Internet
21	Skype
14	YouTube

Table 32: 10 most frequent entities for English

Although the classification of general entities (not domain-specific) is most often correct, we see some degradation with domain-specific terminology (see Table 33). This is particularly true with product and brand names. We see that Facebook, Google or Panda are classified as Organizations. This is true if we consider the cases where these names refer to the company. However, in our user scenario, the names usually refer to product names. Similarly, applications such as Google Chrome or Google Drive, also get the Organization class. Other more serious misclassifications include product names such as Skype or WhatsApp as Location. What this shows is that the classification module is not tuned to deal with product names, which is a known weakness of NERC tools trained on CoNLL corpora.

We noted that the disambiguation of entities (see Section on NED below) is correct even when the classification is not. We can also choose to overlook the NERC classification, and perhaps try to use the NED output to recognize the correct class. Another alternative would be to apply domain adaptation techniques to improve NERC performance on product names.

6.5 Portuguese

6.5.1 Aligned resources

The Portuguese WordNet is aligned to the English WordNet by design as the synsets were manually constructed and aligned with the English equivalents. Accordingly, the evaluation is not an issue here.

Number of occurrences	Entity	Class
90	Windows	MISC
84	Facebook	ORGANIZATION
65	Google	ORGANIZATION
54	PC	ORGANIZATION
31	USB	ORGANIZATION
30	Google Chrome	ORGANIZATION
29	Google Drive	ORGANIZATION
24	Internet	MISC
21	Skype	LOCATION
14	YouTube	ORGANIZATION
14	Portuguese	MISC
13	Panda	ORGANIZATION
13	OK	LOCATION
13	MEO	ORGANIZATION
12	Panda	LOCATION
12	Microsoft	ORGANIZATION
12	Google Play	ORGANIZATION
12	Apple ID	ORGANIZATION
11	WhatsApp	LOCATION

Table 33: 20 most frequent entities with class for English

6.5.2 Lemmatization and PoS tagging

Under a 10-fold cross validation over a reference corpus of ca. 150 Ktokens, the PoS tagger scored an accuracy of 96.87% [Branco and Silva, 2004].

As for the morphological analysis extracting the lemma and inflection features, given the inflection system of Portuguese, with a highly rich morphology for verbs, the task is assigned to different tools, one for nominal and the other for verbal inflection.

With regards nominal analysis, the tool that extracts lemmas has 97.67% f-score [Branco and Silva, 2007], and the tool that extracts inflectional feature values has 91.07% f-score [Branco and Silva, 2006b].

In what concerns verbal analysis, a single tool takes care of both processes, of lemmatization and featurization, and it disambiguates among the various lemma-inflection pairs that can be assigned to a verb form with 95.96% accuracy [Branco et al., 2006].

6.5.3 NERC

The rule-based component of the NERC was evaluated against a manually constructed test-suite including over 300 examples. It scored 85.19% precision and 85.91% recall (85.54 F1). When trained over a manually annotated corpus of approximately 208,000 words and evaluated against an unseen portion with approximately 52,000 words, the other data-based module scored 86.53% precision and 84.94% recall (85.73 F1) [Ferreira et al., 2007].

6.5.4 Domain evaluation

The domain evaluation was performed over a set of 3,000 sentences (ca. 37,300 tokens) from the HF user scenario corpus.

Lemmatizer The lemmatizer works by applying suffix replacement rules. Running it on the HF user scenario domain has little impact on its overall performance. The errors that were found fall into two main categories: (i) word with the wrong POS tag, and (ii) English words.

A word with the wrong POS tag will lead the lemmatizer to apply a different set of suffix replacement rules (e.g. rules for nouns instead of rules for verbs, or vice-versa). For instance, "wifi" is sometimes incorrectly tagged as a verb (this is due to the POS tagger not knowing the word and triggering the suffix-based heuristics for guessing the POS tag). Taking "wifi" as a verb, the lemmatizer applies the suffix replacement rules for verbs and assigns the lemma "wifer".

When the word is in English, and even if the POS tag is correct, the suffix rules of the lemmatizer may be triggered by the suffix of the English word, and produce the wrong lemma. For instance, "backup" is correctly tagged as a common noun and since its suffix does not trigger any replacement rule, the lemma is "backup". The word "addons" is also correctly tagged as a common noun, but since its suffix happens to trigger a replacement rule, the lemma becomes "addom", which is wrong.

An example is shown in Appendix A.5. Note that the lemmatizer does not assign lemmas to words from the closed classes, since these are retrievable through a dictionary lookup. It also does not lemmatize proper names. In the first sentence, "emails" is not properly lemmatized since its suffix does not trigger any rule. In the second sentence, "wifi" is tagged as a verb and lemmatized as "wifer".

POS tagger Overall, the POS tagger shows good performance. However, having been trained over newspaper texts, its accuracy suffers due to the change in domain and style. This is particularly noticeable in the following cases: (i) English words, (ii) words with the wrong capitalization, and (iii) the first word in a sentence.

Much of the domain-specific terminology consists of English words, which are often unknown to the tagger. The unknown word heuristics used by the tagger tend to assign common noun to these words, which is almost always the correct choice. For instance, "password" occurs 39 times, 35 of which are tagged as common noun, 2 as an adjective and 2 as proper name; "email" occurs 56 times, 42 as a common noun and 14 as an adjective; "router" occurs 56 times, 53 as a common noun and 3 as a verb. There are, however, cases like "wifi", which occurs 7 times, 4 as a verb and 3 as a common noun.

Portuguese orthographic conventions indicate that proper names should begin with a capital letter, and the capitalization of the word is a feature used by the tagger. Beginning a word with a capital letter tends to strongly bias the POS towards proper name. Conversely, a word that does not begin with a capital letter is unlikely to be a proper name. The scenario corpus has many cases where the user has not properly capitalized proper names. In these cases the tagger tends to assign common noun instead of proper name. For instance, "Google" occurs 74 times, all correctly tagged as proper name, while "google" occurs 87 times (82 as a common noun and 5 as an adjective). This suggests that it might be ultimately advantageous to include a pre-processing step of orthography normalization, whereby certain pre-defined strings (e.g. "google", "skype", "windows") are

forced to be capitalized.

There are several cases where the first word in the sentence is tagged as a proper name when it should be a verb. Part of the reason is that the capitalization of the first word in the sentence biases the tagger towards proper name. This is further compounded by the fact that the training corpus has few sentences that start with a verb. For instance, there are 141 cases where the first token in the sentence is tagged as a proper name, only 9 of which are correct. Nearly half (69) should have been tagged as a verb. The remaining cases should have been tagger as common noun.

A similar issue occurs with some interrogative pronouns, such as "Como" and "Onde" (Eng: "How" and "Where"), which are frequent in the domain corpus but very rare in the corpus used for training the tagger. As such, their are often tagged with the wrong POS (note that the words "como" and "onde" are ambiguous and occur in the training corpus bearing POS tags other than interrogative pronoun).

An effort of domain adaptation should prove valuable in mitigating these issues. This adaptation could consist of adding to the training data of the tagger a few questions that begin with an interrogative pronoun and a few sentences that begin with a verb.

An example is shown in Appendix A.5. The first word in the example, "Ativar" should have been tagged as a verb. The entity "windows xp" is not capitalized and its tokens were not annotated as a proper name.

NERC The NER detects 2,257 entity mentions, which are aggregated into 833 unique mentions. The tool relies on an underlying statistical model trained over newspaper text. Its performance drops with the domain change, though often the problem is not so much in recognizing the existence of the named entity but in classifying it correctly. For instance, Facebook, Skype, Gmail and Outlook are almost always classified as a location instead of organization or miscellaneous. NERC errors tend to fall into two cases: (i) proper names that have not been annotated as such, and (ii) wrong classification.

When a proper name is not tagged as such, usually due to wrong capitalization, the NERC might not recognize it as being a named entity. For instance, "Windows" occurs 109 times, 107 of which as a proper name that is part of an entity, while "windows" (not capitalized) occurs 103 times, never as a proper name and never as part of an entity. As mentioned in the previous Section, a pre-processing step that forces the capitalization of certain strings could mitigate this issue.

If a domain-specific entity is properly tagged as a proper name, it is recognized (see Table 34 with the 10 most frequent entities). Note that the NER was able to include the year/version as part of the entity (e.g. "Word 2013"). This is probably due to the training corpus also having entities with a similar sequence of tokens, such as "Expo 98" (the Lisbon Word Exposition).

Although entities are successfully recognized, their classification is often wrong, with the entities being marked as either a location or a person, when most of the mentions in the domain corpus refer to a product (see below Table 35 with the 20 most frequent entities, with class).

Note that most of these entities are not known to the NERC model, since the newspaper articles that form the training corpus predate Facebook, Skype, YouTube, Gmail, etc. As with the POS tagger, domain adaptation techniques could be applied to incorporate these entities with the correct classification into the model.

Number of occurrences	Entities
98	Facebook
73	Word 2013
66	PowerPoint 2013
59	Windows
39	Skype
38	Mac
35	Excel 2013
29	PC
29	Android
28	Chrome

Table 34: 10 most frequent entities for Portuguese

6.6 Spanish

6.6.1 Aligned resources

The Spanish WordNet is aligned to the English WordNet by design [Gonzalez-Agirre et al., 2012], so there is no need for further evaluation. In the case of DBpedia for Spanish, the alignment is also native. We did not see any issues in any of those mappings.

6.6.2 Lemmatization and PoS tagging

ixa-pipe-pos module for lemmatization and PoS tagging for Spanish obtained the best results so far with Maximum Entropy models and the same featureset as in Collins [2002]. The models have been trained and evaluated for Spanish using the Ancora corpus; it was randomly divided in 90% for training and 10% for testing. This corresponds to 440K words used for training and 70K words for testing. We obtain a performance of 98.88% (the corpus partitions are available for reproducibility). Giménez and Màrquez [2004] report 98.86%, although they train and test on a different subset of the Ancora corpus.

6.6.3 NERC

ixa-pipe-nerc module for Spanish currently obtains the best results training Maximum Entropy models on the CoNLL 2002 dataset. Our best model obtains 80.16 F1 vs 81.39 F1 of [Carreras et al., 2002], the best result so far on this dataset. Their result uses external knowledge and without it, their system obtains 79.28 F1.

6.6.4 Domain evaluation

Lemmatizer For the Spanish lemmatizer, as for Basque and English, we have seen no difference in performance due to the change in domain. As we can see on the example in Appendix A.6, the lemmatizer performs as expected for the main linguistic changes that occur in Spanish, namely, verbs e.g. “puedo” has been lemmatized as “poder”, number and gender e.g. “los” has been lemmatized as “el”. We see an occasional error such as the verb “quiero” that was not properly lemmatized into its infinitive. Also, we see that the lemmatization of entities is generally correct, e.g. “Windows” has been lemmatized as “Windows”. Specialized terminology does show some occasional error such as the case of the plural noun “emails” which has not been properly lemmatized.

Number of occurrences	Entity	Class
94	Facebook	LOCATION
73	Word 2013	PERSON
66	PowerPoint 2013	PERSON
53	Windows	LOCATION
36	Mac	LOCATION
35	Excel 2013	PERSON
34	Skype	LOCATION
27	Chrome	LOCATION
25	Android	LOCATION
24	Google Docs	PERSON
22	Publisher 2010	PERSON
21	Gmail	LOCATION
18	Dropbox	LOCATION
17	Publisher	PERSON
17	PC	ORGANIZATION
17	2013	LOCATION
16	YouTube	LOCATION
16	Outlook 2010	PERSON
15	ID Apple	PERSON
14	Twitter	ORGANIZATION

Table 35: 20 most frequent entities with class for Portuguese

PoS tagger Just as already noted for some other languages, the PoS tagger for Spanish maintains its high accuracy levels for the domain of the use scenario. As an example (cf. Appendix A.6), we see how a regular sentence is correctly tagged, including domain-specific terminology such as "emails" or "programas", which have been tagged common plural nouns. (Notice that "emails" has been assigned a correct plural PoS tag despite the incorrect lemmatization.) Similarly, domain-specific product names such as Windows seem to be tagged properly as proper single nouns. Once again, we see the occasional PoS error in instances such as "quiero" which has been tagged as a coordinating conjunction, instead of a present tense third person singular verb.

NERC In Batch 1 and Batch 2 of the HF user scenario corpus (4,002 sentences), the NERC module detected 5,204 entity mentions, which were aggregated into 1925 unique entities. After inspecting the recognized entities, we see that the performance of the tool remains at high accuracy levels. We observed that the tool correctly recognizes domain-specific entities (see Table 36 below). We also noticed that it often recognizes user interface strings and paths as well as internet addresses as entities. This is the case of "Inicio" in the Table below, for instance, which has been identified in 46 occasions.

It is worth mentioning the difference in the number of recognized mentions in English and Spanish, 1,893 and 5,204, respectively (2.82% and 7.43% of the total tokens). After reviewing the tool's output, we see that the English tool is capturing fewer mentions per entity. For example, the English NER is capturing 6 mentions for Android and 31 for Skype, whereas the Spanish NER captures 50 and 92 respectively. Also, we have noticed that the Spanish NER captures as entities elements such as UI strings and paths, and URLs much more often that the English NER. In general, we can say that the English

Number of occurrences	Entities
81	Facebook
68	Internet
63	Ajustes
56	Skype
48	USB
48	IP
48	Android
46	Inicio
43	PC
43	Google

Table 36: 10 most frequent entities for Spanish

NER tool has a higher precision and lower recall than the Spanish NER tool.

Although the classification of general entities (not domain-specific) is most often correct, we see degradation with domain-specific terminology (see Table 37). This is particularly true with product and brand names. We see that Facebook, Google and Gmail are classified as Person. We also see that some entities such as Windows or Skype are classified as either Person or Location, which shows the difficulty the NERC tool has with these entities. Given the instructive nature of the texts in our use scenario, imperatives are very frequent. We see that the NER tool incorrectly identifies them as entities and the NERC tool then incorrectly classifies them as Person. What this shows is that the classification module is not tuned to deal with product names or highly instructive text, which is a known weakness of NERC tools trained on CoNLL corpora.

We noted that the disambiguation of entities (see Section on NED below) is correct even when the classification is not. We can also choose to overlook the NERC classification, and perhaps try to use the NED output to recognize the correct class. Another alternative would be to apply domain adaptation techniques to improve NERC performance on product names.

Number of occurrences	Entity	Class
81	Facebook	PERSON
68	Internet	MISC
63	Ajustes	PERSON
56	Skype	PERSON
48	USB	ORGANIZATION
48	IP	ORGANIZATION
48	Android	PERSON
46	Inicio	PERSON
43	PC	ORGANIZATION
43	Google	PERSON
40	Puedo	PERSON
36	Skype	LOCATION
36	Herramientas	MISC
35	Gmail	PERSON
33	Puede	PERSON
33	Haz	PERSON
30	ZON	ORGANIZATION
30	Windows	LOCATION
29	Vaya	PERSON
27	Windows	PERSON

Table 37: 20 most frequent entities with class for Spanish

7 Evaluation of WSD

In this section we introduce the evaluation datasets used for each language. We also describe the results of publicly available tools for WSD in each language, both at the start of the project and at the end of the 2nd Year. We use F1 of precision and recall as the main evaluation measure, but also report precision and recall. We report results for each language in the following sections, with a summary in Table 39, Section 7.7.

7.1 Basque

The ixa-pipe-wsd-ukb module for Basque has been evaluated on the publicly available EPEC-EuSemcor dataset[54]. This dataset is a Basque SemCor corpus, that is, a Basque sense-tagged corpus, which comprises a set of occurrences in the Basque EPEC corpus [Aduriz et al., 2006], which has been annotated with Basque WordNet v1.6 senses [Pociello et al., 2011]. More specifically, it contains 42,615 occurrences of nouns manually annotated, corresponding to the 407 most frequent Basque nouns.

The dataset was split at random by documents in train and test, with 70% of the documents in the training dataset (874 documents) and 30% in the testing dataset (375 documents). We preferred to split the dataset according to documents, rather than according to instances, as it reflects better the performance in real situations. Running ixa-pipe-wsd-ukb on this test corpus yield a precision of 57.2, recall of 57.1 and F1 of 57.2[55]. This was the state-of-the-art of Basque WSD tools at the start of the project. Using the sense distribution in the train part, a most frequent sense baseline would obtain a precision of 75.6, recall of 65.9 and an F1 of 70.4. The recall is lower than precision due to the fact that some words do not occur in the training dataset. When we apply our ixa-pipe-wsd-ukb using the distribution of senses in the training dataset we obtain the best results to date for a publicly available tool in this dataset: precision of 73.5, recall of 73.3 and F1 of 73.4. This is the result presented in Table 39.

Doc. number	Domain
122	Economy
90	Europe
491	Sports
105	World
255	Politics
186	Balanced

Table 38: Basque WSD evaluation dataset: Break out of the number of documents according to domain.

In addition, we split the dataset according to domains. Table 38 shows the distribution of documents according to the domain. The first 5 rows correspond to sections of a Newspaper. The final row corresponds to documents drawn from a balanced corpus. We performed experiments using Sports as the test data, with the rest of the documents being used for training. This is a typical scenario when moving to new domains which are not well represented in the training dataset. In this scenario, the most frequent sense baseline would obtain a precision of 50.7, recall of 43.1 and an F1 of 46.6. The drop in recall is

[54] http://ixa2.si.ehu.es/mcr/EuSemcor.v1.0/EuSemcor_v1.0.tgz

[55] The figure is slightly higher than than reported in D5.7, due to a slight difference in the test split.

due to the fact that some words do not occur in train. Note that the performance for the most frequent sense baseline is sensibly lower than for the random split, due to the different distribution of senses in the test domain. The results for ixa-pipe-wsd-ukb using uniform sense distributions is in this case better, with a precision of 53.4, recall of 53.1 and F1 of 53.2. Finally, when we apply our ixa-pipe-wsd-ukb using the distribution of senses in train we obtain the best results: precision of 78.3, recall of 78.0 and F1 of 78.1.

7.1.1 Domain evaluation

Word disambiguation was performed for 24,691 tokens out of a total of 53,239 present in the Batch 1 and Batch 2 of the QTLeap corpus. This means that 46.38% of the tokens were linked to WordNet and were thus disambiguated. Many disambiguations were correct, and we do not see any performance loss from the expected values. Such is the case of the domain-specific noun *menu*, for instance, which was linked to the synset `30-06493392-n` with a confidence of 0.30, specifying "computer menu". A number of incorrect cases were found, such as domain-specific *mouse*, for instance, which was linked to the `30-02330245-n` with a confidence of 0.52, referring to the animal, instead of the correct synset `30-03793489-n`, with confidence 0.48, which is the specific synset for the IT domain.

7.2 Bulgarian

We have used ixa-pipe-wsd-ukb module for Bulgarian with some extensions of the knowledge graph on the basis of syntactic information from the BulTreeBank treebank. The texts in the treebank were divided in 3/4 for extraction of new relations and 1/4 (about 60000 running words) for testing. The best result by this module is 68.23%. We have the first results from training of supervised module for coarser-grained senses with result about 85% to 92% accuracy. We would like to perform experiments in which the senses for the knowledge-based approach are filtered by coarser-grained ones. This filtering is local in nature and thus could require some changes in the software. Another option without changes in the software is the filtering to be applied after UKB tool. Some of the results are published here — Simov et al. [2015].

7.2.1 Domain evaluation

Word sense disambiguation was performed on Batch 1 of the QTLeap corpus. Then the annotation was manually checked, corrected and extended. The performance on non-domain words in BTB WordNet Was 63.13% which we consider as a good, having in mind that many of the domain words did not received any suggestions from the WordNet. After the manual correction and extension the new domain senses where added to BTB WordNet.

7.3 Czech

We report WSD evaluation results for verbs only, due to the lack of publicly available datasets with Wordnet 3.0 senses assigned. In D5.7 F1-score of 80.47% was reported for WSD on verbs senses using Czech Vallency Lexicon [Dusek et al., 2015] evaluated on Prague Czech-English Dependency Treebank. For D5.9 we have used in-house software

due to the lack of training/evaluation data and performed manual evaluation, which is shown in the next block.

7.3.1 Domain evaluation

Word sense disambiguation was performed for 11,060 tokens out of a total of 71,061 present in the Batch 1 and Batch 2 of the QTLeap corpus. This means that 15.5% of the tokens were linked to Valency Lexicon [Urešová, 2011] and were thus disambiguated.

The second approach to WSD, described in Section 5.4.2 of D5.6, was applied to Europarl parallel corpus. Word senses were assigned to 4,474,614 terms out of 9,094,542 (49.2%). The performance seems reasonable, for example, it produced mappings for words *zasedání* and *rozprava* to synsets `30-07145508-n` and `30-07140978-n`, respectively.

7.4 English

The WSD module ixa-pipe-wsd-ukb has been evaluated on the general domain coarse grained all-words datasets (S07CG) [Navigli et al., 2007]. This dataset uses coarse-grained senses which group WordNet 2.1 senses. We run the WSD system using WordNet 2.1 relations and senses. We used the mapping from WordNet 2.1 senses made available by the authors of the dataset. In order to return coarse-grained senses, we run our algorithm on fine-grained senses, and aggregate the scores for all senses that map to the same coarse-grained sense. We finally choose the coarse-grained sense with the highest score. In D5.7 we reported the results of ixa-pipes-wsd-ukb using a uniform distribution of senses, resulting in a precision of 80.2, a recall of 80.1 and a F1 score of 80.1, as reported in Agirre et al. [2014]. In later work we have tried the use of sense distributions estimated from SemCor, a freely available annotated corpus for English [Miller et al., 1993], with slightly better results (precision 81.4, recall 81.3, F1 81.4). This is the result shown in Table 39. Note that the results of the most frequent sense heuristic learned from SemCor are a precision, recall and F1 of 78.9.

Our results are slightly lower than those of the IMS tool [Zhong and Ng, 2010], which was available prior to the project. In Section 10 we will present the developments on the third year which improve current results.

7.4.1 Domain evaluation

Word disambiguation was performed for 25,069 tokens out of a total of 67,081 present in the Batch 1 and Batch 2 of the HF use scenario corpus. This means that 37.37% of the tokens were linked to WordNet and were thus disambiguated. Many disambiguations were correct, and we don't see any performance loss. Such is the case of the noun *account*, for instance, which was linked to the synset 30-13929037 with a confidence of 0.132461, meaning "a formal contractual relationship established to provide for regular banking or brokerage or business services". A number of incorrect cases were found, such as domain-specific ID, for instance, which was linked to the synset 30-09081213-n with a confidence of 0.389109, referring to Idaho, "a state in the Rocky Mountains".

7.5 Portuguese

Prior to this project, there was no publicly available tool for Portuguese WSD. There have been very few research papers on unsupervised WSD for Portuguese performed

over an existing knowledge-base – which would be comparable with our own approach using the WSD-PT tool. One of the few examples that does exist [Nóbrega and Pardo, 2014] is adapted to cater for WSD across documents, making comparison with our own system difficult, but they do use the 'Mihalcea method' (a similar approach to UKB) as a comparison with their work, reporting a precision of 39.71% and recall of 39.47% with this method. While our reported results using WSD-PT so far are higher than this, comparison is not really possible considering that we use a different (if similar) algorithm for the disambiguation in UKB, we run our WSD over Portuguese-specific lexical resources (the Portuguese MultiWordNet) instead of translating open-class terms into English, and we use a different corpus as a baseline for evaluation than the CSTNews corpus [Cardoso et al., 2011] used by the authors.

The WSD-PT tool was evaluated using a gold-standard, sense-annotated version of the CINTIL International Corpus of Portuguese [Barreto et al., 2006], consisting of 23,825 sentences containing open-class words annotated with synset identifiers from the Portuguese MultiWordNet (45,502 annotated from a total of 193,443 open-class words, or 23.52%).

Comparing the output of the tool against the gold-standard data, 45,386 of these 45,502 manually disambiguated words are also automatically disambiguated by WSD-PT, from a total of 59,190 tagged by the algorithm. WSD-PT assigned the same sense to the word as was chosen by the annotator for 29,540 of the 45,386 words for which a sense was assigned both manually and automatically, giving a precision of 65.09%, recall of 64.92% and F1 of 65.00%.

7.5.1 Domain evaluation

Processing Batch 1 and 2 of the QTLeap using WSD-PT, 6,115 (20.40%) terms were disambiguated from a total of 29,895 open-class words. The low recall seen here is likely to be a result of the lack of domain-specific terms in the Portuguese MultiWordNet, over which WSD-PT performs WSD. Many of the domain-specific terms that were evaluated appear to be correct, suggesting that the tool is performing well, as expected. For example, the Portuguese *rede* (in English, *network*) is linked to `30-008434259-n`, a synset containing *network* and *web* in the sense of "an interconnected system of things or people", while *ligação* (in English, *connection*) is linked to `30-000145218-n`, a synset containing *joining* and *connection* in the sense of "the act of bringing two things into contact (especially for communication)". However, we also found some domain-specific terms to have been disambiguated incorrectly – for example, the Portuguese *instalação* (in English, *installation*) is linked to `30-003315023-n`, a synset containing *facility* and *installation* in the sense of "a building or place that provides a particular service or is used for a particular industry".

7.6 Spanish

The Spanish WSD module was evaluated on SemEval-2007 Task 09 dataset [Màrquez et al., 2007]. The dataset contains examples of the 150 most frequent nouns in the CESS-ECE corpus, manually annotated with Spanish WordNet synsets. We ran the experiment over the test part of the dataset (792 instances).

In D5.7 we reported the results of ixa-pipes-wsd-ukb using a uniform distribution of senses, resulting in a precision, and a F1 score of 79.3, as reported in Agirre et al. [2014]. This was the state-of-the-art of Spanish WSD tools at the start of the project, as no other available tool existed. In later work we have tried the use of sense distributions estimated

from the training data, with better results (precision, recall, and F1 of 82.1). This is the result shown in Table 39. P61

7.6.1 Domain evaluation

Word disambiguation was performed for 21,210 tokens out of a total of 70,037 present in the Batch 1 and Batch 2 of the HF use scenario corpus. This means that 30.28% of the tokens were linked to WordNet and were thus disambiguated. Many disambiguations were correct, and we don't see any performance loss. Such is the case of the domain-specific noun red, for instance, which was linked to the synset 30-03820728 with a confidence of 0.253795, pointing to the domain of "computer science". A number of incorrect cases were found, such as domain-specific "banda", for instance, which was linked to the synset 30-04339291 with a confidence of 0.219025, referring to an "artifact consisting of a narrow flat piece of material ", instead of the correct synset 30-06260628, which is the specific synset for the domain of telecommunications "a band of adjacent radio frequencies (e.g., assigned for transmitting radio or television signals)".

7.7 Results

Table 39 summarizes the results at the start of the project and at the end of the 2nd year for the languages in the project, alongside the publicly available tools used in each case. In some languages there were no published results of publicly available tools at the start of the project, which we signal with a dash. The observation column mentions intermediate results reported in previous deliverables. The table shows that, for three languages, there was no prior publicly available tool. For Basque and Spanish the error has been reduced 39% and 14%, respectively. These are excellent results, taking into account that the same tool is used in all languages. We also note that the results for Basque, English and Spanish have improved with respect to those reported in D5.7.

In the case of English, we provide results which are below those of the state-of-the-art publicly available tool. In Section 10 we will present the developments on the third year which improve the performance of the English. Bulgarian and Czech systems.

The results of all the languages use the ixa-pipe-wsd-ukb tool, except Czech, which does report results for the tool in Section 10.

Language	Start of the project		End of 2nd year		Observations
	tool	result	tool	result	
Basque	wsd-ukb	56.4	wsd-ukb	73.4	D5.7: 56.4
Bulgarian	—	—-	wsd-ukb	68.9	D5.7: 65.8. See Section 10.
Czech	—	—	WSD-CZ	80.5	D5.7: 80.5. See Section 10.
English	IMS	82.6	wsd-ukb	81.4	D5.7: 80.1. See Section 10.
Portuguese	—	—	WSD-PT	65.0	D5.7: 65.0
Spanish	wsd-ukb	79.4	wsd-ukb	82.1	D5.7: 79.4

Table 39: Summary of WSD results as F1 score at the end of the 2nd year. wsd-ukb stands for the ixa-pipe-wsd-ukb tool. The Observation column presents intermediate results, as reported in D5.7.

8 Evaluation of NED

In this section we introduce the evaluation datasets used for each language. We also describe the results of publicly available tools for NED in each language, both at the start of the project and at the end of the 2nd Year. We use F1 of precision and recall as the main evaluation measure, but also report precision and recall. We report results for each language in the following sections, with a summary in Table 40, Section sec:nedresults.

8.1 Basque

DBPedia Spotlight was not suitable for Basque, as it requires[56] OpenNLP models for tokenization, sentence splitting, noun phrase chunking and named entity recognition, which are not available. On top of that, lemmatization is also necessary. We thus opted to use ixa-pipe-ned-ukb, given the good results of UKB for English NED.

The ixa-pipe-ned-ukb module for Basque has been evaluated on the publicly available EDIEC (Basque Disambiguated Named Entities Corpus) dataset.[57] This dataset is a corpus of 1032 text documents with manually disambiguated named entities [Fernandez et al., 2011]. The documents are pieces of news from the 2002 year edition of the Euskaldunon Egunkaria newspaper.

There was no publicly available tool for Basque NED prior to the start of the project. Running ixa-pipe-ned-ukb on this test corpus we obtained a performance of 90.2 in precision, 87.9 in recall and 87.9 in F1 [Pérez de Viñaspre, 2015].

8.1.1 Domain evaluation

For 869 of the total NERC mentions in the QTLeap corpus that we examined, the named entity linking module was able to find a link to DBpedia resources for 252 mentions. Domain-specific entities were mostly correct, and it seems that the tool performed at the expected level. For instance, *Sareko* and *Facebook* were linked to `http://eu.dbpedia.org/resource/Internet` and `http://eu.dbpedia.org/resource/Facebook`, respectively. Even domain-specific products such as *Java* and *MB* were correctly linked to `http://eu.dbpedia.org/resource/Java_(programazio_lengoaia)` and `http://eu.dbpedia.org/resource/Megabyte`. We see, however, some room for improvement with cases such as *PS* which was incorrectly linked to the French Socialist Party `http://eu.dbpedia.org/resource/Frantziako_Alderdi_Sozialista`.

8.2 Bulgarian

The improvement of the NED module implemented during the first year of the project was organized in two ways. First, with more statistics on the basis of additional manually annotated texts. The second improvement is based on the WSD module implemented using UKB system for knowledge-based WSD. The combination is as follows - named entity is connected via DBpedia ontology to the appropriate synsets in WordNet. Then the WSD module is applied. Depending on the assigned WSD, the most frequent link is selected. These two improvements added to the performance of the module 1.8 points — see Table 40. The module is evaluated over a part of the Bulgarian treebank which was

[56] `https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Internationalization`
[57] `http://ixa2.si.ehu.es/ediec/ediec_v1.0.tgz`

annotated with URIs from DBpedia. The division between the part for evaluation and the part for counting the frequency of the URIs in the annotated data is one-to-three.

8.2.1 Domain evaluation

As it was mentioned above in Section 6.2.4., we performed the classification of domain named entities via the WSD task extending WordNet with domain terms on the basis of a domain ontology aligned to the WordNet. Thus the linking depends on the performance of WSD task. We have annotated Batch 1 with the senses from WordNet extended with domain synsets. The performance of the WSD over the correctly recognized named entities (1619 elements) is 74.8%. This is better performance than the general WSD task for Bulgarian (see below). The reason for this is that the domain terms are with lower degree of ambiguity in the Bulgarian WordNet.

8.3 Czech

There was not a publicly available NED tool for Czech prior to the project. We implemented software using interwiki links (i.e. interlingual links between Czech and English articles from Wikipedia) and DBpedia links (i.e. links between articles). There is no evaluation sets for NED task in Czech, so we have provided partial manual evaluation, described in the next block. [58]

8.3.1 Domain evaluation

Domain evaluation of NameTag results in named-entity recognition subtask was presented in Section 7.3.5.3 of D5.4. For 1,715 of total recognized entities, the named-entity linking was able to find a link to 572 DBpedia resources. Domain-specific entities were mostly correct, and it seems that the tool performed at the expected level. For instance, the terms *Gmail* and *Skype* (in any of its inflectional forms) were linked to `http://dbpedia.org/resource/Gmail` and `http://dbpedia.org/resource/Skype`, respectively. There is, however, some room for improvement in cases when NameTag marks some numbers as possible NEs and then the linking algorithm assigns a link to the corresponding page on Wikipedia. Those pages tend to refer to dates or numerical values, which usually does not make much sense in the IT domain.

8.4 English

For the evaluation of the ixa-pipe-ned module, we used the 2010 and 2011 datasets from the TAC KBP editions[59] and the AIDA corpus[60]. Because we focus our study on NED systems, we discard the so-called NIL instances (instances for which no correct entity exists in the Reference Knowledge Base) from the datasets. As the module has several

[58] In D5.7, we reported 80.30% F1 for NERC and explicitly mentioned there is no Czech publicly available test set for NED. Unfortunately, the number 80.30% was included in Table 24 to a wrong column (NED).

[59] Text Analysis Conference (TAC) for the Knowledge Base Population (KBP) track:
`https://www.ldc.upenn.edu/collaborations/current-projects/tac-kbp`
Datasets available on `https://catalog.ldc.upenn.edu/`

[60] `https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/downloads/`

parameters, it was optimized in the TAC 2010 dataset. Using the best parameter combination, the module has been evaluated on two datasets: TAC 2011 and AIDA. The best results obtained on the first dataset were 79.77 in precision and 60.68 in recall. The best performance on the second dataset is 79.67 in precision and 75.94 in recall, 77.76 F1. This last figure is the one reported in Table 40. At the start of the project there was an alternative publicly available software that performed better [Hoffart et al., 2011, Yosef et al., 2011], but the software was not amenable to be used for the rest of the languages. We tested ixa-pipe-ned-ukb, which was giving good results for English, and it provided better results than ixa-pipe-ned (which is based on DBpedia Spotlight), an F1 of 80.0 on AIDA [Agirre et al., 2015]. This is the figure reported in Table 40.

In Section 10 we will present the developments on the third year which improve current results.

8.4.1 Domain evaluation

For 1,893 of the total mentions, the named entity linking module was able to find a link to DBpedia resources for 1,445 (76.33%) mentions. Domain-specific entities were correctly linked to their DBpedia resources, and it seems that the tool performs as expected. For instance, Facebook and Google were linked to `http://dbpedia.org/resource/Facebook` and `http://dbpedia.org/resource/Google`, respectively. Even domain-specific products such as USB were correctly linked to `http://dbpedia.org/resource/Universal_Serial_Bus`. We see, however, some room for improvement with cases such as PC, for instance, which was linked to `http://dbpedia.org/resource/Microsoft_Windows`.

8.5 Portuguese

We have found some previous work which reports results for 'named entity disambiguation in Portuguese' [Santos et al., 2015] but no publicly available tool was used. Our current understanding of this previous work is that it is evaluated against an existing dataset of cross-lingual entities (in English, Spanish and Portuguese), that only entities of the type PERSON are considered, and that it is the cross-lingual linking of the entities that is evaluated as opposed to NED, per se – although we need to clarify this by exploring in greater depth the dataset on which the work is based (XLEL-21) and the task for which it was created (TAC-KBP 2013). Should it be that their results are not comparable with our own, we conclude that our work on NED-PT offers the first results of a Portuguese-specific NED setup evaluated over a gold-standard corpus.

The NED-PT tool was evaluated using a gold-standard, NE-annotated version of the CINTIL International Corpus of Portuguese [Barreto et al., 2006], consisting of 30,493 sentences including named entities linked to appropriated Portuguese Wikipedia entries in DBpedia (16,120 linked from a total of 16,371 entities, or 61.13%).

Comparing the output of the tool against the gold-standard data, 12,160 of these 16,120 manually disambiguated entities are also automatically disambiguated by NED-PT, from a total of 16,486 tagged by the program. NED-PT assigned the same DBpedia entry to the entity as was chosen by the annotator for 9484 of the 12,160 entities for which entities were assigned both manually and automatically, giving a precision of 77.99%, recall of 58.83% and F1 of 67.07%. In counting the accurate results (same DBpedia entry assigned to the entity by both the annotator and by NED-PT) we take into account those for which the assigned DBpedia entries may appear different at first glance, but in

reality redirect either to or from each other in the DBpedia and Portuguese Wikipedia hierarchies.

8.5.1 Domain evaluation

NED-PT was used to process Batch 1 and 2 of the QTLeap corpus (2000 questions and 2000 answers). From 3,799 entities found, 1,868 (49.17%) were disambiguated and linked to their Portuguese Wikipedia entries via DBpedia. Domain-specific entities are mostly correct, suggesting that the tool performs at the expected level. For example, *ISP* was linked to the Portuguese http://pt.dbpedia.org/resource/Fornecedor_de_acesso_à_Internet and subsequently http://dbpedia.org/resource/Internet_service_provider, and *écran* linked to the Portuguese http://pt.dbpedia.org/resource/Monitor_de_vídeo and subsequently http://dbpedia.org/resource/Electronic_visual_display. We do however notice some incorrect cases, most notably where the desired link for a particular entity does not share a Wikipedia/DBpedia entry in both Portuguese and English – for example *reiniciar* was linked to the Portuguese http://pt.dbpedia.org/resource/Reboot_(ficç~ao) and subsequently http://dbpedia.org/resource/Reboot_(fiction), denoting *reboot* in the sense of book and movie franchises. The desired http://dbpedia.org/resource/Reboot_(computing) does not have an equivalent entry in Portuguese, and so was not found.

8.6 Spanish

The Spanish ixa-pipe-ned module has been evaluated on the TAC 2012 Spanish dataset[61]. Starting from 2012 the TAC/KBP conference includes a task on Cross-lingual Entity Linking for Spanish and Chinese. On this setting systems are provided with a document in one language (Spanish or Chinese), and they have to link the mentions to entities belonging to an English Knowledge Base. For evaluating the system we first run Spanish NED over the TAC 2012 Spanish dataset, obtaining entities from the Spanish DBpedia. We then map those entities to the corresponding English counterparts using the interlingual links from Wikipedia[62].

DBpedia Spotlight was available prior to the project and yield a performance of 50.0 F1. This was the only tool for Spanish NED that we are aware of. Tuning the parameters and model used in DBpedia Spotlight on the train part, we were able to improve results to 78.15 in precision, 55.80 in recall, and 65.1 F1, due in part for the use of the English model [Pérez de Viñaspre, 2015]. Table 40 presents those results.

8.6.1 Domain evaluation

For 5,204 of the total mentions, the named entity linking module was able to find a link to DBpedia resources for 3,210 (61.68%) mentions. Domain-specific entities were correctly linked to their DBpedia resources, and it seems that the tool performs as expected. For instance, Facebook and Google were linked to http://es.dbpedia.org/resource/Facebook and http://es.dbpedia.org/resource/Google, respectively. Even domain-specific products such as USB and IP were correctly linked to http://es.dbpedia.org/

[61] Text Analysis Conference (TAC) for the Knowledge Base Population (KBP) track: https://www.ldc.upenn.edu/collaborations/current-projects/tac-kbp
Datasets available on https://catalog.ldc.upenn.edu/

[62] http://www.mediawiki.org/wiki/Interlanguage_links

resource/Universal_Serial_Bus and http://es.dbpedia.org/resource/Dirección_IP. We see, however, some room for improvement with incorrectly recognized entities, such as the imperative verb forms and some UI strings. Although most are not linked to DBpedia resources, some have an homonym noun which results in a link found in DBpedia: "Haz" is linked to an entry for botany http://es.dbpedia.org/resource/Haz_(botánica).

8.7 Results

Table 40 summarizes the results at the start of the project and at the end of the 2nd year for the languages in the project, alongside the publicly available tools used in each case. In some languages there was not published results of publicly available tools at the start of the project, which we signal with a dash. The observation column mentions intermediate results reported in previous deliverables. The table shows that there was no prior publicly available tool for four languages, and at the end of the second year, there was not an evaluation dataset for Czech either. For Spanish the error has been reduced 30%, a very strong result.

In the case of English, we have improved with respect to those reported in D5.7, but our results are below a publicly available tool. In Section 10 we will present the developments on the third year which improve the performance for English, albeit using a different tool.

In fact, from the six project languages, DBpedia spotlight could not be ported to Basque, Bulgarian and Czech. Adaptation to those languages requires[63] OpenNLP models for tokenization, sentence splitting, noun phrase chunking and named entity recognition, which are not available. Given the linguistic typology of these three languages, producing such models is not straightforward, and alternatively, integrating current lemmatizers in DBpedia supposes a large software re-engineering effort, out of the scope of this project.

The improvement of the NED tools for English, Bulgarian and Czech that have been performed for the third year are reported in Section 10.

Language	Start of the project		End of 2nd year		Observations
	tool	result	tool	result	
Basque	—	—	ned-ukb	87.9	D5.7: 87.9.
Bulgarian	—	—	BTB-NED	48.7	D5.7: 46.9. See Section 10.
Czech	—	—	—	—	See Section 10.
English	AIDA	82.5	ned-ukb	80.0	D5.7: 77.8. See Section 10.
Portuguese	—	—	NED-PT	67.1	D5.7: 67.1
Spanish	Spotlight	50.0	ixa-pipe-ned	65.1	D5.7: 65.1

Table 40: Summary of NED results as F1 score at the end of the 2nd year. ned-ukb refers to the ixa-pipe-ned-ukb tool. The Observation column presents intermediate results, as reported in D5.7.

[63] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Internationalization

9 Evaluation of Coreference

In this section we introduce the evaluation datasets used for each language. We also describe the results of publicly available tools for coreference in each language, both at the start of the project and at the end of the 2nd Year. The main evaluation measure used is the F1 score used in CoNLL 2011, which is the unweighted average of MUC, B-CUBED and CEAF F1 scores [Pradhan et al., 2011]. We report results for each language in the following sections, with a summary in Table 41, Section sec:corefresults.

9.1 Basque

The ixa-pipe-coref-eu module has been evaluated on the publicly available EPEC-KORREF dataset.[64] This dataset is a corpus of Basque text documents with manually annotated mentions and coreference chains, which consists of 46,383 words that correspond to 12,792 mentions. The document collection is a subpart of the Basque EPEC corpus (the Reference Corpus for the Processing of Basque) [Aduriz et al., 2006], which is a 300,000 word sample collection of news published in Euskaldunon Egunkaria, a Basque language newspaper. The tool yields 53.67 CoNLL F1 score [Soraluze et al., 2015].

9.2 Bulgarian

We had implemented a rule based module, but it is restricted mainly to relations inside of sentences. We would like to extend the module with implementation of module based on UKB system. The idea is to predict several senses for the anaphoric expression on the basis of the possible antecedents. We are working on a new module exploiting semantic information from WSD, but for the moment we do not have new results.

9.3 Czech

The Treex CR system is used for coreference resolution in Czech. It consists of three components, each of them addressing a specific type of coreference, namely:

- coreference of relative pronouns (`Treex-CR-relat`)
- coreference of reflexive and reflexive possessive pronouns (`Treex-CR-reflex`)
- coreference of personal pronouns and zeros in 3rd person (`Treex-CR-pers`)

We succeeded to significantly improve the performance of the Treex CR system from 48.0% in F-score[65] at the start of the project to the current performance of 55.1%, all measured on the evaluation set of Prague Dependency Treebank 3.0 [Bejček et al., 2013]. The overall performance rise mostly profit from the component focusing on relative pronouns. The original rule-based approach picked the antecedent using only the topology of the dependency syntactic tree. In a new version, it has been replaced by the machine learning approach, taking advantage of the other features as well, e.g. agreement

[64] http://ixa2.si.ehu.es/epec-koref/epec-koref_v1.0.tgz

[65] Calculated from the counts of how often any of the anaphor's antecedent is found, collected over each of the anaphors individually. This evaluation approach is similar to the one presented by Tuggener [2014].

on gender and number, and surface distance. This results in a substantial improvement of the `Treex-CR-relat` component from 56.6% to 73.4% in F-score. Similarly, the original rule-based approach has been substituted with the machine learning also in the `Treex-CR-reflex` component, shifting the F-score from 66.9% to 67.7%. The original machine-learning-based principle of the `Treex-CR-pers` component has remained unchanged, we only allowed ambiguities in pronouns' genders and numbers, increasing the F-score from the original 46% to the current 50.5%. In Table 41 we report the overall score of all the three components.

There also exists a system for resolution of noun phrase coreference (`CUNI-CR-NP`), which however stands apart from the Treex framework. We did no upgrades to this system, so its performance stayed unchanged on 44.4% in F-score.

9.4 English

The ixa-pipe-coref module has been evaluated on the development auto section of the CoNLL 2011 shared evaluation task[66] which uses the English language portion of the OntoNotes 4.0 corpus. We score 56.4 CoNLL F1, around 3 points below Stanford's system, which was publicly available at the start of the project.

9.5 Portuguese

There was not a publicly available tool for Portuguese at the start of the project. Our goal, as originally described in section 5.5.3 of D5.6, was to evaluate (and train) the Portuguese Coreference tool using the Summit Corpus (v3.0) [Collovini et al., 2007], a corpus of coreference for Portuguese constructed from 50 news texts from the 'caderno de Ciência da Folha de São Paulo', but dealing with inconsistencies in Summ-it has been highly problematic, both for training the tool and for later evaluation.

We thus took it upon ourselves to annotate a portion of CINTIL (5 documents so far, with ca. 3,380 tokens in total) with coreference chains and use that to train and evaluate the Portuguese Coreference tool (cf. Section 4.5.3). The model was trained over the largest of the 5 documents (with ca. 1,480 tokens), after balancing the number of true (coreferent) and false instances, and evaluated over each of the other 4 documents to obtain an average MUC F1 score of 45.9.

9.6 Spanish

There were no publicly available tools at the start of the project. The Spanish module of ixa-pipe-coref has been evaluated on the publicly available dataset distributed by the SemEval 2010 task on Multilingual Coreference resolution, in which the AnCora-ES (the Spanish part) corpus is used. The resulting CoNLL F1 is 63.40[67].

9.7 Domain evaluation

From a coreference point of view, the user scenario is quite peculiar. The user-machine interactions generally consist of one user question and one answer. The answer usually

[66] http://conll.cemantix.org/2011/introduction.html

[67] Note that the F1 score reported in D5.7 was erroneously derived, and this is the correct CoNLL F1 figure.

consists of one sentence, but occasionally a few short sentences are displayed. In this context, the number of coreferences present in the texts is low, including a few relative pronouns. As an illustrative case, in Czech, only 65 sentences out of 1000 had a relative pronoun, and only 49 sentences contained zero anaphora, personal or possessive pronouns. In Portuguese, only 2% of the 82496 markable pairs were in fact correferent. Similar statistics apply to all languages.

In addition, we observed that the user scenario text needs to be processed per interaction, that is, each user-machine interaction should be processed separately for coreference annotation.

9.8 Results

Table 41 summarizes the results at the start of the project and at the end of the 2nd year for the languages in the project, alongside the publicly available tools used in each case. In some languages there was not published results of publicly available tools at the start of the project, which we signal with a dash. The observation column mentions intermediate results reported in previous deliverables. The table shows that, for four languages, there was no prior publicly available tool. In the case of English, the ixa-pipe-coref tool provides results which are below a publicly available tool.

The results are in most cases the same as those reported in D5.7. In the case of Spanish, the F1 reported in D5.7 did not correspond to CoNLL F1, so we now provide the CoNLL F1.

The analysis performed in QTLeap, reported in D5.4 and D5.6[68], showed that generic coreference tools like those used in this section have not effect in the QTLeap domain, as coreference phenomena rarely occurs in those documents. The only exception is anaphora resolution, which has a positive effect in English to Czech translation (cf. Deliverable 5.7[69]), and which uses a specialized antecedent resolving system, rather than a generic coreference system. We thus did not improve coreference tools further.

Language	Start of the project		End of 2nd year		Observations
	tool	result	tool	result	
Basque	—	—	ixa-pipe-coref-eu	53.7	D5.7: 53.7
Bulgarian	—	—	BTB-Cor	50.6	D5.7: 50.6
Czech	Treex CR	48	Treex CR	55.1	D5.7: 50.3
English	CoreNLP	59.3	ixa-pipe-coref	56.4	D5.7: 56.4
Portuguese	—	—	LXCoref	45.9	D5.7: none
Spanish	—	—	ixa-pipe-coref	63.4	D5.7: 51.4

Table 41: Summary of coreference results as ConLL F1 score at the end of the 2nd year (except Portuguese, which reports MUC F1). The Observation column presents intermediate results, as reported in D5.7.

[68] http://qtleap.eu/wp-content/uploads/2015/07/QTLEAP-2015-D5.4.pdf
http://qtleap.eu/wp-content/uploads/2015/05/QTLEAP-2015-D5.6.pdf

[69] http://qtleap.eu/wp-content/uploads/2015/11/QTLEAP-2015-D5.7.pdf

10 Improving WSD and NED

The workplan for the WP5 working package includes research on methods to improve the advanced processors until the end of the project. The results reported in the previous sections refer to the status of advanced processors used to produce the resources in D5.8. In this section we report further improvements on some of the advanced processors, including WSD, NED and NERC for several of the project languages. QTLeap will continue to develop advanced processors until the end of the project, and the final results for those improvements will be reported in D5.11.

10.1 WSD

10.1.1 Bulgarian

During the third year of the project we proceed with addition of relations extracted from manually annotated corpora. The the new relations are added on the basis of whole sentences from these corpora. The improvement for Bulgarian is modest from 68.2 % to 68.9 %. In our view, this is because the annotation of Bulgarian treebank with senses is not completed yet. We expect better result after the completion of the annotation.

We have trained SVM classifiers for different word sense categories selected from BTB WordNet. We are using coarser-grained senses to overcome the sparseness of training examples. Another reason to explore this approach is that annotation with very fine-grained sense lexicons (such as the original WordNet) is usually very difficult even for humans and thus characterized by low inter-annotator agreement. This means that automatic approaches using such lexicons are severely constrained with regards to the levels of accuracy they can achieve. Coarsening the sense categories alleviates this problem. Beviá et al. (2015) suggest possible ways to do that. One option is to disambiguate on a very abstract level - that of WordNet domains; another one is to disambiguate at the level of basic level concepts, which may be a sensible compromise between abstractness and concreteness. The accuracy of this module is 85.27 %.

10.1.2 Czech

We trained UKB using Czech Wordnet 1.9 ([Pala et al., 2011], which can be roughly mapped to Wordnet 3.0 with some losses) and evaluated on Czech PDT corpus annotated using Czech WordNet [Pala et al., 2011]. The resulting F1-score on this dataset is 50.74%.

10.1.3 English

The model In this section we describe the experiments carried out to improve the state of the art in WSD. For such purpose we adapt a generative model for NED, which is described in Barrena et al. [2015]. The model in its basic form is Naive Bayes model that provides the synset (e) that maximizes the following probability distribution, given a mention s occurring in context $c = \{w_1, w_2, \ldots, w_n\}$:

$$\underset{e}{\text{argmax}} = P(e)P(s|e)P(c|e) \quad (1)$$

where $P(e)$ represents the probability of generating synset e, $P(s|e)$ is the probability of generating the target word s given synset e and, similarly, $P(c|e)$ is the probability of

generating the surrounding context given the synset e.[70] P71

The model can be easily extended (see the experiments) adding new information sources of more complicated distribution. In this set of experiments, we add the likelihood of the context according to distribution obtained from the posterior of UKB ($P(e|c_{grf})$). Thus, based on MLE method, we calculate the following marginals counting occurrences in Semcor [Miller et al., 1993] (see Barrena et al. [2015] for further details):

- $P(e)$: Synset prior that can be calculate directly from WN or Semcor.
- $P(s|e)$: Probability of the target word given the synset. This can be calculate directly from semcor observations.
- $P(c_{bow}|e)$: Probability of the words given the synset according Semcor counts .
- $P(c_{grf}|e)$: Probability of the words given the synset according personalized PageRank of WordNet.

Weighted Model We add complexity to the model by introducing new free exponential parameters ($\alpha, \beta, \gamma, \delta$) in the basic model: $\text{argmax}_e = P(e)^\alpha P(s|e)^\beta P(c_{bow}|e)^\gamma P(c_{grf}|e)^\delta$. In addition, we extend the model to dynamically adjust the importance of $P(c|e)$ regarding the length of the context by diving γ by number of tokens of the context. We call this *rectified* model.

Datasets Evaluation is carried out in several existing standard benchmarks for WSD. We test the models in Senseval-2 all words (S2AW), Senseval-3 all-words (S3AW), Semeval-2007 fine-grained all words (S07AW), and SemEval-2007 coarse-grained all words (S07CG). Table 42 shows the main characteristics of each dataset. Evaluation is carried in two ways. On one hand, the systems are evaluated on whole datasets (i.e. open-class words), on the other, we evaluate the systems only on nouns, as we suspect that information in SEMCOR is more suitable for nouns.

Dataset	lang	task	wn	inventory	inst	nouns	polysemy
S2AW	English	aw	WN1.7.1	fine	2422	1136	5.2
S3AW	English	aw	WN1.7.1	fine	2041	932	7.1
S07AW	English	aw	WN2.1	fine	486	315	9.2
S07CG	English	aw	WN2.1	coarse	2269	1108	4.5
S07SP	Spanish	ls	WN1.6	fine	7287	7287	2.4

Table 42: Dataset characteristics in terms of language, wordnet version, sense inventory, task (aw:all-words, ls:lexical sample), no. of instances, no. of noun instances, polysemy

Results The top 4 rows in Table 43 show the performance of different combinations among probabilities. The remaining rows show reference systems. BSL stands for a baseline system consisting of assigning the first sense of the target word. The input for BSL is the same as the one for generative model. Due to errors in the preprocessing it would

[70] Naming of the variable might be misleading, but we decide to keep the original notation for an easier comparison with the NED model.

not be able to disambiguate, and consequently the figures might differ from the ones presented in other papers. IMS is the-state-of-the-art WSD algorithm *It Make Sense* [Zhong and Ng, 2010].[71] IMS adopts support vector machines as the classifier and integrates the state of the features extractors including parts-of-speech of the surrounding words, bag of words features, and local collocations as features. IMS provides ready-to-use models trained with examples collected from parallel texts , SEMCOR [Miller et al., 1993], and the DSO corpus [Ng and Lee, 1996]. It is important to note that in comparison to our generative model IMS uses more training data and deploys more sophisticated features. In fact, the extra training data obtained from parallel corpora is not freely available, and, as such, third parties cannot develop improved versions of IMS. We thus think that our system is more suited to be used in research project like QTLeap. In the future, we expect to improve the performance of our system with extra features and more training data.

Model	S2AW		S3AW		S07AW		S07CG	
	all	noun	all	noun	all	noun	all	noun
$P(e)$	58.9	67.95	58.5	64.3	42.3	52.8	77.83	75.9
$P(e)P(s\|e)$	60.6	68.95	61.9	68.8	49.5	62.9	80.39	80.6
$P(e)P(s\|e)P(c_{bow}\|e)$	62.1	72.05	63.7	69.9	50.5	65.4	80.74	80.6
$P(e)P(s\|e)P(c_{bow}\|e)P(c_{grf}\|e)$	63.3	73.75	62.6	69.5	50.1	**66.7**	**82.3**	**82.9**
BSL	61.7	71.60	62.8	70.3	49.4	62.9	78.89	77.4
IMS	**65.6**	**76.20**	**65.7**	**71.0**	**55.9**	62.9	82.10	82.04
PPR_{w2w}	62.8	73.50	63.1	70.6	48.6	63.5	81.37	82.1

Table 43: Results of the Bayesian Generative Model across different WSD datasets. Underline denotes best model among the different distribution combination, and **bold** denotes best system in the dataset.

The results of the generative model suggest that a rather simple WSD system can obtain state of the art results, and can outperform it in a few datasets (S07AW and S07CG). As predicted, the model performs better for nouns, mainly for two reasons: 1) Bag of words features better fit with nouns, and 2) information stored in WordNet is more useful for nouns. We think that the generative model could easily extended to obtain further improvement.

Regarding the combination of generative model, results show that context agnostic models ($P(e)$, $P(e|c)$) obtain competitive results, rivaling with models that incorporate context and WordNet information. Results fo the *full model* are mixed in that is not always clear that the posterior of PPR is useful ($P(e|c_{grf})$.

10.1.4 Spanish

The model For Spanish WSD experiments we use the same model used for English experiments. We hypothesize that most of the parameters learned for WSD are independent of the language and, thus, we can use the model estimated in SEMCOR to disambiguate Spanish text input.

Testing input is first translated to English using QTLeap technology. In addition we keep track of target words via translation alignments and, this way, we simply apply

[71] IMS is a freely available Java implementation http://www.comp.nus.edu.sg/~nlp/software.html

the probability distributions learned on the English side. We only used the Spanish WordNet to set the sense inventory of the target words (which originally are in Spanish). Experiments showed that constraining synset candidates to target language (i.e. Spanish) improve significantly the results.

Datasets Cross-lingual evaluation is carried out on the Spanish all words task of Semeval-2007. Bottom row of Table 43 shows the main characteristics of the whole dataset. The dataset is divided in training and test set, in which the latter includes in-domain and out-of-domain texts. As our model relies on SEMCOR estimates we use both sets as test sets (although we carried out some development on training set).

Evaluation is carried in three ways. In the first one, the systems are evaluated on whole datasets (i.e. all nouns), and as the dataset contains many monosemous target words, we create two additional evaluation: on one hand, we evaluate the systems on polysemous nouns, on the other hand, we selected lemmas that have more than 15 occurrences in training.

Results The top 4 rows in Table 44 show the performance of different combinations among probabilities. The remaining rows show reference systems. BSL stands for a baseline system consisting of assigning the first sense of the target word. Note that BSL is using English WordNet on translated target words, as we do not have reliable counts for the Spanish WordNet. TRAIN-MFS stands for a system that assigns the most frequent sense in the training set. Finally, $\text{PPR}_w 2w$ is the personalized PageRank of WordNet implemented in UKB[72], which was state-of-the-art system up to date .

The bottom 2 rows of the table show the results published in [Màrquez et al., 2007], and due to some differences in the implementation we obtain very different results. SVM is a Support Vector Classifier trained by the organizers of the task, so that is why only have the results for the test set. TRAIN-MFS is their most frequent sense algorithm. They also published the performance on frequent words, but we are not sure how exactly they filter out non frequent words. Similarly, both versions of TRAIN-MFS show very different behavior. Our in-home version of the TRAIN-MFS takes translated test input, which it can be the reason of such differences.

The results of the generative model show that improves by wide margin the state-of-the-art regarding PPR_{w2w}, which could be considered the best generic system for Spanish WSD. Positively, the generative model also outperforms the SVM, specifically trained for the current dataset, which shows that parameters learned in one language (English this case) can accurately used in other language. In a similar vein, the results show that Machine Translation can be useful for cross-lingual semantic tasks. The significant improvement comes when *full model* is applied, which outperform PPR_{w2w} in more than 8 points in the test set, and almost 1.5 points to SVM model. We show that the generative model is useful model to combine different probability sources under the same framework.

In home created TRAIN-MFS shows excellent performances also in the test set. The generative models only outperforms it when evaluating on the whole test set (*all* column), but this is due mainly because TRAIN-MFS obtain very low recall (64.1%) as an effect of incorrect translations. On the other hand, the poor result of BSL shows the importance of having a sense inventory for the target word. BSL being the equivalent of $P(e)P(s|e)$ its performance is about 30 points lower.

[72] xa2.si.ehu.es/ukb/

Model	TRAIN			TEST		
	all	poly	selected	all	poly	selected
$P(e)$	76.8	47.3	59.4	78.8	50.3	60.1
$P(e)P(s\|e)$	79.3	53.0	64.1	78.0	48.5	58.0
$P(e)P(s\|e)P(c_{bow}\|e)$	79.8	54.2	65.4	78.3	49.1	60.4
$P(e)P(s\|e)P(c_{bow}\|e)P(c_{grf}\|e)$	81.2	57.3	**67.5**	**86.5**	**67.4**	**76.5**
BSL	50.2	39.3	46.5	51.6	36.9	45.7
Train-MFS	**88.1**	**73.0**	62.6	71.5	63.4	72.5
PPR_{w2w}	78.6	51.5	62.6	77.9	44.6	56.3
Train-MFS [Màrquez et al., 2007]	-	-	-	84.2	-	61.8*
SVM [Màrquez et al., 2007]	-	-	-	85.1	-	65.2*

Table 44: Results of the Bayesian Generative Model in Spanish all words Underline denotes best model among the different distribution combination, and **bold** denotes best system in the dataset. $*$ selected words published as in [Màrquez et al., 2007].

10.2 NED

10.2.1 Bulgarian

We are proceeding with the improvement of the NED module for Bulgarian extending BTB WordNet with instances from DBpedia and including relations from it.

10.2.2 Czech

CUNI is adapting the ixa-pipe-ukb-ned tool to work with the Czech Wikipedia. In parallel, it is planning the evaluation of the output of the tool, as there are no publicly available NED evaluation corpora for Czech.

10.2.3 English

UPV/EHU is trying to improve the results of the ixa-pipe-ned tool (based on the third-party Spotlight tool) for English. They have explored the use of ixa-pipe-ukb-ned, with positive results, but still slightly below the state-of-the-art for English (cf. Section 7). They are currently studying a fast implementation and public release of their successful prototype Barrena et al. [2015], which reported an 84.9 F1 result, beyond the previous state-of-the-art. At the same time, they are also checking the use of the NED tool developed by DFKI Weissenborn et al. [2015][73], which reports an F1 of 85.1 F1. The advantage of the first approach is that it should be easily portable to other QTLEAP languages, and that it uses Wikipedia and does not require a proprietary knowledge base.

10.3 NERC

Although no improvement for NERC was in the initial plans, UPV/EHU has improved the performance of its ixa-pipe-nerc tool for Basque, English and Spanish, improving the state-of-the-art in all three languages. We thus report the current best results, which can be compared to the tools evaluated in Section 6

[73] https://bitbucket.org/dfki-lt-re-group/mood

10.3.1 Basque

A fraction of the EPEC corpus, consisting in 60.000 tokens, was manually annotated with 4748 named entities. When evaluated over a subset of ca. 15,000 tokens, ixa-pipe-nerc's F1 measure is 76.72% on 3 class evaluation and 75.40 on 4 classes.

10.3.2 English

The ixa-pipe-nerc module has been evaluated on the CONLL 2002[74] and 2003[75] datasets. Trained on local features only obtains F1 84.53, and the models with external knowledge F1 87.11. The Ontonotes CoNLL 4 NE types with local features model obtains F1 86.21. The Ontonotes 3 NE types with local features configuration obtains F1 89.41.

10.3.3 Spanish

ixa-pipe-nerc module for Spanish currently obtains the best results training Maximum Entropy models on the CoNLL 2002 dataset. Our best model obtains 80.16 F1 vs 81.39 F1 of (Carreras et al., 2002), the best result so far on this dataset. Their result uses external knowledge and without it, their system obtains 79.28 F1.

10.3.4 Results

Language	Start of the project		End of 2nd year		Observations
	tool	result	tool	result	
Basque	Eihera	71.35	ixa-pipe-nerc	75.70	current soa
English	Illinois NER tagger	90.57	ixa-pipe-nerc	91.36	current soa
Spanish	Freeling	81.39	ixa-pipe-nerc	84.30	current soa

Table 45: Summary of improvements NERC results. Soa stand for state-of-the-art.

[74] http://www.clips.ua.ac.be/conll2002/ner/
[75] http://www.clips.ua.ac.be/conll2003/ner/

11 Final remarks

This deliverable reports on the LRTs curated and produced in WP5 in the project, as described in D1.3, D1.7 and D1.10, comprising 6 languages (BG Bulgarian, CS Czech, EN English, ES Spanish, EU Basque, PT Portuguese). This deliverable describes the LRTs in D5.8, and includes the materials in earlier releases (D5.3 and D5.6). It includes three multilingual corpora:

- QTLeap WSD/NED corpus. It contains annotated versions of Batches 1 and 2 of the QTLeap corpus, including BG, CS, EN, ES, EU and PT.
- Europarl-QTLeap WDS/NED corpus. In addition to BG, CS, EN, ES and PT subsets of the Europarl parallel corpus, it contains an EN-EU parallel corpus from non-Europarl sources.
- News-QTLeap WSD/NED corpus. It contains annotated versions of WMT 2011 and 2012 test corpora in BG, CS, EN, ES, EU and PT.

The three multilingual corpora have been annotated with WSD, NED and coreference information. The goal was to provide 1M tokens from parallel corpora, and 10M tokens from comparable corpora. All languages focused on parallel corpora, as it enables better machine translation, and in some cases exceeded the required sizes.

This deliverable also describes the basic processing tools for the 6 target languages, including PoS taggers, lemmatizers and NERC. These tools have been evaluated in standard datasets, and we also provide some qualitative analysis of their output in the IT domain scenario.

Regarding advanced tools like NED, WSD and Coreference, QTLeap has made a large effort in searching, adapting and in some cases developing publicly available tools for all languages. The effort includes, in some languages, the annotation of evaluation corpora. The results of the evaluation show that in many cases tools for most of the languages were missing at the start of the project. For the cases where tools were available at the start of the project, the QTLeap tools provide similar, or, in most of the cases, better results than other tools.

Finally, QTLeap is still improving some of the advanced tools. Section 10 reported the interim improvements on some advanced processors. The final results for those improvements will be reported in D5.11.

A Examples of annotations

This appendix presents the output examples of lemmatizer and PoS tagger for different languages when run on the user scenario texts.

A.1 Basque

Lemmatizer

Ez	lemma="ez"
dakit	lemma="jakin"
Wi-Fi	lemma="Wi-Fi"
sarearen	lemma="sare"
pasahitza	lemma="pasahitz"
zein	lemma="zein"
den	lemma="izan"
.	lemma="."
Facebook	lemma="Faceboo"
aplikazioa	lemma="aplikazio"
ez	lemma="ez"
dabil	lemma="ibili"
nire	lemma="ni"
iPhone-an	lemma="iphone"
.	lemma="."

PoS tagger

Ez	pos="PRT EGI"	(truth partiple)
dakit	pos="ADT"	(synthetic verb)
Wi-Fi	pos="IZE IZB"	(proper noun)
sarearen	pos="IZE ARR"	(common noun)
pasahitza	pos="IZE ARR"	(common noun)
zein	pos="DET NOLGAL"	(interrogative determiner)
den	pos="ADT"	(synthetic verb)
.	pos="PUNT_PUNT"	(full stop)
Facebook	pos="IZE IZB"	(proper noun)
aplikazioa	pos="IZE ARR"	(common noun)
ez	pos="PRT EGI"	(truth partiple)
dabil	pos="ADT"	(synthetic verb)
nire	pos="IOR PERARR"	(personal pronoun)
iPhone-an	pos="IZE ARR"	(common noun)
.	pos="PUNT_PUNT"	(full stop)

A.2 Bulgarian

Lemmatizer

Idete	otida
v	v
System	system
Tools	tools
>	>
System	system
Restore	restore
i	i
sledvajte	sledvam
instrukciite	instrukciya
.	.

PoS tagger

Idete	Vppiz–2p
v	R
System	Np
Tools	Np
>	punct
System	Np
Restore	Hmsi
i	Cp
sledvajte	Vpitz–2p
instrukciite	Ncfpd
.	punct

A.3 Czech

Lemmatizer

Omylem	omyl
jsem	být
odstranil	odstranit_:W
soubor	soubor
z	z-1
Google	Google_;K
Drive	drive_;c_,t
.	.
Mohu	moci_^(mít_možnost_[něco_dělat])
jej	on-1
získat	získat_:W
zpět	zpět
?	?
Zkuste	zkusit
na	na-1
webových	webový_,t
stránkách	stránka
(	(
https	https
:	:
/	/
/	/
drive.google.com	drive.google.co
)	)
zkontrolovat	zkontrolovat_:W
,	,
jestli	jestli
nebude	být
na	na-1
kartě	karta
Bin	bin-
(	2_,t_^(angl._koš,_válec)
Koš	(
)	koš
.	)
	.

PoS tagger P80

Jak	pos="D"	morphofeat="Db————-"
mohu	pos="V"	morphofeat="VB-S—1P-AA–1"
ve	pos="R"	morphofeat="RV–6———-"
Photoshopu	pos="N"	morphofeat="NNIS6——A—-"
uložit	pos="V"	morphofeat="Vf———A—-"
obrázek	pos="N"	morphofeat="NNIS4——A—-"
jako	pos="J"	morphofeat="J,————-"
jpeg	pos="N"	morphofeat="NNIS4——A—-"
místo	pos="R"	morphofeat="RR–2———-"
png	pos="N"	morphofeat="NNFXX——A—8"
?	pos="Z"	morphofeat="Z:————-"

A.4 English

Lemmatizer

My	lemma="my"
Gmail	lemma="Gmail"
shortcut	lemma="shortcut"
icon	lemma="icon"
has	lemma="have"
disappeared	lemma="disappear"
from	lemma="from"
the	lemma="the"
desktop	lemma="desktop"
.	lemma="."
There	lemma=”there”
is	lemma=”be”
no	lemma=”no”
sound	lemma=”sound”
coming	lemma=”come”
from	lemma=”from”
the	lemma=”the”
speakers	lemma=”speaker”
.	lemma=”.”

PoS tagger

My	pos="Q" morphofeat="PRP$"
Gmail	pos="R" morphofeat="NNP"
shortcut	pos="N" morphofeat="NN"
icon	pos="N" morphofeat="NN"
has	pos="V" morphofeat="VBZ"
disappeared	pos="V" morphofeat="VBN"
from	pos="P" morphofeat="IN"
the	pos="D" morphofeat="DT"
desktop	pos="N" morphofeat="NN"
.	pos="O" morphofeat="."

A.5 Portuguese

Lemmatizer

Restaurar	RESTAURAR
um	_
backup	BACKUP (English word tagged as common noun)
de	_
os	_
emails	EMAILS (English word tagged as common noun)
para	_
o	_
Outlook	_

Mudar	MUDAR
nome	NOME
de	_
a	_
rede	REDE
wifi	WIFER (English word tagged as verb)

PoS tagger

Ativar	PNM (proper name)
modo	CN (common noun)
de	PREP (preposition)
hibernar	V (verb)
em	PREP (preposition)
o	DA (definite article)
windows	CN (common noun)
xp	ADJ (adjective)

A.6 Spanish

Lemmatizer

No	lemma="no"
puedo	lemma="poder"
acceder	lemma="acceder"
a	lemma="a"
los	lemma="el"
emails	lemma="emails"
.	lemma="."
Quiero	lemma="quiero"
desinstalar	lemma="desinstalar"
algunos	lemma="alguno"
programas	lemma="programa"
de	lemma="de"
Windows	lemma="Windows"
.	lemma="."

PoS tagger

No	pos="A" morphofeat="RN"
puedo	pos="V" morphofeat="VMIP1S0"
acceder	pos="V" morphofeat="VMN0000"
a	pos="P" morphofeat="SPS00"
los	pos="D" morphofeat="DA0MP0"
emails	pos="N" morphofeat="NCMP000"
.	pos="O" morphofeat="FP"
Quiero	pos="C" morphofeat="CC"
desinstalar	pos="V" morphofeat="VMN0000"
algunos	pos="D" morphofeat="DI0MP0"
programas	pos="N" morphofeat="NCMP000"
de	pos="P" morphofeat="SPS00"
Windows	pos="R" morphofeat="NP00000"
.	pos="O" morphofeat="FP"

B Summary of availability

Name of LRT	language	QTLeap	License	URL
AnCora (Lemma./PoS)	ES	No	*Check with authors*	http://clic.ub.edu/corpus/en/ancora
BulTreeBank (Lemma./PoS, CR, NERC)	BG	No	CC-BY-NC-SA 4.0	http://www.bultreebank.org/dpbtb/
BulTreeBank-DB (NED, WSD)	BG	Yes	CC-BY-NC-SA v4.0	http://www.bultreebank.org/QTLeap/
CoNLL 2002 (NERC)	ES	No	*Check with authors*	http://www.cnts.ua.ac.be/conll2002/ner.tgz
CoNLL 2003 (NERC)	EN	No	*Check with authors*	http://www.cnts.ua.ac.be/conll2003/ner.tgz
CoNLL 2011 (CR)	EN	No	*Check with authors*	http://conll.cemantix.org/2011/data.html
Czech Named Entity Corpus 2.0 (NERC, NED)	CS	No	CC BY-NC-SA 3.0	http://hdl.handle.net/11858/00-097C-0000-0023-1B22-8
EPEC (CR)	EU	No	CC BY 4.0	http://ixa2.si.ehu.es/epec-koref/epec-koref_v1.0.tgz
EuSemcor (WSD)	EU	No	CC BY 3.0	http://ixa2.si.ehu.es/mcr/EuSemcor.v1.0/EuSemcor_v1.0.tgz
Euskaldunon Egunkaria (NERC)	EU	No	CC BY 4.0	http://ixa2.si.ehu.es/eiec/eiec_v1.0.tgz
Euskaldunon Egunkaria (NED)	EU	No	CC BY 4.0	http://ixa2.si.ehu.es/ediec/ediec_v1.0.tgz
Prague Dependency Treebank 3.0 (Lemma./PoS, CR, WSD)	CS	No	CC BY-NC-SA 3.0	http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3
Semeval 2010 (CR)	ES	No	*Check with authors*	http://www.lsi.upc.edu/~esapena/downloads/index.php?id=1
TAC 2010/2011 (NED)	EN	No	LDC User Agreement for NonMembers	https://catalog.ldc.upenn.edu/LDC2014T16
TAC 2012 (NED) registered TAC 2012	ES	No	Restricted to	http://www.nist.gov/tac/2012/KBP/data.html
WSJ Treebank (Lemma./PoS)	EN	No	*Check with authors*	http://www.cis.upenn.edu/ treebank/

Table 46: Summary of publicly available LRTs mentioned: Datasets. QTLeap column for those LRTs which have been (partially) funded by QTLeap.

Name of LRT	language	QTLeap	License	URL
Basque DBpedia 3.9	EU	No	CC BY-SA 3.0	http://downloads.dbpedia.org/3.9/eu/
Bulgarian DBpedia	BG	No	CC BY-SA 3.0	http://downloads.dbpedia.org/3.9/bg
Bulgarian WordNet	BG	Yes	CC BY 3.0	http://www.bultreebank.org/QTLeap/ http://compling.hss.ntu.edu.sg/omw/
Czech DBpedia	CS	No	CC BY-SA 3.0	http://downloads.dbpedia.org/3.9/cs/
Czech WordNet	CS	No	CC BY-NC-SA 3.0	http://hdl.handle.net/11858/00-097C-0000-0001-4880-3
English DBpedia 3.9	EN	No	CC BY-SA 3.0	http://downloads.dbpedia.org/3.9/en/
Mapping WordNet DBpedia	EN	No	CC BY-NC-SA 3.0	http://ixa2.si.ehu.es/mcr/mapping_wn_dbpedia_v1.0.tgz
Spanish DBpedia 3.9	ES	No	CC BY-SA 3.0	http://downloads.dbpedia.org/3.9/es/
WordNet 3.0	EU, EN, ES	No	WordNet license CC BY-NC-SA 3.0 CC BY 3.0	http://adimen.si.ehu.es/web/files/mcr30/mcr30.tar.bz2

Table 47: Summary of publicly available LRTs: Ontologies. QTLeap column for those LRTs which have been (partially) funded by QTLeap.

Name of LRT	language	QTLeap	License	URL
Europarl-QTLeap WSD/NED	BG,CS,EN, ES,EU,PT	Yes	CC-BY v4.0	http://metashare.metanet4u.eu/go2/europarl-qtleap-wsdned-corpus https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1477
QTLeap WSD/NED	BG,CS,EN, ES,EU,PT	Yes	CC-BY-NC-SA v4.0	http://metashare.metanet4u.eu/go2/qtleap-wsdned-corpus https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1476
News-QTLeap WSD/NED corpus	BG,CS,EN, ES,EU,PT	Yes	CC-BY-NC v4.0	http://metashare.metanet4u.eu/go2/news-qtleap-wsdned-corpus
SETIMES corpus	BG, EN	Yes	CC-BY-NC-SA 4.0	http://www.bultreebank.org/EMP/
Bulgarian Radio QTLeap WSD/NED corpus	BG, EN	Yes	CC-BY-NC-SA 4.0	http://metashare.metanet4u.eu/go2/radio-bulgaria-wsdned-corpus
Wikipedia corpus	BG	Yes	CC-BY-NC-SA 4.0	http://metashare.metanet4u.eu/go2/bulgarian-english-wikipedia-wsdned-corpus

Table 48: Summary of publicly available LRTs: Annotated corpora. QTLeap column for those LRTs which have been (partially) funded by QTLeap. QTLeap corpus are also available through CLARIN Lindat (https://lindat.mff.cuni.cz/)

Name of LRT	language	QTLeap	License	URL
Bulgarian NLP pipeline	BG	Yes	GPL v3.0	http://www.bultreebank.org/QTLeap/
ixa-pipe-coref	EN, ES	No	APL 2.0	https://bitbucket.org/Josu/corefgraph
ixa-pipe-ned	EN, ES	No	GPL v3.0	https://github.com/ixa-ehu/ixa-pipe-ned
ixa-pipe-nerc	EN, ES, EU	No	APL 2.0	https://github.com/ixa-ehu/ixa-pipe-nerc/
ixa-pipe-pos	EN, ES	No	APL 2.0	https://github.com/ixa-ehu/ixa-pipe-pos/
ixa-pipe-pos-eu	EU	Yes	GPL v3.0	http://ixa2.si.ehu.es/ixa-pipes/eu/ixa-pipe-pos-eu.tar.gz http://metashare.metanet4u.eu/go2/ixa-pipe-pos-eu
ixa-pipe-wsd-ukb	EN, ES	No	GPL v3.0	http://ixa2.si.ehu.eus/ixa-pipes/eu/ixa-pipe-wsd-ukb.tar.gz
ixa-pipe-ned-ukb	EU	Yes	GPLv3.0	http://ixa2.si.ehu.eus/ixa-pipes/eu/ixa-pipe-ned-ukb.tar.gz
ixa-pipe-wsd-ukb	EU	No	GPLv3.0	http://ixa2.si.ehu.eus/ixa-pipes/eu/ixa-pipe-wsd-ukb.tar.gz
ixa-pipe-coref-eu	EU	Yes	GPLv3.0	http://ixa2.si.ehu.eus/ixa-pipes/eu/ixa-pipe-coref-eu.tar.gz
MorphoDita	CS, EN	No	CC BY-NC-SA 3.0	http://ufal.mff.cuni.cz/morphodita
MorphoDiTa Treex wrapper	CS	Yes	GPLv3.0 + Perl Artistic	https://github.com/ufal/treex/
NameTag	CS, EN	No	CC BY-NC-SA 3.0	http://ufal.mff.cuni.cz/nametag
NameTag Treex wrapper	CS	Yes	GPLv3.0 + Perl Artistic	https://github.com/ufal/treex/
Portuguese coreference tool	PT	Yes	Apache License 2.0	http://nlx-server.di.fc.ul.pt/~jsilva/Coref-PT.tgz

Table 49: Summary of publicly available LRTs: Processing tools. QTLeap column for those LRTs which have been (partially) funded by QTLeap. QTLeap corpus are also available through CLARIN Lindat (https://lindat.mff.cuni.cz/)

References

I. Aduriz, M. Aranzabe, J. M. Arriola, A. Atutxa, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, M. Oronoz, A. Soroa, and R. Urizar. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for automatic processing. *Language and Computers*, 56(1):1–15, 2006.

R. Agerri, J. Bermudez, and G. Rigau. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), may 2014.

E. Agirre and A. Soroa. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41. Association for Computational Linguistics, 2009.

E. Agirre, O. López de Lacalle, and A. Soroa. Random Walks for Knowledge-based Word Sense Disambiguation. *Comput. Linguist.*, 40(1):57–84, March 2014. ISSN 0891-2017.

E. Agirre, A. Barrena, and A. Soroa. Studying the Wikipedia Hyperlink Graph for Relatedness and Disambiguation. 2015. URL http://arxiv.org/abs/1503.01655.

I. Alegria, M. Aranzabe, A. Ezeiza, N. Ezeiza, and R. Urizar. Robustness and customisation in an analyser/lemmatiser for Basque. In *LREC-2002 Customizing knowledge in NLP applications workshop, pages 1-6, Las Palmas de Gran Canaria, 28th May 2002"*, 2002.

A. Barrena, E. Agirre, and A. Soroa. UBC entity linking at TAC-KBP 2013: random forests for high accuracy. In *Proceedings of the Sixth Text Analysis Conference, TAC 2013, Gaithersburg, Maryland, USA, November 18-19, 2013*. NIST, 2013.

A. Barrena, A. Soroa, and E. Agirre. Combining Mention Context and Hyperlinks from Wikipedia for Named Entity Disambiguation. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 101–105, Denver, Colorado, June 2015. Association for Computational Linguistics.

F. Barreto, A. Branco, E. Ferreira, A. Mendes, M. F. Bacelar Nascimento, F. Nunes, and J. Silva. Open Resources and Tools for the Shallow Processing of Portuguese: The TagShare Project. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC '06, pages 1438–1443, 2006.

E. Bejček, E. Hajičová, J. Hajič, P. Jínová, V. Kettnerová, V. Kolářová, M. Mikulová, J. Mírovský, A. Nedoluzhko, J. Panevová, L. Poláková, M. Ševčíková, J. Štěpánek, and S. Zikánová. Prague Dependency Treebank 3.0, 2013.

E. Bejček, J. Panevová, J. Popelka, P. Straňák, M. Ševčíková, J. Štěpánek, and Z. Žabokrtský. Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In *Proceedings of COLING 2012: Technical Papers*, Mumbai, 2012.

O. Bojar, Z. Žabokrtský, M. Janíček, V. Klimeš, J., D. Mareček, V. Novák, M. Popel, and J. Ptáček. CzEng 0.9, 2009.

O. Bojar, Z. Žabokrtský, O. Dušek, P. Galuščáková, M. Majliš, D. Mareček, J. Maršík, M. Novák, M. Popel, and A. Tamchyna. The joy of parallelism with CzEng 1.0. In *Proceedings of LREC 2012*, Istanbul, Turkey, 2012. European Language Resources Association.

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

A. Branco and J. Silva. Dedicated Nominal Featurization of Portuguese. In *PROPOR*, 2006a.

A. Branco and J. Silva. Very High Accuracy Rule-based Nominal Lemmatization with a Minimal Lexicon. In *Actas do XXI Encontro Anual da Associaccao Portuguesa de Linguistica*, 2007.

A. Branco and J. R. Silva. A Suite of Shallow Processing Tools for Portuguese: LX-suite. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, EACL '06, pages 179–182, Stroudsburg, PA, USA, 2006b. Association for Computational Linguistics.

A. Branco, F. Nunes, and J. Silva. Verb Analysis in an Inflective Language: Simpler is better. In *University of Lisbon, Department of Informatics, NLX-Natural Language and Speech Group*, 2006.

P. C. F. Cardoso, E. G. Maziero, M. L. R. Castro Jorge, E. M. R. Seno, A. Di Felippo, L. H. M. Rino, M. das Graças V. Nunes, and T. A. S. Pardo. CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In *Proceedings of the Third Annual RST and Text Studies Workshop*, pages 88–105, 2011.

X. Carreras, L. Màrquez, and L. Padró. Named Entity Extraction Using AdaBoost. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

S. Cinková. From PropBank to EngValLex: adapting the PropBank-Lexicon to the valency theory of the functional generative description. In *Proceedings of LREC 2006, Genova, Italy*, 2006.

M. Collins. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

S. Collovini, T. I. Carbonel, J. T. Fuchs, J. C. Coelho, L. Rino, and R. Vieira. Summit: Um Corpus Anotado com Informações Discursivas Visando à Sumarização Automática. In *Proceedings of the 5th Workshop on Information and Human Language Technology*, TIL'2007, Rio de Janeiro, Brazil, 2007.

J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, pages 121–124. ACM, 2013.

J. G. C. de Souza, P. N. Gonçalves, and R. Vieira. Learning Coreference Resolution for Portuguese Texts. In A. Teixeira, V. de Lima, L. de Oliveira, and P. Quaresma, editors, *Computational Processing of the Portuguese Language*, volume 5190 of *Lecture Notes in Computer Science*, pages 153–162. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-85979-6.

P. Denis and J. Baldridge. A Ranking Approach to Pronoun Resolution. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1588–1593, 2007.

O. Dušek, J. Hajic, and Z. Uresova. Verbal Valency Frame Detection and Selection in Czech and English. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 6–11. Association for Computational Linguistics, 2014.

O. Dusek, E. Fukicova, J. Hajic, M. Popel, J. Sindlerova, and Z. Uresova. Using Parallel Texts and Lexicons for Verbal Word Sense Disambiguation. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 82–90. Uppsala University, 2015. ISBN 978-91-637-8965-6.

C. Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.

I. Fernandez, I. Alegria, and N. Ezeiza. Semantic Relatedness for Named Entity Disambiguation Using a Small Wikipedia. In Ivan Habernal and Václav Matoušek, editors, *Text, Speech and Dialogue*, volume 6836 of *Lecture Notes in Computer Science*, pages 276–283. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-23537-5.

S. Fernando and M. Stevenson. Mapping WordNet synsets to Wikipedia articles. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

E. Ferreira, J. Balsa, and A. Branco. Combining rule-based and statistical methods for named entity recognition in portuguese. In *In V Workshop em Tecnologia da Informação e da Linguagem Humana*, pages 1615–1624, 2007.

A. Fokkens, A. Soroa, Z. Beloki, N. Ockeloen, G. Rigau, W. Robert van Hage, and P. Vossen. NAF and GAF: Linking Linguistic Annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, 2014.

G. Georgiev, V. Zhikov, P. Osenova, K. Simov, and P. Nakov. Feature-rich Part-of-speech Tagging for Morphologically Complex Languages: Application to Bulgarian. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 492–502, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. ISBN 978-1-937284-19-0.

J. Giménez and L. Màrquez. Svmtool: A general pos tagger generator based on support vector machines. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 43–46, 2004.

A. Gonzalez-Agirre, G. Laparra E., and Rigau. Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. 2012. ISSN ISBN 978-80-263-0244-5.

J. Hajič, M. Ciaramita, R. Johansson, D. Kawahara, M. A. Martí, L. Màrquez, A. Meyers, J. Nivre, S. Padó, J. Štěpánek, P. Straňák, M. Surdeanu, N. Xue, and Y. Zhang. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL-2009*, Boulder, Colorado, USA, 2009.

J. Hajič, J. Panevová, Z. Urešová, A. Bémová, V. Kolářová, and P. Pajas. PDT-VALLEX: creating a large-coverage valency lexicon for treebank annotation. In *Proceedings of The 2nd Workshop on Treebanks and Linguistic Theories*, volume 9, page 57–68, 2003.

J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský, M. Ševčíková Razímová, and Z. Urešová. *Prague Dependency Treebank 2.0*. Number LDC2006T01. LDC, Philadelphia, PA, USA, 2006.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009. ISSN 1931-0145.

J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust Disambiguation of Named Entities in Text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 793–803, Edinburgh, Scotland, UK, 2011. The Association for Computational Linguistics.

V. Honetschläger. Using a Czech valency lexicon for annotation support. In *Text, Speech and Dialogue*, pages 120–125. Springer, 2003.

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT.

H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky. Stanford's Multi-pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 28–34, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 9781937284084.

H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky. Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules. *Comput. Linguist.*, 39(4):885–916, 2013.

L. Màrquez, M. A. Villarejo, T. Martí, and M. Taulé. SemEval-2007 Task 09: Multilevel Semantic Annotation of Catalan and Spanish. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007) in conjunction with ACL*, pages 42–47, Prague, Czech Republic, 2007.

G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker. A Semantic Concordance. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, pages 303–308, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. ISBN 1-55860-324-7.

C. Müller and M. Strube. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany, 2006.

MultiWordNet. The MultiWordNet project. http://multiwordnet.fbk.eu/english/home.php, n.d. Accessed: 2015-01-13.

R. Navigli and S. P. Ponzetto. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Artif. Intell.*, 193: 217–250, 2012. ISSN 0004-3702.

R. Navigli, K.C. Litkowski, and O. Hargraves. SemEval-2007 Task 07: Coarse-grained English All-words Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 30–35. Association for Computational Linguistics, 2007.

H. T. Ng and H. Lee. Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-based Approach. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 40–47, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics.

G. L. Nguy. Návrh souboru pravidel pro analýzu anafor v českém jazyce. Master's thesis, MFF UK, Prague, Czech Republic, 2006. In Czech.

G. L. Nguy, V. Novák, and Z. Žabokrtský. Comparison of classification and ranking approaches to pronominal anaphora resolution in Czech. In *Proceedings of the SIGDIAL 2009 Conference*, London, UK, 2009. The Association for Computational Linguistics. ISBN 978-1-932432-64-0.

F. A. A. Nóbrega and T. A. S. Pardo. General Purpose Word Sense Disambiguation Methods for Nouns in Portuguese. *Computational Processing of the Portuguese Language*, 8775:94–101, 2014.

M. Novák and Z. Žabokrtský. Resolving Noun Phrase Coreference in Czech. *Lecture Notes in Computer Science*, 7099:24–34, 2011. ISSN 0302-9743.

K. Pala, T. Čapek, B. Zajíčková, D. Bartůšková, K. Kulková, P. Hoffmannová, E. Bejček, P. Straňák, and J. Hajič. Czech WordNet 1.9 PDT, 2011. URL http://hdl.handle.net/11858/00-097C-0000-0001-4880-3.

M. Pazienza, M. Pennacchiotti, and F. Zanzotto. Mixing WordNet, VerbNet and PropBank for studying verb relations. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA), 2006.

E. Pociello, E. Agirre, and I. Aldezabal. Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, 45(2):121–142, 2011. ISSN 1574-020X.

S. Pradhan, L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, and N. Xue. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

J. Pérez de Viñaspre. Wikipedia eta anbiguetate lexikala. Technical report, Computer Science Faculty, University of the Basque Country, 2015.

K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning. A Multi-pass Sieve for Coreference Resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 492–501, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

J.T.L. Santos, I.M. Anastácio, and B.E. Martins. Named entity disambiguation over texts written in the portuguese or spanish languages. *IEEE Latin America Transactions*, 13 (3):856–862, 2015.

J. Semecký. Verb valency frames disambiguation. *The Prague Bulletin of Mathematical Linguistics*, (88):31–52, 2007.

K. Simov, P. Osenova, and M. Slavcheva. BTB:TR03: BulTreeBank morphosyntactic tagset BTB- TS version 2.0. 2004.

K. I. Simov and P. N. Osenova. A Hybrid System for MorphoSyntactic Disambiguation in Bulgarian. In *In: Proc. of the RANLP 2001 Conference, Tzigov chark*, pages 5–7, 2001.

Kiril Simov, Alexander Popov, and Petya Osenova. Improving word sense disambiguation with linguistic knowledge from a sense annotated treebank. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 596–603, 2015. URL http://www.aclweb.org/anthology/R15-1077.

A. Soraluze, O. Arregi, X. Arregi, and A. Díaz de Ilarraza. Coreference Resolution for Morphologically Rich Languages. Adaptation of the Stanford System to Basque. *Procesamiento del Lenguaje Natural*, 55:23–30, 2015. ISSN 1989-7553.

J. Straková, M. Straka, and J. Hajič. *"Text, Speech, and Dialogue: 16th International Conference, TSD 2013, Pilsen, Czech Republic, September 1-5, 2013. Proceedings"*, chapter A New State-of-The-Art Czech Named Entity Recognizer, pages 68–75. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-40585-3.

J. Straková, M. Straka, and J. Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

J. Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218. European Language Resources Association (ELRA), 2012.

K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

D. Tuggener. Coreference Resolution Evaluation for Higher Level Applications. In G. Bouma and Y. Parmentier, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 231–235. The Association for Computer Linguistics, 2014.

Z. Urešová. *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia, ISBN 978-80-904571-1-9, 375 pp., 2011.

Z. Urešová, O. Dušek, E. Fučíková, J. Hajič, and J. Šindlerová. Bilingual English-Czech Valency Lexicon Linked to a Parallel Corpus. In *Proceedings of the The 9th Linguistic Annotation Workshop (LAW IX 2015)*, pages 124–128, Stroudsburg, PA, USA, 2015. Association for Computational Linguistics, Association for Computational Linguistics. ISBN 978-1-941643-47-1.

P. Vossen, editor. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA, 1998. ISBN 0-7923-5295-5.

Dirk Weissenborn, Leonhard Hennig, Feiyu Xu, and Hans Uszkoreit. Multi-objective optimization for the joint disambiguation of nouns and named entities. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 596–605, Beijing, China, July 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P15-1058.

M.A. Yosef, J. Hoffart, M. Spaniol, and G. Weikum. AIDA: An online tool for accurate disambiguation of named entities in text and tables. In H. V. Jagadish, José Blakeley, Joseph M. Hellerstein, Nick Koudas, Wolfgang Lehner, Sunita Sarawagi, and Uwe Röhm, editors, *Proceedings of the 37th International Conference on Very Large Data Bases (VLDB 2011)*, Proceedings of the VLDB Endowment, pages 1450–1453, Seattle, USA, 2011. VLDB Endowment.

T. Zhang and D. Johnson. A Robust Risk Minimization Based Named Entity Recognition System. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 204–207, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

Z. Zhong and H. T. Ng. It Makes Sense: A Wide-coverage Word Sense Disambiguation System for Free Text. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 78–83, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.