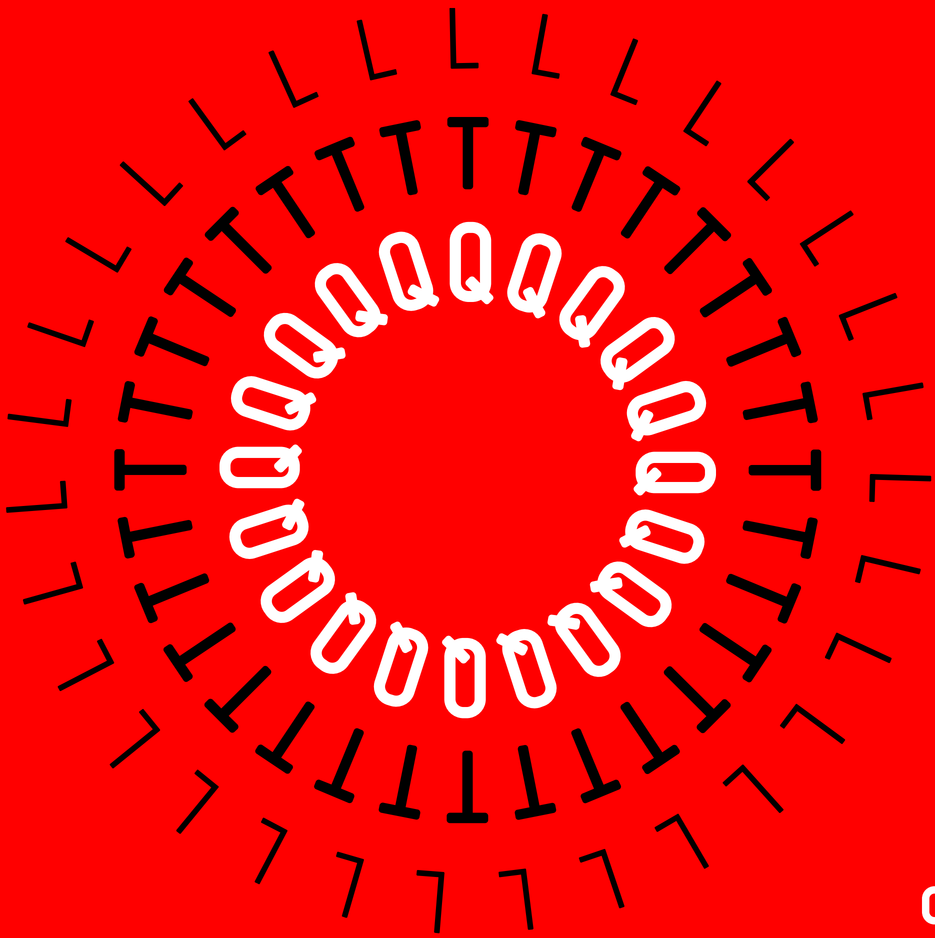

qtleap
quality
translation
by deep
language
engineering
approaches

REPORT ON LANGUAGE RESOURCES AND TOOLS FOR SEMANTIC LINKING AND RESOLVING

DELIVERABLE D5.4

VERSION 2.0 | 2015 JUN 15

QTLeap

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

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

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

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

The deeper the processing of utterances the less language-specific differences remain between therepresentation 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 areopened 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	author	organisation	description
0.1	2014 MAY 20	Gorka Labaka, Eneko Agirre	UPV/EHU	Structure
Meeting	2014 JUNE 19	Eneko Agirre, Gorka Labaka, António Branco, Joao Silva, Martin Popel, Kiril Simov, Petya Osenova	UPV/EHU FCUL CUNI IICT-BAS	
0.2	2014 SEPT 30	Arantxa Otegi, Nora Aranberri, Eneko Agirre	UPV/EHU	Sections 3 and English parts of 4,5,6,7
0.3	2014 OCT 8	João Silva	FCUL	Sections 3.7, 4.6 and 7.6
0.4	2014 OCT 21	Nora Aranberri, Arantxa Otegi, Gorka Labaka	UPV/EHU	All sections (Basque and Spanish)
0.5	2014 OCT 22	Petya Osenova, Kiril Simov	IICT-BAS	Sections 3.4, 4.3, 5.3, 6.2 and 7.3
0.6	2014 OCT 22	Ondřej Bojar, Ondřej Dušek, Michal Novák, Martin Popel	CUNI	Sections 3.6, 4.1.2, 4.5, 7.1.2 and 7.5
1.0	2014 OCT 24	Arantxa Otegi, Eneko Agirre	UPV/EHU	Stable Draft for internal review
Review	2014 OCT 29	Gertjan Jan Van Noord (reviewer)	UG	Improvements suggested by internal review
1.1	2014 OCT 30	Michal Novák	CUNI	Sections 4.5.2 and 7.5.5.3
1.2	2014 OCT 30	Arantxa Otegi, Eneko Agirre	UPV/EHU	Prefinal version
Review	2014 OCT 31	António Branco (Project coordinator)	FCUL	Feedback
1.3	2014 OCT 31	Arantxa Otegi, Eneko Agirre	UPV/EHU	Version first submitted
1.4	2014 NOV 18	Petya Osenova, Kiril Simov, Arantxa Otegi, Eneko Agirre	IICT-BAS, UPV/EHU	Resubmitted version, extended with appendix with the collection of narrative descriptions of LRTs
2.0	2015 JUN 15	Arantxa Otegi , Eneko Agirre, Steve Neal, Petya Osenova, Martin Popel, João Silva, Roman Sudarikov	CUNI FCUL UPV/EHU IICT-BAS	Review comments addressed

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 LANGUAGE RESOURCES AND TOOLS FOR SEMANTIC LINKING AND RESOLVING

DOCUMENT QTLEAP-2014-D5.4
EC FP7 PROJECT #610516

DELIVERABLE D5.4

completion

FINAL

status

SUBMITTED

dissemination level

PUBLIC

responsible

ENEKO AGIRRE (WP5 COORDINATOR)

reviewer

GERTJAN VAN NOORD (UG)

contributing partners

UPV/EHU, FCUL, IICT-BAS, CUNI

authors

ENEKO AGIRRE, NORA ARANBERRI, GORKA LABAKA, ARANTXA OTEGI, STEVE NEAL, JOÃO SILVA, PETYA OSENOVA, KIRIL SIMOV, ONDŘEJ BOJAR, ONDŘEJ DUŠEK, MICHAL NOVÁK, MARTIN POPEL, ROMAN SUDARIKOV

Contents

1 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 data sets and processing tools 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”) and Deliverable 5.1 (”State of the art”).

The present deliverable documents the language resources and tools that compose deliverable D5.3 ”Pilot version of language resources and tools (LRTs) enhanced to support semantic linking and resolving”.

Deliverable D1.3 describes the resources and tools in deliverable D5.3, as follows:

- Datasets for Named Entity Recognition and Classification (NERC)/Named Entity Disambiguation (NED) and Coreference Resolution (CR) for all languages in WP5 (Basque, Bulgarian, Czech, English, Portuguese and Spanish).
- Lexical ontologies aligned, for all languages in WP5.
- Sense annotated corpora, for 2 languages besides English (Bulgarian and Spanish): 100Ktokens aligned, 1Mtokens comparable.
- NERC tools at state of the art performance for all languages in WP5.
- Intrinsic evaluation 1 of Word Sense Disambiguation (WSD), NERC/NED and CR tools, for two languages besides English (Bulgarian and Spanish).

A few of the LRTs in D5.5 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 stored in our internal repository.

Note that English, Spanish and Bulgarian were selected to perform initial development, aimed at preparing subsequent handling of LRTs for all the remaining languages in WP5. The rest of the languages in WP5 (Basque, Czech and Portuguese) need to have all tools available by M16, so they also have the aligned corpora ready for MT Pilot 2.

This deliverable will be followed by D5.6 (due M18) and D5.9 (due M30). D5.6 will extend the NED, WSD and CR to Basque, Czech and Portuguese, and explore crosslingual ambiguity resolution. D5.9 will report on the final versions of the language resources and tools of WP5.

2 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 data sets and processing tools 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). In particular Pilot 2 will be devoted to check the contribution of the tools in this WP to MT.

This deliverable documents the language resources and tools (LRTs) for 6 languages (BG Bulgarian, CS Czech, EN English, ES Spanish, EU Basque, PT Portuguese) that compose deliverable D5.3 "Pilot version of language resources and tools (LRTs) enhanced to support semantic linking and resolving". These LRTs are described in Appendix B, which summarizes the resources. These resources and tools will be used to improve the quality of machine translation.

Deliverable D1.3 "Language resources and tools (LRTs) management plan" describes the resources and tools that belong to deliverable D5.3, as follows:

- Lexical ontologies aligned, for all languages: Section 3 presents our alignment strategy. In the case of English it is based on WordNet, DBpedia and a mapping between WordNet synsets and DBpedia instances. The lexical ontologies for the rest of languages are aligned to either the English WordNet, the English DBpedia or both, as explained in Section 3. The quality of the alignments is reported in Section 7.

- NERC tools at state of the art performance for all languages: Section 4 presents the lemmatization, Part of Speech (PoS) tagging and NERC tools for all languages.

- Intrinsic evaluation 1 of WSD, NERC/NED and CR tools, for two languages besides English: Section 5 presents the WSD, NED and CR tools for English, Spanish and Bulgarian. Section 7 presents the evaluation of those tools for the three languages.

- Sense annotated corpora, for 2 languages besides English: Section 6 presents the corpora which we annotated not only with word senses, but with all available tools for English, Spanish and Bulgarian. Beyond the 100Ktokens from parallel corpora and 1Mtokens from comparable corpora, we have processed 4M tokens from parallel corpora for the pair EN-ES.

- Datasets for NERC/NED and CR for all languages: Section 7 presents the texts that, once annotated, will be used to evaluate the QTLeap tools for NERC, NED and CR. These are reported in the evaluation sections for the corresponding tools. For instance, for English it corresponds to Sections 7.4.3, 7.4.4, 7.5.6 (NERC, NED and CR, respectively), with similar sections for Spanish and Bulgarian. For the rest of the languages, annotated corpora for NERC and texts to be used to evaluate the future tools are described separately. For instance, in the case of Basque, the NERC dataset is covered in Section 7.1.3, and the texts for NED and CR are covered in Section 7.1.4.

Note that Spanish and Bulgarian were the two languages in WP5 selected to perform initial experimental work on LRTs aimed at preparing the subsequent handling of LRTs for all the remaining languages in WP5. All the LRTs for all languages in WP5 follow the management plan for language resources and tools set up in D1.3.

This deliverable is organized as follows. It starts with the Executive Summary and this introduction. The Sub-sections are organized by language. We first present the aligned ontologies for all languages (Section 3). Section 4 presents the basic processing tools for all languages in WP5, including PoS tagging, lemmatization and NERC. The next two sections described LRTs for the three pilot languages. Section 5 describes the WSD, NED and CR tools for English, Spanish and Bulgarian. Section 6 describes the sense-annotated corpora for English, Spanish and Bulgarian. Section 7 reports the pertinent evaluation: aligned ontologies, lemmatization, PoS tagging and NERC for all languages in WP5; NED, WSD and CR for English, Spanish and Bulgarian; evaluation of the tools when applied to domain texts from user scenarios. Section 7 discusses harmonisation issues. Finally, Section 8 presents the conclusions. Appendix A presents the output examples of lemmatizer and PoS tagger for different languages when run on the user scenario texts. Appendix B summarizes the LRTs describes in this deliverable, alongside availability information.

3 Aligned Ontologies

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

3.1 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 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 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

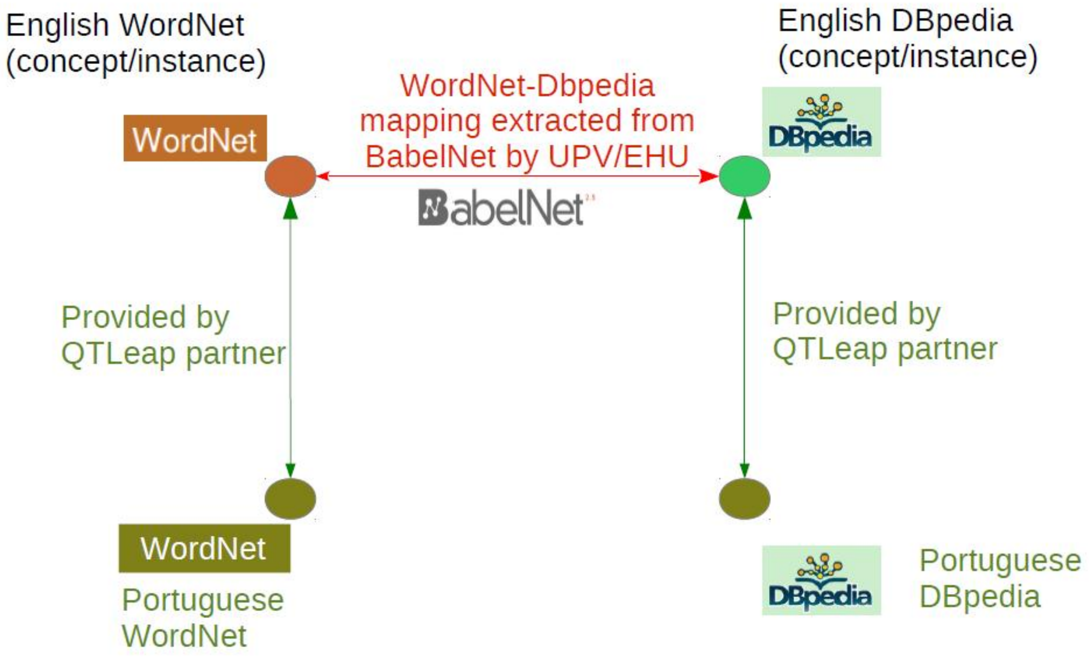

Figure 1: Example figure of the ontology alignment procedure for a sample QTLeap partner for a language (Portuguese shown for illustration). The design for the other languages is analogous.

These resources will be frozen, to allow for comparability alongside project development. Note that the Statistical Machine Translation (SMT) pilots 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. All languages have access to wordnets which are aligned to the English WordNet.

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

All publicly available ontologies and alignment resources are listed in Appendix B.

3.2 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:

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

- 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.

3.3 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:

- WordNet 3.0 contains 4,999 synsets, 6,783 words and 9,056 senses. It covers 100% of the Core WordNet[2].
- 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.

3.4 Czech

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

- Czech BabelNet contains 646 Klemmas, 410 Ksynsets, 897 word senses[3].
- Czech DBpedia contains 225 K 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.

3.5 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

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

[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).

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 DBpedia by automatically acquiring a mapping between WordNet senses and DBpedia 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 publicly available in a text file in the QTLeap repository with the following format:

- WordNet 3.0 URI
- 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 favourably.

3.6 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.

3.7 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:

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

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

- 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.

4 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 Appendix B.

4.1 Basque

4.1.1 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.

4.1.2 NERC

The module ixa-pipe-nerc is multilingual Named Entity Recognition and Classification tagger, and is part of IXA pipes tool (see Section 4.4.1). 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

[8] http://ixa2.si.ehu.es/ixa-pipes/eu/ixa-pipe-pos-eu.tar.gz
[9] http://www.clips.ua.ac.be/conll2 002/ner/
[10] http://www.clips.ua.ac.be/conll2003/ner/

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.

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.

4.2 Bulgarian

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

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

4.2.1 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.

4.2.2 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.

4.3 Czech

4.3.1 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

4.3.2 NERC

NameTag13 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.

4.4 English and Spanish

4.4.1 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.

4.4.1.1 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 for English 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 lookups 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 lookup in WordNet-3.0 (Fellbaum, 1998) via the JWNL API.[19]

Regarding to Spanish, lemmatization is performed via 2 different dictionary lookup 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.

[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/

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

4.4.1.2 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.

4.4.2 Treex

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

4.4.2.1 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.

4.4.2.2 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.

4.5 Portuguese

4.5.1 PoS tagger and lemmatizer

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

[20] https://github.com/ixa-ehu/ixa-pipe-pos
[21] http://www.clips.ua.ac.be/conll2 002/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

LX-Chunker: Marks sentence boundaries with <s>...</s>, and paragraph boundaries with <p>...</p>. Unwraps sentences split over different lines. An f-score of 99.94% was obtained when testing on a 12,000 sentence corpus accurately hand tagged with respect to sentence and paragraph boundaries.

LX-Tokenizer: 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.

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)

This tool achieves a f-score of 99.72% (Branco and Silva, 2003).

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 and an accuracy of 96.87% was obtained (Branco and Silva, 2004).

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

This tool has 91.07% f-score (Branco and Silva, 2006b).

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

This tool has 97.67% f-score (Branco and Silva, 2007).

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, achieving 95.96% accuracy (Branco, Nunes and Silva, 2006).

4.5.2 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. It scored 85.19% precision and 85.91% recall. The name-based component is based on Hidden Markov Models technology and was trained over a manually annotated corpus of approximately 208,000 words. When evaluated against an unseen portion with approximately 52,000 words, it scored 86.53% precision and 84.94% recall (Ferreira, Balsa and Branco, 2007).

5 NED, WSD and Coreference tools

This section describes the NED, WSD and CR tools for the languages that were selected to act as pilots in the current phase of the WP5 activities, namely Bulgarian, English and Spanish. Similar tools for the rest of languages in WP5 are due M16. Tools for NED, WSD and CR will be ready in time to annotate the aligned corpora for MT Pilot 2 for all languages in WP5.

All these tools are listed in Appendix B.

5.1 Bulgarian

We have adopted two approaches in the project. 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.

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

5.1.1 NED

The annotation follows the same approach as the disambiguation module of the Bulgarian pipeline (see 4.2.2) but here the DBpedia classes are used. DBpedia's ontological hierarchy determines the more general categories for DBpedia instances (City, Politician, etc.) as subclasses of Person, Location and Organization. For other kinds of instances we rely on the most general category provided by the classification of the instance according to DBpedia. Then the standard module is adapted to use the new categories. In case the selected categories in the annotation are not sufficient for disambiguating among DBpedia instance URIs, we store all of them in the annotation.

It is an unfortunate fact that DBpedia Spotlight[26] does not support Bulgarian.

The input is the result from the PoS tagger and the lemmatizer for Bulgarian.

The output is converted to NAF similar to English and Spanish modules, presented above.

5.1.2 WSD

The basic version of WSD is implemented on the assumption of one sense per discourse and bigram statistics. In the next phase of the project more advanced system will be

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

implemented using additional semantic resources like ontologies, base concepts of WordNet as well as syntactic structure of the sentences.

5.1.3 Coreference

We have implemented a basic version of a coreference resolution module, using paths in the dependency tree of each sentence. By using path patterns, we are mainly performing 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 and we group them together. For more advanced processing we plan on exploiting the RelaxCor system[27].

RelaxCor solves cofererence resolution in several steps, which include mention detection (detection of possible coreferents, such as noun phrases, named entities, pronouns, etc.), generation of feature vectors for each mention pair (for instance, morphological features for gender and number agreement), application of a set of constraints to the pairs, and detection of coreferences using relaxation labelling over a weighted graph with the mentions as nodes. An edge weight is the sum of the weights of the constraints that apply to that mention pair. A feature vector of over a hundred binary features is defined for each pair by extracting information from the preprocessed input.

The input to RelaxCor is preprocessed data in the CoNLL format used for the Semeval 2010 task. Besides the standard CoNLL columns, it includes information about named entities and predicates. The minimum information required by RelaxCor is tokenization, part-of-speech tag, and dependency parsing, while named entities are optional, but beneficial for good performance.

5.2 English and Spanish

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

5.2.1 NED

The ixa-pipe-ned module performs the Named Entity Disambiguation task based on DBpedia Spotlight[28]. 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[29]. The tool has been developed independently from QTLeap.

5.2.2 WSD

UKB is a collection of programs for performing graph-based Word Sense Disambiguation[30]. 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.

[27] http://nlp.lsi.upc.edu/relaxcor/

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

[29] https://github.com/ixa-ehu/ixa-pipe-ned

[30] 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[31]. The tool has been developed independently from QTLeap.

5.2.3 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 favours 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; Lee et al., 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)[32]. The tool has been developed independently from QTLeap.

6 Annotated corpora

This section describes the corpora which have been automatically annotated with the tools mentioned in the previous sections for Bulgarian, English and Spanish. The corpora for the rest of languages are due in M18.

Given the availability of large parallel corpora, we decided to go beyond the 100K tokens planned in the DoW to be annotated from parallel corpora targeted at D1.3. We have processed 4M tokens from parallel corpora for EN-ES and 500K tokens from parallel corpora for EN-BG. Both parallel corpora come from Europarl. Fortunately, the overlap of English sentences between the EN-BG and EN-ES corpora is very high, that is, 93% of the sentences in the English part of EN-ES are also present in the BG-EN corpus.

Regarding the 1M tokens from comparable corpora planned in the DoW and targeted in D1.3, the parallel corpora provides data of better quality in larger numbers, so we decided to focus on parallel corpora alone.

Given the analysis of the output of the processors in the specific domain of QTLeap, we also decided to start checking comparable corpora from the same domain. The Bulgarian team is making a first step in this direction, having gathered and annotated comparable corpora on the target domain automatically extracted from Wikipedia.

All the annotated corpora is listed in Appendix B.

[31] http://ixa2.si.ehu.eus/ixa-pipes/eu/ixa-pipe-wsd-ukb.tar.gz

[32] https://bitbucket.org/Josu/corefgraph

6.1 Bulgarian-English

- Parallel corpus (100K tokens of, SETIMES are PoS tagged and Dependency parsed) Processed for D5.3 (582,376 tokens from SETIMES, 24561 sentences). English part with the pipeline of Newsreader European project which is using the tools provided by UPV/EHU. This corpus is entirely developed within the project. It is distributed under the license CC-BY-NC-SA 4.0.

language	EN	BG
tokens	578,405	582,376
terms	578,405	582,376
linked to WordNet	227,370 (39.04%)	142,638 (24.5%)
entities	43,077	57,585
linked to DBpedia	36,379 (84.45%)	22,935 (39.8%)
coreference chains	26,039	39,637

Table 1: Statistics on SETIMES annotated parallel corpora (Bulgarian-English)

- Comparable corpus (1 Mtokens)
We have extracted interlinked English-Bulgarian Wikipedia articles. The total number of aligned articles is more than 36000. From them about 3000 are in the technical domain and to some extent are related to the domain of the real user scenario. The data has been processed. Additionally, we plan to process domain parallel corpora. This corpus is entirely developed within the project. It is distributed under the license CC-BY-NC-SA 4.0.

language	EN	BG
tokens	1,997,667	887,968
terms	1,997,667	887,968
linked to WordNet	781,723 (39.1%)	182,985 (20.6%)
entities	98,021	52,903
linked to DBpedia	83,914 (85.6%)	18,544 (35.1%)
coreference chains	68,455	39,519

Table 2: Statistics on annotated comparable domain corpora (Bulgarian-English)

- Monolingual Bulgarian resources
Since there are not freely available resources for Bulgarian to support semantic annotation with senses and instance identifiers, we have created our own resources on the base of Bulgarian Treebank. We have annotated all open class words in the treebank with appropriate senses. Where possible, we selected the senses from two resources: (1) Bulgarian WordNet, mentioned above, and (2) Definitions from a machine readable dictionary. In many cases the annotators added their own definitions. The already completed part covers 78 308 words (annotated by two annotators). There are about 30000 more cases annotated by just one annotator. The selected senses that are not in WordNet are being added

and mapped to the English WordNet. Bulgarian WordNet was partially extented within the project. It is distributed under the license CC BY 3.0.

We have annotated also the Treebank with URIs of DBpedia instances. The number of the annotated named entities is 10855. This dataset was entirely developed within the project. It is distributed under the license CC-BY-NC-SA 4.0.

Both resources will be used for training of more advanced tools for named entities.

6.2 Spanish-English

- Parallel corpus (100 Ktokens):
  - Europarl-QTLeap WSD/NED corpus: Processed 4M tokens for D5.3 of Europarl v7.0 parallel corpus. The corpus is distributed under the license CC BY 4.0, and has been released through meta-share[33] and CLARIN Lindat[34].

language	EN	ES
tokens	4,244,573	4,351,530
terms	4,244,573	4,351,530
linked to WordNet	1,858,851 (43.79%)	1,525,516 (35.06%)
entities	146,202	176,671
linked to DBpedia	133,880 (92.57%)	144,859 (81.99%)
coreference chains	142,799	76,043

Table 3: Statistics on annotated Europarl-QTLeap WSD/NED corpus (Spanish-English)

  - QTLeap WSD/NED corpus: batch 1 and 2. The corpus is distributed under the license BY-NC-SA 4.0, and has been released through meta-share[35] and CLARIN Lindat[36].

language	EN	ES
tokens	67,081	70,037
terms	67,081	70,037
linked to WordNet	25,069 (37.37%)	21,210 (30.28%)
entities	1,893	5,204
linked to DBpedia	1,445 (76.33%)	3,210 (61.68%)
coreference chains	2,370	774

Table 4: Statistics on annotated QTLeap WSD/NED corpus (Spanish-English)

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

[34] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1477

[35] http://metashare.metanet4u.eu/go2/qtleap-wsdned-corpus

[36] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1476

- Comparable corpus (1 Mtokens)
  We used 4Mtokens from parallel corpus instead, as it provides better quality translations. We plan to extract interlinked English-Spanish Wikipedia articles.

7 Evaluation

In this section we report on the evaluation of all tools and resources mentioned in this deliverable. 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 Appendix B). In the case of aligned resources, we provide a qualitative statement. Note that according to the planning in the Dow, for the rest of the languages in WP5, that is Basque, Czech and Portuguese, only basic tools are due to be reported at this milestone.

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) for each one of the three pilot languages, that is Bulgarian, English and Spanish. In the next phase of WP5 activities, we will extend the study to the discrepancies between the results of the tools among language pairs, specially between English and each of the QTLeap languages. Some early conclusions have been drawn in the conclusion section.

7.1 Basque

7.1.1 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.

7.1.2 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%.

7.1.3 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.

7.1.4 Datasets for NED/WSD and coreference

The evaluation of NED/WSD and coreference for Basque is due M21. At this stage, we uploaded to the repository the NED and CR corpora that will be used for testing in the future, as well as the WSD corpora.

Since there is no standard Basque corpus defined for the NED evaluation task, we have generated a repository for that purpose using pieces of news of the 2002 year edition of the *Euskaldunon Egunkaria* newspaper. In order to build the test-corpus, we collected news paragraphs with at least one entity. For each NE in this example set, the corpus was manually disambiguated, linking each NE occurrence to its corresponding DBpedia entry, when possible. This test corpus was divided into two groups in order to use one for the tuning process (Corpus dev) and the second one for evaluation (Corpus eval). Final version on the corpus consists on 1032 entities (532 on Corpus dev and 500 on Corpus eval).

The test corpus used to evaluate correference is a subpart of EPEC corpus consisting of 46,383 words that correspond to 12,792 mentions. First of all, automatically tagged mentions obtained by our mention detection system (Soraluze et al., 2012) have been corrected; then, coreferent mentions have been linked in clusters. This work has been carried out using the MMAX2 annotation tool (Müller and Strube, 2006).

EuSemcor is the Basque semantic concordance (Basque Semcor), comprising a set of occurrences of nouns in the Basque EPEC corpus which has been annotated with Basque WordNet V1.6 senses. It will be used to test the performance of the WSD tools for Basque. The corpus contains 42,615 occurrences of nouns annotated by hand, corresponding to the 407 most frequent Basque nouns, in XML format. The release was produced in 2008, as reported in Pociello et al. (2011). It is not freely available yet, although it can be checked online in http://ixa2.si.ehu.es/eusemcor/.

7.1.5 Domain evaluation

7.1.5.1 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.

7.1.5.2 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.

7.1.5.3 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 5). 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	Entity
171	Windows
119	Wi-Fi
98	Google
78	Skype
45	Gmail
42	internet
39	Facebook
39	Android
38	Dropbox
34	Word

Table 5: 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 6). 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 6: 20 most frequent entities with class for Basque

7.2 Bulgarian

7.2.1 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 Bulgarian WordNet as a new synset and then the new synset is aligned to English WordNet. The alignment between Bulgarian DBpedia and English DBpedia is provided within DBpedia itself. The entities missing in DBpedia that were created on the basis of Wikipedia are also checked by two people.

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

7.2.2 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 and new words.

7.2.3 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 %. 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.

7.2.4 NED

We have reused the same data for measurement as in the case of NERC. The gold standard annotations of DBpedia instances are 667. The automatic procedure annotated 391 instances. The intersection of the annotated instances is 248. Thus the precision of the tool is 63.43 % and the recall is 37.18 %. The low results are due to the small coverage of the Bulgarian DBpedia. In order to solve this problem in the next phase of the project we plan to extend the coverage of the Bulgarian DBpedia in the following ways: (1) using Bulgarian Wikipedia articles that are not in the Bulgarian DBpedia but have linked corresponding instances in the English DBpedia. In this case we will automatically transfer the ontological classification from the English DBpedia to the new Bulgarian instances; (2) using transliteration rules, we will transliterate English instance names into Bulgarian ones. In the first case we will be able to refer to both Bulgarian and English Wikipedia articles. In the second case we will be able to refer only to English ones. The second approach could possibly introduce errors due to cases of wrong transliteration or ambiguous Bulgarian names.

7.2.5 WSD

Again, we have reused the same data for measurement as in the case of NERC. The gold standard sense annotations are 3118. The automatic procedure annotated 2727 cases. The annotations in common are 1925. Precision is 70.6 % and recall is 61.7 %. The result is relatively good, bearing in mind the limited size of the Bulgarian WordNet, which was used in the annotation. We plan on improving the result by extending the coverage of the WordNet and by exploiting a better tool for WSD.

7.2.6 Coreference

The same corpus was used for evaluating the module for coreference resolution. The human-annotated coreference chains are 337. The automatic annotation yielded 53 chains, a difference that is too large. One problem was identified as the source of this apparent bad performance: the automatic procedure selected coreference chains that were too long, because they extended beyond the boundaries of individual texts. In order to overcome this problem, we constructed lists of the coreferent words in each chain and used those to calculate performance. The gold standard annotations contain 903 related words. The automatic procedure returns 563 related words, out of which 371 match the gold standard data. Measured in this way, precision is 65.9 and recall is 41.1 %. We hope we can achieve better results by exploiting the RelaxCor system.

7.2.7 Domain evaluation

The evaluation of 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 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 "кликам" (to click). Such words have mainly 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).

7.3 Czech

7.3.1 Aligned resources

The link between the Czech and English DBpedias 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.

7.3.2 Lemmatization and PoS tagging

Czech has standard resources with manual morphological annotation, i.a. the Prague Dependency Treebank[37]. 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.

7.3.3 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[38] is 80.30% for the coarse-grained 7-classes classification and 77.22% for the fine-grained 42-classes classification (Straková et al. 2014).

7.3.4 Datasets for NED/WSD and coreference

The Prague Dependency Treebank (PDT) lends itself well also to the task of WSD. All verbs and some nouns in PDT are annotated with valency frames from the Valency Lexicon of Czech Verbs (Vallex)[39]. Valency frames correspond to senses. The dataset was already used for a shared task in WSD, namely CONLL 2009 SRL Joint Task.

PDT has been also exploited for coreference resolution. Particularly, it served as training data for a system that targets personal pronouns in 3rd person introduced by Nguy et al. (2009) and re-implemented by Bojar et al. (2012). It achieves 50% in pairwise F-score measured on the evaluation part of PDT. In addition, several rules have been designed to cover coreference of relative and reflexive pronouns in Czech.

CUNI addressed the task of NED in (Hálek et al., 2011) where we mined Wikipedia for translations of named entities into a morphologically rich language (namely Czech). The work is currently not used by CUNI's MT systems but can be applied with some small effort.

[37] http://ufal.mff.cuni.cz/pdt3.0

[38] http://ufal.mff.cuni.cz/cnec

[39] http://ufal.mff.cuni.cz/vallex

7.3.5 Domain evaluation

7.3.5.1 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 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.2).

7.3.5.2 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.2, 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).

7.3.5.3 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 7 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.

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

Table 7: 10 most frequent entities for Czech

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 8. As mentioned in Section 4.5.2, 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 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.

7.4 English

7.4.1 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 use string matching between the names of the articles, as the automatic construction of DBpedia ensures a one-to-one mapping.

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 8: 20 most frequent entities with class for Czech

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.

7.4.2 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).

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

7.4.3 NERC

The *ixa-pipe-nerc* module based on the CONLL 2002[40] and 2003[41], 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.

7.4.4 NED

For the evaluation of the ixa-pipe-ned module, we used the 2010 and 2011 datasets from the TAC KBP editions[42] and the AIDA corpus[43]. 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 parameters, it was optimized in 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.

7.4.5 WSD

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. The overall result obtained is F1 80.1. An analysis of the performance according to the PoS shows that this module performs better particularly on nouns, obtaining F1 83.6 (results for the rest of PoS: 71.1 for verbs, 83.1 for adjectives and 82.3 for adverbs).

7.4.6 Coreference

The ixa-pipe-coref module has been evaluated on the development auto section of the CoNLL 2011 shared evaluation task[44] 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. The difference could be because their system's CoNLL adaptation and post-processing is better. For example, Stanford recognizes the speaker (speaker sieve) which ixa-pipe-coref does not.

7.4.7 Domain evaluation

7.4.7.1 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.3, 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

[40] http://www.clips.ua.ac.be/conll2002/ner/

[41] http://www.clips.ua.ac.be/conll2003/ner/

[42] 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/

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

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

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”.

7.4.7.2 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.3), 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.

7.4.7.3 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 9 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	Entity
90	Windows
84	Facebook
65	Google
54	PC
31	USB
30	Google Chrome
29	Google Drive
24	Internet
21	Skype
14	YouTube

Table 9: 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 10). 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.

7.4.7.4 NED

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.

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 10: 20 most frequent entities with class for English

7.4.7.5 WSD

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”.

7.4.7.6 Coreference

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

From this first pilot, we have learned that the user scenario text needs to be processed per interaction, that is, each user-machine interaction should be processed separately for coreference annotation.

7.5 Portuguese

7.5.1 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.

7.5.2 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 take 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, Nunes and Silva, 2006).

7.5.3 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. 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 (Ferreira, Balsa and Branco, 2007).

7.5.4 Datasets for NED/WSD and coreference

The corpus used for NERC has been uploaded to the repository. This corpus is composed of 30,509 sentences (688,962 tokens) taken from CINTIL Corpus. The corpus uploaded to the repository contains a subset of the linguistic information in CINTIL Corpus, namely PoS annotation and information on named-entities.

The Portuguese portion of the MultiWordNet ontology has been uploaded to the repository. This ontology will be used to support the task of WSD. It comprises 17,200 manually validated concepts/synsets in the subontologies under the concepts of Person, Organization, Event, Location, and Artworks.

Several Portuguese corpora with coreference information have been gathered and uploaded to the repository. The Summ-it corpus (Collovini et. al., 2007) consists of 50 texts taken from the Brazilian newspaper Folha de São Paulo. Among its layers of annotation, it includes semi-automatic annotation of co-reference information in MMAX format (Muller and Strube, 2001). The Collovini corpus was developed by Collovini (2005). It consists of 24 texts taken from the Brazilian newspaper Folha de São Paulo and, and like the Summ-it corpus, it includes a layer of co-reference information in MMAX format. The LinkPeople corpus was developed by Garcia and Gamallo (2014). It consists of 97 documents taken from newspapers and Wikipedia. The co-reference annotation follows the SemEval-2010 Task #1 format (Recasens et al., 2010).

7.5.5 Domain evaluation

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

7.5.5.1 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.4. 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".

7.5.5.2 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.4. 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.

7.5.5.3 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 11 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 12 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	Entity
98	Facebook
73	Word 2013
66	PowerPoint 2013
59	Windows
39	Skype
38	Mac
35	Excel 2013
29	PC
29	Android
28	Chrome

Table 11: 10 most frequent entities for Portuguese

7.6 Spanish

7.6.1 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.

7.6.2 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 et al., 2004) report 98.86%, although they train and test on a different subset of the Ancora corpus.

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 12: 20 most frequent entities with class for Portuguese

7.6.3 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.

7.6.4 NED

The Spanish ixa-pipe-ned module has been evaluated on the TAC 2012 Spanish dataset[45]. 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 NED Spanish over the TAC 2012 Spanish dataset, which outputs entities from Spanish DBpedia. We then map those entities to the corresponding English counterparts using the

[45] 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/

interlingual links from Wikipedia[46]. We tried the same best set of parameters as used for the English experiments in the evaluation dataset (TAC 2010). We obtained a performance of 78.15 in precision and 55.80 in recall.

7.6.5 WSD

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) and obtained F1 79.3.

7.6.6 Coreference

ixa-pipe-coref Spanish module has been evaluated on the publicly available datasets distributed by the SemEval 2010 task on Multilingual Coreference resolution, in which the AnCora-ES (the Spanish part) corpus is used. These are the results we obtained on the closed gold type of evaluation for different F1 metrics: 70.67 CEAFm F1, 43.58 MUC F1, 75.94 B3 F1 and 61.42 BLANC F1.

7.6.7 Domain evaluation

7.6.7.1 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.5, 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.

7.6.7.2 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.5), 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.

7.6.7.3 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 13 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.

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

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

Table 13: 10 most frequent entities for Spanish

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 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 14). 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.

7.6.7.4 NED

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/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 botanics http://es.dbpedia.org/resource/Haz_(botánica).

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 14: 20 most frequent entities with class for Spanish

7.6.7.5 WSD

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)".

7.6.7.6 Coreference

The same comments as English apply here (cf. Section 7.4.7.6).

8 Harmonisation

The 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 harmonising the output of linguistic tools, which is reused in this project, as follows:

- PoS tags and syntactic tags: HamleDT[47] provides harmonised treebanks for all project languages.
- NERC tags: all languages and annotations schemes provide three common tags, person, location and organization.
- NED and WSD: the alignment strategy for ontologies set up in Section 3.

9 Conclusions and future work

This deliverable reports on the LRTs that compose deliverable D5.3 "Pilot version of language resources and tools (LRTs) enhanced to support semantic linking and resolving".

With respect to all the six languages in WP5, the partners have prepared basic processing tools, namely PoS taggers, lemmatizers and NERC modules. Those tools are on the state-of-the-art when compared to freely available NLP pipelines

As planned in the DoW, and further detailed in Deliverable 1.3., the consortium gathered and/or produced an extensive array of basic resources and tools for the six languages in WP5 (Basque, Bulgarian, Czech, English, Portuguese and Spanish). With respect to the three languages used for pilot work, Bulgarian, English and Spanish, tools for NED, WSD and Coreference have been put in place, which were evaluated in standard datasets and in datasets belonging to the domain of the real usage scenario assumed in the project. These tools have shown performance and the level of the state of the art .

A large corpus of EN-BG and EN-ES has been annotated with word senses, as well as all the basic and advanced tools. In the case of EN-ES the corpus is a subset of 4 Million tokens from Europarl, and for EN-BG around 2 Million tokens from Wikipedia and SETIMES corpora.

In addition to the goals set in D1.3, we have also performed in-domain evaluation, analyzing the quality of the output of the tools when applied over domain texts.

In the next phase of WP5 activities, the analysis of the domain results will be very useful to improve all tools and resources. This domain analysis will be extended to the full set of

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

tools. In addition, we will analyze the sense-annotated corpora. This analysis will inform the following tasks:

- Harmonization of the output of the tools across languages. As the in-domain analysis has shown, some of the tools provide differing results across languages: for instance, the number and type of named-entities across languages are different (Faceboook is detected 81 times in Spanish, only 39 times in the Basque translations). The harmonization for the two languages in a pair should produce more consistent results which could lead to better SMT results.
- Adaptation of some of the tools to the domain. For instance, some terms like PC and USB are detected as named-entities.
- Improve the state-of-the-art on NED, WSD and CR using crosslingual techniques.
- Design of the strategies to improve the quality of translation in Pilot 2 at M24.

References

Aduriz I., Aranzabe M., Arriola J., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., Urizar R. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. Corpus Linguistics Around the World. Book series: Language and Computers. Vol 56 (pag 1- 15). Ed. Andrew Wilson, Paul Rayson, and Dawn Archer. Rodopi. Netherlands. 2006

Agirre E., Lopez de Lacalle O., Soroa A. Random Walks for Knowledge-Based Word Sensce Disambiguation. Computational Linguistics, 40:1. 2014.

Alegria I., Aranzabe M., Ezeiza A., Ezeiza N., Urizar R. Robustness and customisation in an analyser/lemmatiser for Basque. LREC-2002 Customizing knowledge in NLP applications Workshop. 2002

Alegria I., Arregi O., Ezeiza N., Fernandez I. Lessons from the Development of a Named Entity Recognizer Procesamiento del Lenguaje Natural, 36, 25-37. 2006

Barrena A., Agirre E., Soroa A. UBC Entity Linking at TAC-KBP 2013: random forests for high accuracy. Text Analysis Conference, Knowledge Base Population 2013

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

Branco A., Silva J. Contractions: breaking the tokenization-tagging circularity. Lecture Notes in Artificial Intelligence 2721, pages 167—170, Spinger. 2003.

Branco A., Silva J. Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). 2004. ELRA, pages 507—510. 2004.

Branco A., Silva J. Dedicated Nominal Featurization of Portuguese. Lecture Notes in Artificial Intelligence 3960. Springer. 2006.

Branco A., Silva J. A Suite of Shallow Processing Tools for Portuguese: LX-Suite. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06). 2006b.

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

Bojar O., Žabokrtský Z., Dušek O., Galuščáková P., Majliš M., Mareček D., Maršík J., Novák M., Popel M., Tamchyna A. The Joy of Parallelism with CzEng 1.0. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), İstanbul, Turkey, ISBN 978-2-9517408-7-7, pages. 3921-3928, 2012

Carreras X., Màrquez L., Padró L. 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. Association for Computational Linguistics. 2002.

Collins M. Discriminative training methods for Hidden Markov Models: Theory and experiments with Perceptron algorithms. In Proceedins of the ACL-02 conference on Empirical methods in Natural Language Processing, volume 10, pages 1-8. 2002.

Collovini, S. Análise de Expressões Referenciais em Corpus Anotado da Língua Portuguesa. MSc Thesis, Universidade do Vale do Rio dos Sinos. 2005.

Collovini, S., Carbonel T.I., Thiesen Fuchs J., Coelho J.C., Rino L., Vieira R. Summ-it: Um Corpus Anotado com Informações Discursivas Visando à Sumarização Automática. In Anais do XXVII Congresso da Sociedade Brasileira de Computação – V Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2007), pages 1605-1614. 2007.

Daiber J., Jakob M., Hokamp C., Mendes P. Improving Efficiency and Accuracy in Multilingual Entity Extraction. Proceedings of the 9th International Conference on Semantic Systems (I-Semantics). 2013.

Fellbaum C., editor. WordNet. An Electronic Lexical Database. The MIT Press, 1998.

Fernando S., Stevenson M. Mapping WordNet synsets to Wikipedia articles In proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), May 23-25 Istanbul, Turkey. 2012.

Ferreira E., Balsa J., Branco A. Combining Rule-based and Statistical Methods for Named Entity Recognition in Portuguese. TIL2007 - V Workshop em tecnologia da Informaccao e da Linguagem Humana, Anais do XXVII Congresso da Sociedade Brasileira de Computac cao, 1615—1624. 2007.

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

Garcia, M., Gamallo P. Multilingual Corpora with Coreference Annotation of Person Entities. In Proceedings of the Language Resources and Evaluation Conference, pages 3229-3233. 2014.

Georgi G., Zhikov V., Osenova P., Simov K., Nakov P. Feature-rich part-of-speech tagging for morphologically complex languages: Applica-tion to Bulgarian. In Proceedings of the 13th Conference of the European Chapter ofthe Association for Computational Linguistics, EACL ’12, pages 492–502, Stroudsburg, PA, USA, 2012.

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

Gonzalez-Agirre A., Laparra E. and Rigau G. Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base . In Proceedings of the Sixth International Global WordNet Conference (GWC’12). Matsue, Japan. 2012.

Hálek O., Rudolf R., Tamchyna A., Bojar O. Named Entities from Wikipedia for Machine Translation. In: Information Technologies – Applications and Theory, Copyright ©

Univerzita Pavla Jozefa Šafárika v Košiciach, Košice, Slovakia, ISBN 978-80-89557-02-8, ISSN 1613-0073, pp. 23-30, 2011.

Lee H., Peirsman Y., Chang A., Chambers N., Surdeanu M., Jurafsky D. 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, page 28–34. 2011.

Lee H., Chang A., Peirsman Y., Chambers N., Surdeanu M., Jurafsky D. Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics 39(4). 2013.

Màrquez, L., Villarejo M.A., Martí T., Taulé M. 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.

Navigli R., Litkowski K.C., Hargraves O. SemEval-2007 Task 07: Coarse-Grained English All-Words Task. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007) in conjunction with ACL, pages 30–35, Prague, Czech Republic. 2007.

Müller, C., Strube M. MMAX: A Tool for the Annotation of Multi-modal Corpora. In Proceedings of the 17th International Joint Conference on Artificial Intelligence, pages 45-50. 2001.

Müller C. and Strube M. Multi-level annotation of linguistic data with MMAX2. Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods. 2006

Navigli R., Ponzetto S. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier, pp. 217-250. 2012.

Nguy G., Novák Václav, Žabokrtský Zdeněk: Comparison of Classification and Ranking Approaches to Pronominal Anaphora Resolution in Czech. In: Proceedings of the SIGDIAL 2009 Conference, Copyright © The Association for Computational Linguistics, London, UK, ISBN 978-1-932432-64-0, pp. 276-285, 2009.

Pociello, E., Agirre, E. and Aldezabal. Methodology and construction of the Basque WordNet. Language Resources and Evaluation, Volume 45, Issue 2, pp 121-142, 2011.

Raghunathan K., Lee H., Rangarajan S., Chambers N., Surdeanu M., Jurafsky D., Manning C. A multi-pass sieve for coreference resolution. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, page 492–501. 2010.

Ratnaparkhi A. Learning to Parse Natural Language with Maximum Entropy Models. Machine Learning, 34, 151-175. 1999.

Simov K., Osenova P. A hybrid system for morphosyntactic disambiguation in Bulgarian. In: Proc. of the RANLP 2001 Conference, Tzigov chark, pages 5–7, 2001.

Recasens, M., Màrquez L., Sapena E., Martí M.A., Taulé M., Hoste V., Poesio M., Versley Y. SemEval-2010 Task 1: Coreference Resolution in Multiple Languages. In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10), pages 1-8. 2010.

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

Straková J., Straka M., Hajič J. A New State-of-The-Art Czech Named Entity Recognizer. Text, Speech, and Dialogue, eds. Habernal I., Matoušek V., Lecture Notes in Computer Science, vol. 8082, pages 68-75. 2013.

Straková J., Straka M., 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.

Toutanova K., Klein D., Manning C., Singer Y. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL, pages 252–259. 2003.

Zhang T, Johnson D. 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, pages 204-207. 2003.

Appendix A: Examples of annotations

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

9.1 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)

9.2 A.2 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

```
	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:-------------"

9.3 A.3 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="."

9.4 A.4 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)

9.5 A.5 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"

Appendix B: Summary of availability

Name of LRT	Language(s)	QTLeap	License	URL
Datasets				
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 Non-Members	https://catalog.ldc.upenn.edu/LDC2014T16
TAC 2012 (NED)	ES	No	Restricted to registered TAC 2012	http://www.nist.gov/tac/2012/KBP/data.html

			participants	
WSJ Treebank (Lemma./PoS)	EN	No	*Check with authors*	http://www.cis.upenn.edu/~treebank/
Ontologies				
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
Annotated corpora				
Europarl-QTLeap WDS/NED corpus	EN, ES	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 WDS/NED corpus	BG, EN, ES	Yes	CC-BY-NC-SA 4.0	http://metashare.metanet4u.eu/go2/qtleap-wsdned-corpus https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1476
SETIMES corpus	BG, EN	Yes	CC-BY-NC-SA 4.0	http://www.bultreebank.org/QTLeap/ , http://www.bultreebank.org/EMP/
Wikipedia corpus	BG, EN	Yes	CC-BY-NC-SA 4.0	http://www.bultreebank.org/QTLeap/

Processing tools				
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
MorphoDita	CS, EN	No	CC BY-NC-SA 3.0	http://ufal.mff.cuni.cz/morphodita
NameTag	CS, EN	No	CC BY-NC-SA 3.0	http://ufal.mff.cuni.cz/nametag

Tabla 15: Summary of publicly available LRTs mentioned in D5.4. QTLeap column for those LRTs which have been (partially) funded by QTLeap.