Deliverable D4.2
Linguistically annotated parallel corpora, using a variety of adapted NLP tools, across a range of languages and domains

7th March, 2011

Grant Agreement number: 247762

Project acronym: FAUST

Project title: Feedback Analysis for User Adaptive Statistical Translation

Funding Scheme: FP7-ICT-2009-4 STREP

Name, title and organisation of the scientific representative of the project's coordinator:
Dr William Byrne, Reader in Information Engineering
Department of Engineering, University of Cambridge, UK

Tel: +44 (0)1223 332651

Fax: +44 (0)1223 332662

E-mail: bill.byrne@eng.cam.ac.uk

Project website address: `http://faust-fp7.eu`

Contents

1 Overview

This document reports on the linguistic annotation of static text collections which will be used for the development of MT systems and evaluation methods in the FAUST project. This work is being carried out within Task T4.2 with objectives summarised as follows (taken from project Annex I, Description of Work):

T4.2 Robust linguistic annotation of static text collections for a range of languages and domains.

> Tools developed in T4.1 will be applied to static monolingual and parallel text collections used to build MT systems and evaluation methods. This is a static, off-line task. Dependencies on other tasks: T3.3

We have performed linguistic annotation of Catalan, Czech, English, French, Romanian and Spanish text. The level of processing varies among languages, and may include tokenization, part-of-speech tagging, lemmatization, base phrase chunking, dependency parsing, constituency parsing, named entity recogition, semantic role labeling and discourse analysis.

In the following sections, we list the linguistic processors used (Section 2), annotated corpora (Section 3), data sets (Section 4), data representation formats (Section 5), the progress of annotation as of February 2011 (Section 6) and the final list of file for public release (see Appendix A).

2 Linguistic Processors

Several tools[1] have been used for the linguistic annotation of corpora:

- TectoMT
- SVMTool
- ZPar
- BIOS
- MALT Parser
- C&C Tools

[1] `http://www.faust-fp7.eu/faust/Main/PublicTools`

- SwiRL
- JointParser

In the following, we provide a brief description of these tools and their role in the annotation pipeline in this project.

2.1 TectoMT

TectoMT [Pv10] is a multi-purpose open-source NLP framework[2]. It allows for fast and efficient development of NLP applications by exploiting a wide range of software modules already integrated in TectoMT, such as tools for sentence segmentation, tokenization, morphological analysis, POS tagging, shallow and deep syntax parsing, named entity recognition, anaphora resolution, tree-to-tree translation, natural language generation, word-level alignment of parallel corpora, and other tasks. One of the most complex applications of TectoMT is the English-Czech machine translation system with transfer on deep syntactic (tectogrammatical) layer. Several modules are available also for other languages (German, Russian, Arabic). Where possible, modules are implemented in a language-independent way, so they can be reused in many applications.

We use TectoMT for the syntactic analysis of Czech and English. For both languages we used integrated tagger Morce [SHV$^+$07] and McDonald's maximum spanning tree parser [MPRH05].

2.2 SVMTool

The SVMTool [GM04a, GM04b] is a simple and effective generator of sequential taggers based on Support Vector Machines[3]. The SVMTool has been successfully applied to the problem of part-of-speech (PoS) tagging, achieving a very competitive accuracy of 97.2% for English on the Wall Street Journal corpus, which is comparable to the best taggers reported up to date. The SVMTool is robust and flexible for feature modelling (including lexicalization), trains efficiently with almost no parameters to tune, and is able to tag thousands of words per second, which makes it really practical for real NLP applications.

We use the SVMTool for the PoS tagging and lemmatization of Catalan, English, French, Romanian[4] and Spanish.

[2] `http://ufal.mff.cuni.cz/tectomt/`

[3] `http://www.lsi.upc.edu/~nlp/SVMTool/`

[4] No lemmary was available for Romanian.

2.3 ZPar

ZPar [ZC11] is a statistical natural language parser, currently supporting multiple languages, including Chinese, English and Romanian[5]. ZPar has specific language support for Chinese and English, while all other languages are currently implemented with a generic method. ZPar does word segmentation, part-of-speech tagging and parsing in constituent and dependency grammars.

We have used ZPar for the syntactic parsing of Romanian and English.

2.4 MALT Parser

MaltParser [NHN+07] is a system for data-driven dependency parsing, which can be used to induce a parsing model from treebank data and to parse new data using an induced model[6]. Parsers developed using MaltParser have achieved state-of-the-art accuracy for a number of languages. However, please note that MaltParser is a complex system with many parameters that need to be optimized.

We use MALT for dependency parsing of Catalan, English, French, Romanian and Spanish.

2.5 BIOS

Bios [STC05] is a suite of syntactico-semantico analyzers that include the most common tools needed for the shallow analysis of English text[7]. The following tools are included:

- Smart tokenizer that recognizes abbreviations, SGML tags etc
- Part-of-speech (POS) tagger. The POS tagger is implemented as a a wrapper around the TNT tagger by Thorsten Brants
- Syntactic chunking using the labels promoted by the CoNLL chunking evaluations
- Named-Entity Recognition and Classification (NERC) for the CoNLL entity types plus an additional 11 numerical entity types

BIOS can be configured for very high accuracy but slower execution (using Yamcha) or for high speed and slightly lower accuracy (using its own implementation of an asymmetric

[5] `http://www.cl.cam.ac.uk/~yz360/zpar.html`

[6] `http://maltparser.org/`

[7] `http://www.surdeanu.name/mihai/bios/`

Perceptron). It has built-in models for both case-sensitive and case-insensitive text. It can be retrained on annotated corpora. It has a clean and easy to use Java API.

For the purpose of this project we are currently using BIOS for Syntactic Chunking and Named Entity Recognition of English. This is required as a preprocessing step by the SwiRL semantic role labeling tool.

2.6 C&C Tools

The C&C tools [CCB07] are built around a wide-coverage Combinatory Categorial Grammar (CCG) parser [8]. The parser not only recovers the local dependencies output by treebank parsers, but also the long-range depdendencies inherent in constructions such as extraction and coordination. CCG is a lexicalized grammar formalism, so that each word in a sentence is assigned an elementary syntactic structure, in CCG's case a lexical category expressing subcategorisation information. Statistical tagging techniques can assign lexical categories with high accuracy and low ambiguity. The combination of finite-state supertagging and highly engineered C++ leads to a parser which can analyse up to 30 sentences per second on standard hardware. The C&C tools also contain a number of Maximum Entropy taggers, including the CCG supertagger, a POS tagger, chunker, and named entity recogniser. The taggers are highly efficient, with processing speeds of over 100,000 words per second. Finally, the various components, including the morphological analyser morpha[MCP01], are combined into a single program. The output from this program —a CCG derivation, POS tags, lemmas, and named entity tags— is used by the module Boxer to produce interpretable structure in the form of Discourse Representation Structures.

We use the C&C Tools for the syntactic and discoursive analysis of English.

2.7 SwiRL

SwiRL [ST05, MSCT05] is a Semantic Role Labeling (SRL) system for English constructed on top of full syntactic analysis of text [9]. The syntactic analysis is performed using Eugene Charniak's parser. SwiRL trains one classifier for each argument label using a rich set of syntactic and semantic features. The classifiers are learned using one-vs-all AdaBoost classifiers.

SwiRL has state-of-the-art performance: currently its F1 on the WSJ corpus is 77+, and on the Brown corpus it is 66+ points. SwiRL ranks fifth among the systems that participated

[8] `http://svn.ask.it.usyd.edu.au/trac/candc`

[9] `http://www.surdeanu.name/mihai/swirl/`

at the CoNLL shared task evaluation, but all the systems that scored higher were actually combinations of several individual models. SwiRL is fairly robust, it can work with case-sensitive and case-insensitive text. It can also be retrained on annotated corpora. It has an easy to use API, so you can easily integrate it with your software.

We have used SwiRL for the semantic role labeling of English.

2.8 JointParser

The jointparser [LM08, LBM09] is a parser that jointly annotates syntax and semantics [10]. It performs syntactic parsing, shallow semantic parsing and predicate identification. And it is one of the few parsers that simultaneously learns and annotates syntax and semantics. The jointparser extends the Eisner algorithm to annotate semantics by assigning semantic links at each dependency scoring step. The learning is based on an averaged Perceptron. For efficiency reasons, some syntax-based features used in the semantic classifier are pre-computed. The predicate identification is done as a previous step. The implementation of the model does not give us an optimal syntactic-semantic tree as we pre-compute some features.

We have used the jointparser for the semantic role labeling of Catalan, English and Spanish.

3 Annotated Corpora

Some of the linguist analyzers have been retrained for new languages. In the following we provide a brief description of the corpora used.

3.1 Ancora Corpus

AnCora[11] consist of a Catalan corpus (AnCora-CA) and a Spanish corpus (AnCora-ES), each of them of 500,000 words from the news domain [MMMB08]. The corpora are annotated at different levels:

- Lemma and Part of Speech
- Syntactic constituents and functions
- Argument structure and thematic roles

[10] `http://nlp.lsi.upc.edu/jointparser/demo/`

[11] `http://clic.ub.edu/corpus/ancora`

- Semantic classes of the verb
- Denotative type of deverbal nouns
- Nouns related to WordNet synsets
- Named Entities
- Coreference relations
- AnCora corpus is mainly based on journalist texts.

3.2 French Treebank

The French treebank [12] is made of 12,531 sentences from the Le Monde newspaper, annotated for morphology and phrase-structure [ACcT03]. Further, some of the nodes are labeled with a grammatical function. This is necessary because a given structural position may correspond to different grammatical relations.

Morphosyntactic annotation →

- Part of speech (POS) based on 15 lexical categories
- Subcategorization
- Inflection
- Lemma (canonical form)
- Parts (with similar morphosyntactic tags) for compounds

Constituent annotation →

- Surface and shallow annotations, compatible with various syntactic frameworks
- The phrasal tagset includes 10 categories
- Discontinuous constituents are not annotated
- Headless phrases are allowed

Function annotation →

- Grammatical functions associated with major constituents are annotated
- The functional tagset includes 8 function types
- No more than one fucntion can be tagged on a constituent, except for verbal nucleus which bear all the functions of their pronominal clitics.

[12] `http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-en.php`

- Only surface functions are encoded
- The link between the dependent and the head is not coded, so long distance dependencies are not taken into account.

3.3 Penn Treebank

The Penn Treebank [13] consists of over 4.5 million words of American English from the news domain annotated with part-of-speech information and syntactic structure [MSM93].

3.4 Prague Dependency Treebank

The Prague Dependency Treebank 2.0 (PDT 2.0) [14] contains a large amount of Czech texts with complex and interlinked morphological (2 million words), syntactic (1.5 MW) and complex semantic annotation (0.8 MW); in addition, certain properties of sentence information structure and coreference relations are annotated at the semantic level [Haj04].

3.5 Romanian Dependency Treebank

The Romanian Treebank [15] developed in the RORIC-LING project [16] consists of 36,150 tokens (punctuation excluded) and comprises newspapers articles, mostly on political and administrative subjects [HP03]. It contains 4,042 sentences, having a mean sentence length of 8.94 tokens per sentence. The type/token ratio of 0.245 indicates a rather frequent repetition of certain types. The texts were chosen so as to offer a representative sample of modern written standard Romanian. However, texts including complex ambiguities were avoided as much as possible and removed from the treebank.

The texts were annotated with part-of-speech (POS) tags and information about the dependency relations (annotation of the head and the dependent) and dependency labels. All the dependency graphs for the sentences in the treebank are connected, projective, rooted, acyclic and any node has at most one head.

[13] http://www.cis.upenn.edu/~treebank/

[14] http://ufal.mff.cuni.cz/pdt2.0/

[15] http://www.phobos.ro/roric/texts/xml/

[16] http://phobos.cs.unibuc.ro/roric/

4 Data Sets

We have compiled a static corpora collection to be used for training and testing our MT Systems and/or evaluation methods. We are using publicly available data from several sources. A brief description is provided below.

4.1 CzEng 0.9

The CzEng 0.9 (Czech-English Parallel Corpus, version 0.9)[17] is the third release of a sentence-parallel Czech-English corpus compiled at the Institute of Formal and Applied Linguistics (ÚFAL) freely available for non-commercial and research purposes. CzEng 0.9 contains 8.0 million parallel sentences (93 million English and 82 million Czech tokens) from seven different types of sources automatically annotated at surface and deep (a- and t-) layers of syntactic representation. The number of sentences and nodes of a given layer and language per data source is given Table 1.

		English		**Czech**	
Source	**Sentences**	**a-layer**	**t-layer**	**a-layer**	**t-layer**
Movie Subtitles	3,549,367	26,550,305	16,615,991	22,175,284	16,675,187
EU Legislation	1,589,036	31,725,089	19,458,544	28,484,512	19,310,396
Technical Doc.	1,212,494	9,099,748	6,339,129	8,460,491	6,512,247
Fiction	1,036,952	17,045,233	10,861,341	15,031,926	11,102,760
Parallel Web Pages	464,522	4,946,552	3,666,149	4,750,757	3,667,297
News	140,191	3,196,303	2,019,758	2,945,777	2,220,789
Project Navajo	37,239	612,826	385,292	539,659	405,484
Total	8,029,801	93,176,056	59,346,204	82,388,406	59,894,160

Table 1: CzEng 0.9 parallel corpus description

4.2 El Periódico

This corpus has been extracted from the bilingual newspaper 'El Periódico' [18] using data from 2001 to 2010. This corpus is available through ELRA (Catalog Reference : W0053) [19]. A brief numerical description of the corpus is available in Table 2.

[17] http://ufal.mff.cuni.cz/czeng/czeng09/

[18] http://www.elperiodico.com/

[19] http://catalog.elra.info/product_info.php?products_id=1122

Language	#Sentences	#Tokens	Vocabulary
Spanish	2,478,130	63,200,201	500,961
Catalan	2,478,130	65,789,528	459,077

Table 2: "El Periódico" corpus description

4.3 WMT 2010

We used a subset of the data sets included in the Fifth Workshop on Statistical Machine Translation (WMT10) [20]. A brief numerical description of monolingual and parallel corpora selected is available in Tables 3 and 4, respectively.

Corpus	Language	#Sentences	#Tokens	Vocabulary
europarl-v5	English	1,843,035	49,681,693	129,617
europarl-v5	French	1,855,589	53,700,804	149,357
europarl-v5	Spanish	1,822,021	51,240,123	193,342
news.shuffled	English	48,653,884	1,128,574,893	2,130,940
news.shuffled	French	15,670,745	377,022,425	1,171,689
news.shuffled	Spanish	3,857,414	107,071,320	613,144
news.shuffled	Czech	13,042,040	205,278,456	1,761,355
news-commentary10	Czech	103,227	2,205,529	131,146
news-commentary10	English	125,879	2,972,589	59,514
news-commentary10	French	97,033	2,675,189	59,669
news-commentary10	Spanish	107,862	2,948,913	73,895

Table 3: WMT10 data description (monolingual corpora)

Moreover, in order to speed up the experimental cycle, we have created an additional resource, *'wmt10.select_es-en'*, a Spanish-English parallel corpus consisting of 2 million sentence pairs randomly selected from the various sources (europarl-v5_es-en, news-commentary10_es-en and united_nations_es-en) keeping their respective proportion. A brief description is provided in Table 5.

4.4 EuroParl

Since Romanian translation is not addressed in WMT, for this language we used directly EuroParl Release v6 data[21]. A brief numerical corpora description is available in Table 6.

[20] http://www.statmt.org/wmt10/

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

Corpus	Language	#Sentences	#Tokens	Vocabulary
europarl-v5_es-en	Spanish	1,650,152	47,711,764	186,468
europarl-v5_es-en	English	1,650,152	45,665,702	124,017
europarl-v5_fr-en	French	1,683,156	50,431,815	144,081
europarl-v5_fr-en	English	1,683,156	46,744,790	124,832
news-commentary10_cz-en	Czech	94,742	2,054,462	126,658
news-commentary10_cz-en	English	94,742	2,250,434	52,197
news-commentary10_es-en	Spanish	98,598	2,717,797	70,949
news-commentary10_es-en	English	98,598	2,388,829	54,237
news-commentary10_fr-en	French	84,624	2,377,740	56,595
news-commentary10_fr-en	English	84,624	2,064,152	50,461
united_nations_es-en	Spanish	6,222,450	213,337,370	470,445
united_nations_es-en	English	6,222,450	187,308,015	443,280

Table 4: WMT10 data description (parallel corpora)

Corpus	Language	#Sentences	#Tokens	Vocabulary
wmt10.select_es-en	English	1,994,058	52,667,251	224,945
wmt10.select_es-en	Spanish	1,994,058	57,611,897	269,413

Table 5: Description of ‘wmt10.select’ corpus

5 Data Representation Formats

The exchange data representation format for the FAUST project is CoNLL 2009 format [HCJ+09] (see Subsection 5.1). However, for convenience, we use several representation formats as required by the different production pipelines of each partner (see Subsections 5.3 and 5.2). Scripts for the transformation between all the formats and the CoNLL format have been implemented and will be distributed with annotated corpora.

In the following we provide a brief description of the different formats employed with particular focus on the CoNLL format.

Corpus	Language	#Sentences	#Tokens	Vocabulary
europarl-v6_ro-en	Romanian	222,854	5,865,513	73,012
europarl-v6_ro-en	English	222,854	5,908,150	45,016

Table 6: Numerical description of the Europarl-v6 Romanian-English parallel corpus

5.1 CoNLL 2009 Representation Format

Data adheres to the following rules:

- Data files contain sentences separated by a blank line.
- A sentence consists of one or more tokens, each one starting on a new line.
- Each line consists of ten fields:

  ID FORM LEMMA PLEMMA POS PPOS FEAT PFEAT HEAD PHEAD DEPREL PDEPREL FILLPRED PRED APREDs

  Fields are separated by a single tab character. Space/blank characters are not allowed in within fields.

  ID, FORM, LEMMA, POS, FEAT, HEAD and DEPREL are the same as in the CoNLL-2006 and CoNLL-2007 Shared Tasks.

  FEAT is a set of morphological features (separated by —) defined for a particular language, e.g. more detailed part of speech, number, gender, case, tense, aspect, degree of comparison, etc.

  The P-columns (PLEMMA, PPOS, PFEAT, PHEAD and PDEPREL) are the automatically predicted variants of the gold-standard LEMMA, POS, FEAT, HEAD and DEPREL columns. They are produced by independently (or cross-)trained taggers and parsers.

  PRED is the same as in the 2008 English data. APREDs correspond to 2008's ARGs. FILLPRED contains Y for lines where PRED is/should be filled.

  A detailed column description is available in Tables 7 (input columns) and 8 (output columns).
- All data files will contains these ten fields, although only the ID, FORM, CPOSTAG, POSTAG, HEAD and DEPREL columns are guaranteed to contain non-dummy (i.e. non-underscore) values for all languages.
- Data files are UTF-8 encoded (Unicode).

Common Values

UNDERSCORE ('_') is used for "unknown", "unannotated", "unfilled" etc. value (simply for all those cells in the large data table that do not display a defined label from the label sets described for each language in the documentation). The same character is also used in all cells of those columns which are completely unfilled because (e.g.) data for the particular language is not available.

Field number	**Field name**	**Description**
1	ID	Token counter, starting at 1 for each new sentence.
2	FORM	Word form or punctuation symbol.
3	LEMMA	Lemma or stem (depending on particular data set) of word form, or an underscore if not available.
4	CPOSTAG	Coarse-grained part-of-speech tag, where tagset depends on the language.
5	POSTAG	Fine-grained part-of-speech tag, where the tagset depends on the language, or identical to the coarse-grained part-of-speech tag if not available.
6	FEATS	Unordered set of syntactic and/or morphological features (depending on the particular language), separated by a vertical bar or an underscore if not available
7	HEAD	Head of the current token, which is either a value of ID or zero ('0'). Note that depending on the original treebank annotation, there may be multiple tokens with an ID of zero.
8	DEPREL	Dependency relation to the HEAD. The set of dependency relations depends on the particular language. Note that depending on the originale treebank annotation, th dependency relation may be meaningfull or simply 'ROOT'.

Table 7: CoNLL 2009 Data Representation Format (input fields)

Field number	Field name	Description
9	PHEAD	Projective head of current token, which is either a value of ID or zero ('0'), or an underscore if not available. Note that depending on the original treebank annotation, there may be multiple tokens an with ID of zero. The dependency structure resulting from the PHEAD column is guaranteed to be projective (but is not available for all languages), whereas the structures resulting from the HEAD column will be non-projective for some sentences of some languages (but is always available).
10	PDEPREL	Dependency relation to the PHEAD, or an underscore if not available. The set of dependency relations depends on the particular language. Note that depending on the original treebank annotation, the dependency relation may be meaningfull or simply 'ROOT'.
11	PRED	Rolesets of the semantic predicates in this sentence. This includes both nominal and verbal predicates. The split-form tokens that are not semantic predicates must be marked with "_". We use the same roleset names as the PropBank and NomBank frames.
12+	ARG	Columns with argument labels for the each semantic predicate following textual order, i.e., the first column corresponds to the first predicate in PRED, the second column to the second predicate, etc. Note that, because this algorithm uniquely identifies the ID of the corresponding predicate, it is sufficient to store the label of the argument here. The argument labels for verbal predicates follow the PropBank conventions. Labels of arguments to nominal predicates use NomBank conventions. The differences between PropBank and NomBank labels are discussed here.

Table 8: CoNLL 2009 Data Representation Format (output fields)

5.2 Prague Markup Language

Thee Prague Markup Language (PML) is a generic data representation format for linguistic annotation based on XML[22]. PML has been used for the representation of morphology, syntax and deep syntactic annotation in the Prague Dependency Treebank (PDT). Czech annotation is divided into four interlinked layers:

1. word layer (w-layer) - tokenized sentence
2. morpological layer (m-layer) - lemma and tag is assigned to the each token
3. analytical layer (a-layer) - surface syntax, dependency tree and dependency relations (analytical functions)
4. tectogrammatical layer (t-layer) - deep syntax, semantic functions and other attributes, coreferences

5.3 Other Formats

ZPar and the C&C Tools have specific representation formats. Further details may be found in their respective websites.

[22] `http://ufal.mff.cuni.cz/jazz/PML/index_en.html`

6 Annotation Status as of February 2011

Table 9 provides a summary of the current status of corpora annotation, i.e., the level of annotation for each corpora and the tools employed[23].

The following abbreviations have been used:

pos part-of-speech tagging

lem lemmatization

sc base phrase syntactic chunking

dp dpendency parsing

ne named entity recognition and classification

sr semantic role labeling

tecto full TectoMT annotation

dr discourse representation

All corpora have been tokenized using the tokenizer provided inside the Moses open source toolkit for statistical MT [KSF$^+$06].

[23] Dependency Parsing ('dp'), Semantic Role Labeling ('sr') and Discourse Representation ('dr') annotation of most of the corpora is currently undergoing. The reason is that the linguistic processors employed do not allow for fast massive text processing. These annotations are expected to be completed along the next few months.

Language	**Corpus**	**#Sentences**	**Annotation**	**Tools**
Catalan	el_periodico	2,478,130	pos+lem+dp+sr	SVMTool+JointParser
Czech	news.shuffled	13,042,040	tecto	TectoMT
Czech	CzEng 0.9	8,029,801	tecto	TectoMT
Czech	news-commentary10	103,227	tecto	TectoMT
Czech	news-commentary10_cz-en	94,742	tecto	TectoMT
English	news.shuffled	48,653,884	pos+lem+dp+sc+ne+sr+dr	SVMTool+MALT+ZPar+BIOS+SwiRL+C&C
English	CzEng 0.9	8,029,801	tecto	TectoMT
English	united_nations_es-en	6,222,450	pos+lem+dp+sc+ne+sr+dr	SVMTool+MALT+ZPar+BIOS+SwiRL+C&C
English	europarl-v5	1,843,035	pos+lem+dp+sc+ne+sr+dr	SVMTool+MALT+ZPar+BIOS+SwiRL+C&C
English	europarl-v5_es-en	1,650,152	pos+lem+dp+sc+ne+sr+dr	SVMTool+MALT+ZPar+BIOS+SwiRL+C&C
English	europarl-v5_fr-en	1,683,156	pos+lem+dp+sc+ne+sr+dr	SVMTool+MALT+ZPar+BIOS+SwiRL+C&C
English	europarl-v6_ro-en	222,854	pos+lem+dp+sc+ne+sr+dr	SVMTool+MALT+ZPar+BIOS+SwiRL+C&C
English	news-commentary10	125,879	pos+lem+dp+sc+ne+sr+dr	SVMTool+MALT+ZPar+BIOS+SwiRL+C&C
English	news-commentary10_cz-en	94,742	pos+lem+dp+sc+ne+sr+dr	SVMTool+MALT+ZPar+BIOS+SwiRL+C&C
English	news-commentary10_es-en	98,598	pos+lem+dp+sc+ne+sr+dr	SVMTool+MALT+ZPar+BIOS+SwiRL+C&C
English	news-commentary10_fr-en	84,624	pos+lem+dp+sc+ne+sr+dr	SVMTool+MALT+ZPar+BIOS+SwiRL+C&C
English	wmt10.select_es-en	1,994,058	pos+lem+dp+sr	SVMTool+ZPar+JointParser
French	news.shuffled	15,670,745	pos+lem+dp	SVMTool+MALT
French	europarl-v5	1,855,589	pos+lem+dp	SVMTool+MALT
French	europarl-v5_fr-en	1,683,156	pos+lem+dp	SVMTool+MALT
French	news-commentary10	97,033	pos+lem+dp	SVMTool+MALT
French	news-commentary10_fr-en	84,624	pos+lem+dp	SVMTool+MALT
Romanian	europarl-v6_ro-en	222,854	pos+dp	SVMTool+MALT+ZPar
Spanish	united_nations_es-en	6,222,450	pos+lem+dp+sr	SVMTool+JointParser
Spanish	news.shuffled	3,857,414	pos+lem+dp+sr	SVMTool+JointParser
Spanish	el_periodico	2,478,130	pos+lem+dp+sr	SVMTool+JointParser
Spanish	europarl-v5	1,822,021	pos+lem+dp+sr	SVMTool+JointParser
Spanish	europarl-v5_es-en	1,650,152	pos+lem+dp+sr	SVMTool+JointParser
Spanish	news-commentary10	107,862	pos+lem+dp+sr	SVMTool+JointParser
Spanish	news-commentary10_es-en	98,598	pos+lem+dp+sr	SVMTool+JointParser
Spanish	wmt10.select_es-en	1,994,058	pos+lem+dp+sr	SVMTool+JointParser

Table 9: Corpora Annotation

References

[ACcT03] Anne Abeillé, Lionel Clément, and François Toussenel. Building a Treebank for French. In Anne Abeillé, editor, *Treebanks*. Kluwer, Dordrecht, 2003.

[CCB07] James Curran, Stephen Clark, and Johan Bos. Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[GM04a] Jesús Giménez and Lluís Màrquez. Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited. In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing III*, Current Issues in Linguistic Theory (CILT), pages 153–162, Amsterdam, 2004. John Benjamin Publishers. ISBN 90-272-4774-9.

[GM04b] Jesús Giménez and Lluís Màrquez. SVMTool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC)*, pages 43–46, 2004.

[Haj04] Jan Hajič. Complex Corpus Annotation: The Prague Dependency Treebank. Bratislava, Slovakia, 2004. Jazykovedný ústav Ľ. Štúra, SAV.

[HCJ$^+$09] Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[HP03] Florentina Hristea and Marius Popescu. A Dependency Grammar Approach to Syntactic Analysis with Special Reference to Romanian. In Florentina Hristea and Marius Popescu, editors, *Building Awareness in Language Technology.* University of Bucharest Publishing House, 2003.

[KSF$^+$06] Philipp Koehn, Wade Shen, Marcello Federico, Nicola Bertoldi, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Ondrej Bojar, Richard Zens, Alexandra Constantin, Evan Herbst, and Christine Moran. Open Source Toolkit for Statistical Machine Translation. Technical report, Johns Hopkins University Summer Workshop. `http://www.statmt.org/jhuws/`, 2006.

[LBM09] Xavier Lluís, Stefan Bott, and Lluís Màrquez. A Second-Order Joint Eisner Model for Syntactic and Semantic Dependency Parsing. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 79–84, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[LM08] Xavier Lluís and Lluís Màrquez. A Joint Model for Parsing Syntactic and Semantic Dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 188–192, Manchester, England, August 2008. Coling 2008 Organizing Committee.

[MCP01] Guido Minnen, John Carroll, and Darren Pearce. Applied morphological processing of English. *Natural Language Engineering*, 7:207–223, September 2001.

[MMMB08] M. A. Martì, M.Taulé, L. Màrquez, and M. Bertran. ANCora: A Multilingual and Multilevel Annotated Corpus. 2008.

[MPRH05] Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 523–530, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[MSCT05] Lluís Màrquez, Mihai Surdeanu, Pere Comas, and Jordi Turmo. Robust Combination Strategy for Semantic Role Labeling. In *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*, 2005.

[MSM93] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[NHN$^+$07] Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13:95–135, September 2007.

[Pv10] Martin Popel and Zdeněk Žabokrtský. TectoMT: Modular NLP Framework. In Hrafn Loftsson, Eiríkur Rögnvaldsson, and Sigrún Helgadóttir, editors, *Advances in Natural Language Processing*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304. Springer Berlin / Heidelberg, 2010. 10.1007/978-3-642-14770-8_33.

[SHV+07] Drahomíra ”johanka” Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007*, pages 67–74, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[ST05] Mihai Surdeanu and Jordi Turmo. Semantic Role Labeling Using Complete Syntactic Analysis. In *Proceedings of CoNLL Shared Task*, 2005.

[STC05] Mihai Surdeanu, Jordi Turmo, and Eli Comelles. Named Entity Recognition from Spontaneous Open-Domain Speech. In *Proceedings of the 9th International Conference on Speech Communication and Technology (Interspeech)*, 2005.

[ZC11] Yue Zhang and Stephen Clark. Syntactic Processing using the Generalized Perceptron and Beam Search, 2011.

A Files

We have created a different file for each corpus and language, as shown in Table 9. Filenames follow this rule: "{`<corpus_name>.<language>.<format>.gz''`}, where language is indeed an abbreviation → { *'ca'* for Catalan, *'cz'* for Czech, *'en'* for English, *'fr'* for French, *'es'* for Spanish and *'ro'* for Romanian }, and format indicates the data representation format (currently only *'conll'*). This applies for all corpora except CzEng 0.9 for which we deliver the files from the original release[24] and provide a script for transformation into CoNLL's format[25]. Finally, the 'gz' extension indicates that all files have been compressed using *gzip*, a standard tool, so as to save both disk space and bandwidth.

Table 10 shows the list of files we have made available for public release through the FAUST project wiki[26].

[24] http://ufal.mff.cuni.cz/czeng/czeng09/

[25] http://www.faust-fp7.eu/faust/pub/Main/DataReleases/czeng_to_conll.pl.txt

[26] http://www.faust-fp7.eu/faust/Main/DataReleases

Language	Corpus	File Name
Catalan	el_periodico	`ftp://mi.eng.cam.ac.uk/data/faust/el_periodico_ca-es.ca.conll.gz`
Czech	news.shuffled	`ftp://mi.eng.cam.ac.uk/data/faust/news.shuffled.cz.conll.gz`
Czech	CzEng 0.9	`http://ufal.mff.cuni.cz/czeng/czeng09/`
Czech	news-commentary10	`ftp://mi.eng.cam.ac.uk/data/faust/news-commentary10.cz.conll.gz`
Czech	news-commentary10_cz-en	`ftp://mi.eng.cam.ac.uk/data/faust/news-commentary10_cz-en.cz.conll.gz`
English	news.shuffled	`ftp://mi.eng.cam.ac.uk/data/faust/news.shuffled.en.conll.part1.rar`
		...
		`ftp://mi.eng.cam.ac.uk/data/faust/news.shuffled.en.conll.part8.rar`
English	CzEng 0.9	`http://ufal.mff.cuni.cz/czeng/czeng09/`
English	united_nations_es-en	`ftp://mi.eng.cam.ac.uk/data/faust/united_nations_es-en.en.conll.gz`
English	europarl-v5	`ftp://mi.eng.cam.ac.uk/data/faust/europarl-v5.en.conll.gz`
English	europarl-v5_es-en	`ftp://mi.eng.cam.ac.uk/data/faust/europarl-v5_es-en.en.conll.gz`
English	europarl-v5_fr-en	`ftp://mi.eng.cam.ac.uk/data/faust/europarl-v5_fr-en.en.conll.gz`
English	europarl-v6_ro-en	`ftp://mi.eng.cam.ac.uk/data/faust/europarl-v6_ro-en.en.conll.gz`
English	news-commentary10	`ftp://mi.eng.cam.ac.uk/data/faust/news-commentary10.en.conll.gz`
English	news-commentary10_cz-en	`ftp://mi.eng.cam.ac.uk/data/faust/news-commentary10_cz-en.conll.gz`
English	news-commentary10_es-en	`ftp://mi.eng.cam.ac.uk/data/faust/news-commentary10_es-en.conll.gz`
English	news-commentary10_fr-en	`ftp://mi.eng.cam.ac.uk/data/faust/news-commentary10_fr-en.conll.gz`
English	wmt10.select_es-en	`ftp://mi.eng.cam.ac.uk/data/faust/wmt10.select_es-en.en.conll.gz`
French	news.shuffled	`ftp://mi.eng.cam.ac.uk/data/faust/news.shuffled.fr.conll.part1.rar`
		`ftp://mi.eng.cam.ac.uk/data/faust/news.shuffled.fr.conll.part2.rar`
		`ftp://mi.eng.cam.ac.uk/data/faust/news.shuffled.fr.conll.part3.rar`
French	europarl-v5	`ftp://mi.eng.cam.ac.uk/data/faust/europarl-v5.fr.conll.gz`
French	europarl-v5_fr-en	`ftp://mi.eng.cam.ac.uk/data/faust/europarl-v5_fr-en.fr.conll.gz`
French	news-commentary10	`ftp://mi.eng.cam.ac.uk/data/faust/news-commentary10.fr.conll.gz`
French	news-commentary10_fr-en	`ftp://mi.eng.cam.ac.uk/data/faust/news-commentary10_fr-en.fr.conll.gz`
Romanian	europarl-v6_ro-en	`ftp://mi.eng.cam.ac.uk/data/faust/europarl-v6_ro-en.ro.conll.gz`
Spanish	united_nations_es-en	`ftp://mi.eng.cam.ac.uk/data/faust/united_nations_es-en.es.conll.gz`
Spanish	news.shuffled	`ftp://mi.eng.cam.ac.uk/data/faust/news.shuffled.es.conll.gz`
Spanish	el_periodico	`ftp://mi.eng.cam.ac.uk/data/faust/el_periodico_ca-es.es.conll.gz`
Spanish	europarl-v5	`ftp://mi.eng.cam.ac.uk/data/faust/europarl-v5.es.conll.gz`
Spanish	europarl-v5_es-en	`ftp://mi.eng.cam.ac.uk/data/faust/europarl-v5_es-en.es.conll.gz`
Spanish	news-commentary10	`ftp://mi.eng.cam.ac.uk/data/faust/news-commentary10.es.conll.gz`
Spanish	news-commentary10_es-en	`ftp://mi.eng.cam.ac.uk/data/faust/news-commentary10_es-en.es.conll.gz`
Spanish	wmt10.select_es-en	`ftp://mi.eng.cam.ac.uk/data/faust/wmt10.select_es-en.es.conll.gz`

Table 10: Released Files