Computational Linguistics in the Netherlands Journal 3 (2013) 103-120 Submitted 05/2013; Published 12/2013

LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit

Marjan Van de Kauter Marjan.Vandekauter@UGent.be
Geert Coorman Geert.Coorman@ieee.org
Els Lefever Els.Lefever@UGent.be
Bart Desmet Bart.Desmet@UGent.be
Lieve Macken Lieve.Macken@UGent.be
Véronique Hoste Veronique.Hoste@UGent.be

LT^3, Language and Translation Technology Team - Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent
Belgium

Abstract

This paper presents the LeTs Preprocess Toolkit, a suite of robust high-performance preprocessing modules including Part-of-Speech Taggers, Lemmatizers and Named Entity Recognizers. The currently supported languages are Dutch, English, French and German.
We give a detailed description of the architecture of the LeTs Preprocess pipeline and describe the data and methods used to train each component. Ten-fold cross-validation results are also presented. To assess the performance of each module on different domains, we collected real-world textual data from companies covering various domains (a.o. automotive, dredging and human resources) for all four supported languages. For this multi-domain corpus, a manually verified gold standard was created for each of the three preprocessing steps. We present the performance of our preprocessing components on this corpus and compare it to the performance of other existing tools.

1. Introduction

This paper reports on the LeTs Preprocess Toolkit, a suite of linguistic preprocessing modules that currently supports four languages, viz. Dutch, English, French and German. The toolkit includes modules for Part-of-Speech Tagging, Lemmatization and Named Entity Recognition and has been developed in the framework of the TExSIS terminology extraction system (Macken et al. 2013). The TExSIS system aims to automatically extract monolingual and bilingual domain-specific glossaries on the basis of monolingual and parallel text.

The TExSIS system architecture consists of the following components. In a first step, a monolingual corpus or each part of a bilingual corpus is separately preprocessed, meaning that it is split into sentences, tokenized, Part-of-Speech tagged, lemmatized, and finally named entities are extracted. The linguistically preprocessed corpora are then fed to the monolingual term extraction module, which starts by generating candidate terms from syntactically motivated chunks. Next, a set of well-known statistical filters such as frequency, Log-Likelihood Ratio and C-value, are combined to determine the specificity of the candidate terms. In a bilingual term extraction setup, sentence, word and chunk alignment is performed in order to find translational correspondences between the two monolingual term lists.

The performance of the terminology extraction system highly depends on the quality of the linguistic preprocessing. Therefore we decided to build robust high-performance Part-of-Speech Taggers, Lemmatizers and Named Entity Recognizers for the four supported languages. Accurate Part-of-Speech (PoS) Tagging is necessary to identify valid candidate terms based on predefined PoS patterns. Lemmatization allows us to determine the specificity of the candidate terms by calculating

frequencies and other statistical measures both on word form and lemma level. Finally, the TExSIS system not only aims at the extraction of domain-specific terms, but also at the identification of named entities such as persons, organizations, etc.

The development of a new preprocessing toolkit (instead of using different existing modules) was motivated by the ease of integrating one dedicated machine learning algorithm (viz. CRF++ in our case, see Section 2.1) for all preprocessing modules in a practical application, such as the TExSIS terminology extraction system. To our knowledge, there is only one multilingual preprocessing tool, namely TreeTagger (Schmid 1994, Schmid 1995), that supports all four considered languages. TreeTagger, however, does not perform Named Entity Recognition (NER). As NER is a linguistic preprocessing step considered necessary for various NLP tasks, we included a NER module for all four supported languages in the LeTs Preprocess Toolkit.

In this paper, we present in detail the LeTs Part-of-Speech Taggers, Lemmatizers and Named Entity Recognizers that were developed for Dutch, English, French and German. Furthermore, we evaluate these modules in a multi-domain setting. Most preprocessing tools have been evaluated on the same type of data that was used for training (usually general-domain text such as newswire). In real-world scenarios, however, these tools are applied to texts from various domains. For this reason, we constructed an evaluation corpus of real-world textual data from different companies and domains for all four of the supported languages. For this corpus, a manually verified gold standard was created for Part-of-Speech Tagging, Lemmatization and Named Entity Recognition. We present the performance of the LeTs preprocessing modules on this multi-domain gold standard corpus and compare it to the performance of other existing tools.

The remainder of this paper is structured as follows. Section 2 gives a detailed description of the three modules for preprocessing (viz. Part-of-Speech Tagging, Lemmatization and Named Entity Recognition) that were developed. In Section 3, we elaborate on the multi-domain evaluation of these modules. Finally, Section 4 concludes this paper and gives some directions for future work.

2. Preprocessing pipeline

Figure 1 presents the different components of the LeTs Preprocess pipeline. The LeTs Toolkit accepts input files in various character encoding formats and first converts these files to UTF-8 encoding. The texts are then split into sentences and tokenized by rule-based methods which we based on the Sentence Splitter and Tokenizer developed resp. for the Europarl Parallel Corpus (Koehn 2005) and for TreeTagger (Schmid 1994, Schmid 1995). In this paper, we will focus on the subsequent steps of the preprocessing pipeline, viz. the LeTs components that were developed for Part-of-Speech Tagging, Lemmatization and Named Entity Recognition. We will not discuss the LeTs Chunking modules, which are based on the research performed by Macken and Daelemans (2010) and start from the output of the LeTs Part-of-Speech Tagging and Lemmatization components.

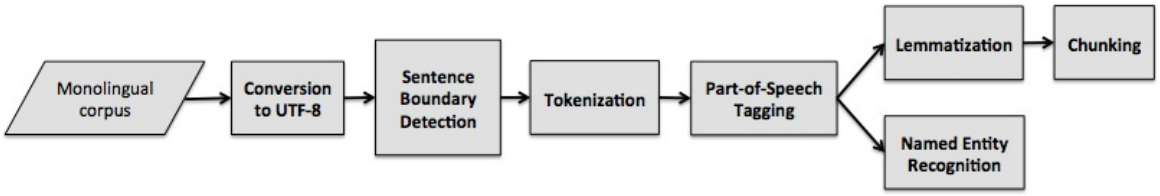

Figure 1: Architecture of the LeTs Preprocess pipeline

The output of the LeTs Preprocess components is written to tab-delimited text files. This column format can easily be converted to other linguistic annotation formats such as FoLiA (van Gompel 2012).

2.1 Machine learning framework

To develop the linguistic preprocessing modules described in this paper, a supervised machine learning approach was taken. Part-of-Speech Tagging, Named Entity Recognition and even Lemmatization are Natural Language Processing (NLP) tasks that can be regarded as classification problems, where each token (i.e. word in a text) is to be assigned a class tag, viz. a PoS, named entity or lemma tag. In supervised machine learning approaches, a classifier learns from a set of manually labeled example tokens, and is then able to predict a class label for unseen examples. Several types of machine learning algorithms have been successfully applied to these tasks, for example Hidden Markov Models (HMMs) (Kupiec 1992, Zhou and Su 2002), Support Vector Machines (SVMs) (Isozaki and Kazawa 2002, Giménez and Màrquez 2004, Chrupała 2006), Memory-Based Learning (Daelemans et al. 1996, van den Bosch and Daelemans 1999, Tjong Kim Sang 2002b), etc.

To train the modules of the LeTs Preprocess Toolkit, we made use of an implementation of Conditional Random Fields, CRF++[1]. Conditional Random Fields (CRFs) are used to segment and label sequential data (Lafferty et al. 2001). They have been successfully applied to a variety of NLP tasks, including Part-of-Speech Tagging (Lafferty et al. 2001) and Named Entity Recognition (McCallum and Li 2003). For example, in the research conducted by Desmet and Hoste (2010a) for Dutch Named Entity Recognition, it is shown that the CRF++ algorithm outperforms several other classifiers (viz. SVMs and MBL algorithms). The decision to use Conditional Random Fields for the LeTs Preprocess Toolkit was also motivated by the fact that the CRF++ toolkit is available under a LGPL/BSD dual license.

In the following sections, we describe the data and methods applied to train Part-of-Speech Taggers, Lemmatizers and Named Entity Recognizers for Dutch, English, French and German using version 0.57 of CRF++.

2.2 Part-of-Speech Tagging

Part-of-Speech Tagging is the task of assigning each token in a text its correct grammatical category (e.g. noun, verb, adjective, adverb, etc.) depending on its context. To train the LeTs Part-of-Speech Taggers, we collected text corpora labeled with manually determined or corrected PoS tags. The PoS tagsets applied to label these corpora are widely used in the NLP domain.

For Dutch, we used the manually verified Dutch part of the Dutch Parallel Corpus (DPC) (Paulussen et al. 2013) and the texts from the Lassy Small treebank (van Noord et al. 2009) that were not included in the DPC collection. These corpora cover various text types (e.g. journalistic, instructive, administrative texts, etc.) and both have been annotated using the D-Coi/CGN PoS tagset. This tagset is characterized by a high level of granularity: for each token, the main word class is determined (e.g. *V* for verbs), followed by a number of morphosyntactic features (e.g. the tense of the verb). A detailed overview of the tagset, which consists of over 300 tags, can be found in Van Eynde (2005). For two word classes, namely *VNW* (pronouns) and *LID* (articles), we removed certain morphosyntactic features from the tagset[2] which were deemed less important for the remaining steps of the preprocessing pipeline, viz. Lemmatization, Named Entity Recognition and Chunking. Discarding these features resulted in a reduction of the tagset to 146 tags.

The English training corpus is composed of approximately one million words taken from the Penn Treebank (Marcus et al. 1993), complemented by the manually corrected English part of the DPC. As mentioned above, the Dutch Parallel Corpus contains a variety of text types. The Penn Treebank consists mainly of newspaper articles from the Wall Street Journal and a small set of transcripts from the Air Travel Information System (ATIS) corpus (Hemphill et al. 1990). Both

1. http://crfpp.googlecode.com/svn/trunk/doc/index.html
2. For articles, the features case and NPAGR (denoting the type of noun phrase an article can occur with) were removed from the tagset. For tokens of the pronoun category, we discarded the following morphosyntactic features: case, status (e.g. stressed), position (e.g. prenominal), declension, NPAGR, the degree of comparison and finally for nominally used determiners the number (i.e. with or without -n suffix).

corpora are enriched with PoS tags from the Penn Treebank tagset (Santorini 1990). Some tokens in the Penn Treebank had received multiple PoS tags. For these tokens, only the first tag was retained for training the Part-of-Speech Tagger.

The manually verified French part of the Dutch Parallel Corpus served as the training data for French. In the DPC project, French texts were labeled using the GRACE PoS tagset (Paroubek 2000), which is made up of over 300 tags. Since the manually verified part of the DPC consists of only 337,000 tokens for French, and preliminary experiments showed that more data is needed to achieve acceptable results on such an extensive PoS tagset, the number of labels was reduced. The GRACE PoS tags assigned to each token in the DPC collection were mapped to the TreeTagger tagset[3], and one extra label was added for foreign words (*FW*). Mapping the PoS tags of the GRACE tagset to the TreeTagger tagset was a relatively straightforward process, with the exception of a few PoS categories. For example, tokens labeled as nominal numerals sometimes needed to be mapped to the *NUM* category for numerals, but other times to the *NOM* category for nouns. In order to map PoS tags of these categories, we needed to establish certain token-based rules (e.g. the word *millions* is to be labeled as *NOM*, but the word *100* needs to receive the PoS tag *NUM*). This was also the case for symbols and punctuation marks. Establishing these rules was not always easy, since extensive guidelines for PoS annotation using the TreeTagger tagset are not available. For this reason, the French training corpus was also processed using the TreeTagger tool. In case of doubt about which mapping procedure to follow, we looked at the conventions adopted by TreeTagger[4].

Finally, we made use of the NEGRA Corpus (Skut et al. 1998), the TIGER Treebank (Brants et al. 2002) and the Tüba-D/Z Treebank (Telljohann et al. 2004) to train the German Part-of-Speech Tagger. These three corpora are composed of newspaper text annotated with the Stuttgart-Tübingen-Tagset (STTS) (Schiller et al. 1999). However, the STTS tagset is used slightly differently in the TIGER Treebank (Smith 2003). In this dataset, attributive indefinite pronouns occurring with and without determiners are assigned the same PoS tag (*PIAT*), whereas they receive different tags in the NEGRA Corpus and Tüba-D/Z Treebank (resp. *PIDAT* and *PIAT*). When combining the three corpora, all tokens labeled *PIDAT* were assigned the tag *PIAT*.

Table 1 shows the size of the training data used for Part-of-Speech Tagging in each of the supported languages.

	number of tokens	number of PoS tags
Dutch	1,326,444	146
English	1,552,940	45
French	337,143	34
German	2,608,975	53

Table 1: Properties of the training corpora for Part-of-Speech Tagging

In the training corpora, each token is represented by a feature vector listing the features, i.e. properties of the token that we consider to be useful information to help predict its PoS tag, and a manually verified PoS label. When training CRF++ on these corpora, models are created to predict the PoS tag of new tokens. The features extracted for Part-of-Speech Tagging in all four languages are the following:

3. `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html`
4. Although we compared the GRACE PoS tags in the DPC to the PoS tags assigned to the tokens by TreeTagger for certain PoS categories, the conventions we followed in mapping the PoS tags sometimes differed from the rules adopted by TreeTagger (see Section 3.2).

1. **basic features**
    - the original form of the token
    - the lowercased form of the token
    - the number of characters in the token
2. **orthographic features**
    - does the token start with a capital letter?
    - does the token contain any capital letters, apart from the first character?
    - is the entire token capitalized?
    - is the entire token lowercased?
    - does the token contain any numeric characters?
    - does the token contain numeric as well as alphabetic characters?
    - does the token consist of only numeric characters?
    - does the token contain any punctuation marks?
    - does the token consist of only punctuation marks?
    - is there a hyphen inside the token?
    - is the token an initial? (e.g. *J.R.*)
    - is the token a URL?
3. **affix features**
    - prefixes consisting of the first 1 to 3 characters of the token
    - suffixes consisting of the last 1 to 3 characters of the token
4. **sentence features**
    - is the token the first token of the sentence?
    - does every token of the sentence start with a capital letter?
    - is the entire sentence capitalized?
5. **context features**
    - a context window of two words (left and right)

For each of the four supported languages, CRF++ was used to create a Part-of-Speech Tagging model from the corresponding training corpus. Depending on the number of tokens and PoS tags in the training data, the f parameter of CRF++, which sets the cut-off threshold for the features, was set to 1 (the default value) for French, 2 for English, and 3 for Dutch and German.

2.3 Lemmatization

Lemmatization is the process of mapping a token to its base form, i.e. lemma. The LeTs Lemmatizers make use of lexicons listing word forms with their possible word classes and respective lemmas. Word forms may have different lemmas depending on the grammatical category at hand. For example, the verb form *does* is lemmatized as *do*, but the base form of the plural noun *does* is *doe*. Using the word class of a token as a feature for Lemmatization allows for more accurate results.

For Dutch, English and German, the training data consists of word forms taken from the CELEX lexical database (Baayen et al. 1995). For French, we made use of the Morphalou lexicon (Romary

et al. 2004), complemented by entries from the Lefff lexicon (Sagot 2010). When necessary, word forms were added to the lexicons or lemmas were altered to attain consistency with the Lemmatization guidelines applied in the D-Coi and Spoken Dutch Corpus for Dutch (Van Eynde 2005), the Dutch Parallel Corpus for English and French (Paulussen et al. 2013) and the TIGER Treebank for German (Crysmann et al. 2005).

Lemmatization is performed in several steps and takes as input tokenized text labeled with Part-of-Speech tags from the tagsets discussed in Section 2.2. The PoS tag of a token is used as a feature for Lemmatization in order to disambiguate between multiple possible lemmas based on word class. It also allows for more accurate Lemmatization of unseen word forms.

Step 1: Changing the casing of the tokens

A rule-based method is adopted to change the casing of a token if necessary, for example in a capitalized title in which not every token is a proper noun. The rules that are applied are language-specific and based on the Part-of-Speech tag of a token. In German, for example, all nouns (common as well as proper) start with a capital letter, but prepositions should be lowercased. Some PoS categories can contain upper- as well as lowercased word forms. This is the case for adjectives in Dutch and English. In general, these adjectives are lowercased (e.g. *short*), but when derived from a proper noun, they usually retain a capital first letter (e.g. *English*). For these Part-of-Speech categories, the rule-based approach is supported by a lexicon of word forms listed with their grammatical category and corresponding casing.

Step 2: Mapping the Part-of-Speech tags of the tokens

The Part-of-Speech tags assigned by the LeTs Taggers (see Section 2.2) are more fine-grained than the grammatical categories used as features in the training lexicons for Lemmatization. Before the CRF++ models trained on these lexicons can be successfully applied to the output of the LeTs Part-of-Speech Taggers, mapping of the PoS tags is necessary.

Step 3: Labeling the tokens with lemma tags

At this point, the actual Lemmatization process starts, which makes use of CRF++ models trained on the lemmatized lexicons described above.

In order to tackle Lemmatization as a classification task, the lemmas in the training lexicons needed to be converted to class tags. Treating each individual lemma as a separate class tag would not have been a valid solution, since the number of possible output tags would be too high for CRF++ to handle, and the generated models would not be able to predict the lemma for unseen words. Since Lemmatization in Dutch, English, French and German mainly involves removing and/or adding characters at the end of a word form, the lemmas in the training lexicons were converted to lemma tags specifying the suffixes to be removed and/or added in order to lemmatize the word form at hand. A similar, but slightly more complex approach has been followed by Chrupała (2006) for several languages, including Dutch, French and German. What follows is an example of an English lexicon entry, containing a word form with its PoS category and corresponding lemma tag.

satisfied VBD +Died+Iy

In order to lemmatize this token, the characters preceded by *+D* (*delete*) are to be removed from the end of the word form, after which the characters preceded by *+I* (*insert*) need to be added. This results in the lemma *satisfy*. If a word form equals its lemma, it is assigned the lemma tag */*. These are the same tags used in MBLEM (van den Bosch and Daelemans 1999). Irregular word forms are assigned unique lemma tags, for example the lemma tag of the English verb form *is*:

is VBZ +Dis+Ibe

Table 2 lists the number of (unique) lexicon entries and lemma tags in the training data for Lemmatization based on the removal and addition of suffixes.

	number of tokens	number of lemma tags
Dutch	206,036	1,092
English	54,239	416
French	460,455	1,475
German	352,445	2,212

Table 2: Properties of the training lexicons for Lemmatization through suffix removal and addition

In these training lexicons, each token is represented by a feature vector consisting of the elements listed below:

1. **basic features**
   - the original form of the token
   - the (reduced) PoS tag of the token
2. **suffix features**
   - suffixes consisting of the last 1 to 8 characters of the token

To train machine learning models for Lemmatization based on these data, the c parameter provided by CRF++ was set to 3 for all four languages. This parameter trades the balance between overfitting and underfitting. Depending on the number of tokens and lemma tags in the training data, the f parameter, which sets the cut-off threshold for the features, was set to 1 for English, 3 for Dutch, 4 for French and 5 for German. Using these settings, we trained CRF++ on the lexicons listed in Table 2 and created models which can be applied to lemmatize Dutch, English, French and German word forms through suffix removal and addition. For the vast majority of tokens in English and French, such a model is sufficient for accurate Lemmatization. In Dutch and German however, Lemmatization sometimes requires removing characters at the beginning or in the middle of a word. This is the case for many past participles (in both Dutch and German) and all 'zu' infinitives[5] (only in German). For example, the correct lemmatized form of the Dutch past participle *lesgegeven* (*taught*) is obtained by removing the infix *ge*. For these grammatical categories, two additional CRF++ models were created, viz. one for Dutch and one for German. To train these models, the lemma of each past participle and 'zu' infinitive in the Dutch and German lexicons described above was converted to a lemma tag specifying the prefix or infix to be removed from the word form. For example, the lexicon entry for *lesgegeven* looks as follows:

lesgegeven WW(vd,zonder) 3ge

The lemma tag at the end of the lexicon entry specifies which characters are to be removed from the word form and at which position in the string (starting from 0). Applying this operation to the past participle *lesgegeven* results in the lemma *lesgeven*. Again, the lemma tag / is used for word forms that require no (prefix or infix) alterations. Table 3 shows the size of the lexicons used to train the CRF++ models that perform Lemmatization through prefix and infix removal.

In these training lexicons, each token is represented by the following features:

5. An example of a 'zu' infinitive is the word form *auszuführen* (*to carry out*, *to take out*, *to export*), which is to be lemmatized as *ausführen*.

	number of tokens	number of lemma tags
Dutch	19,026	16
German	10,312	29

Table 3: Properties of the training lexicons for Lemmatization through prefix and infix removal

1. **basic features**
   - the original form of the token
   - the (reduced) PoS tag of the token
2. **prefix features**
   - prefixes consisting of the first 1 to 10 characters of the token

When training CRF++ on these data, the c parameter for over/underfitting was set to 3 for both languages.

For some past participles, Lemmatization involves the removal of a prefix or infix as well as the removal and/or addition of a suffix. Both CRF++ models are therefore applied to tokens belonging to this grammatical category, after which the output of the two models is combined. The German past participle *geändert* (*changed*), for example, is assigned the lemma tag *+Dt+In* by the CRF++ model for Lemmatization through suffix addition/removal, and receives the label *0ge* from the Lemmatization model for prefix and infix removal. Performing the operations specified in both lemma tags on the token results in the lemma *ändern*.

Finally, we note that in some PoS categories, tokens never require Lemmatization. This is the case for example for singular nouns in English (*NN*), interjections, punctuation and symbols, etc. Word forms belonging to these categories are not included in the training data, and are assigned an identical lemma.

Note that although Conditional Random Fields are aimed at sequential tagging tasks, we chose to apply CRF++ to Lemmatization as well so that only one classification framework is used throughout the entire LeTs Preprocess pipeline. In the Lemmatization process described above, no sequential tagging takes place since each word is treated separately and no context features are used.

Step 4: Deriving the tokens' lemmas from the lemma tags

In a final step, the lemma of a token is determined by removing and/or adding the characters specified in the lemma tag(s) assigned to the token in step 3, unless the token belongs to a grammatical category for which Lemmatization is not necessary. In the same process, we also add or remove punctuation marks and diacritics if required by the spelling conventions of the language at hand. In Dutch, for example, a diaeresis is sometimes used to indicate that two vowels should be pronounced separately, e.g. in the past participle *geëgaliseerd* (*evened*, *smoothed*). When determining the lemma for this word form based on its lemma tags, the prefix *ge* is removed and the diaeresis on the second *e* is discarded as well, resulting in the lemma *egaliseren*.

2.4 Named Entity Recognition

Named Entity Recognition is aimed at the extraction of names (e.g. of persons, organizations) from text and is often treated as a classification task. Several annotated corpora are available in which tokens are assigned a tag specifying whether a token is part of a named entity and, if so, the exact type of that named entity. For example, the tags *B-ORG* and *I-ORG* denote that a token is part of an organization name. The tag *O* is used for tokens that are not contained in a named entity. The possible formats for named entity annotation are discussed at the end of this section.

For German and English, we made use of the CoNLL-2003 shared task data for Named Entity Recognition (Tjong Kim Sang and De Meulder 2003). These data contain newspaper text in which four types of named entities are annotated: person (*PER*), location (*LOC*) and organization names (*ORG*), and miscellaneous named entities (*MISC*). For French, a new training corpus for NER was created by manually annotating a subset of the French Treebank (Abeillé et al. 2003) using the CoNLL guidelines for named entity annotation[6]. Finally, the SoNaR 1-million-word corpus (Schuurman et al. 2009) served as the training data for Dutch. This corpus contains annotations of the same types of NEs covered in the CoNLL shared task and two additional categories: event (*EVE*) and product names (*PRO*) (Desmet and Hoste 2010b). Table 4 presents an overview of the corpora used to train the LeTs NER models. These models take as input tokenized text labeled with Part-of-Speech tags from the tagsets used by the LeTs PoS Taggers (see Section 2.2). Note that the PoS tags assigned by the LeTs taggers are more fine-grained than the grammatical categories present in the training data for Named Entity Recognition. Before the CRF++ models trained on these data can be successfully applied to the output of the LeTs Part-of-Speech Taggers, mapping of the PoS tags is necessary.

	number of tokens	number of NE tags	number of named entities
Dutch	1,000,437	13	62,643
English	203,621	9	23,499
French	188,355	9	7,933
German	206,931	9	11,851

Table 4: Properties of the training corpora for Named Entity Recognition

The selection of features for Named Entity Recognition was based on the research conducted by Desmet and Hoste (2010a). The same orthographic features extracted for Part-of-Speech Tagging (see Section 2.2) are used again for NER. Furthermore, the following features were added to the feature vectors of the tokens for all four languages:

1. **basic features**
   - the original form of the token
   - the (reduced) PoS tag of the token.
     The training data taken from the French Treebank for French and the CoNLL datasets for English and German already contained PoS tags. For the Dutch training corpus, PoS tags were generated by the LeTs Part-of-Speech Tagger and mapped to the main word classes of the tokens (e.g. *N* for nouns).
   - the number of characters in the token
2. **orthographic features**
   - the word shape of the token.
     This feature is based on the orthographic features described in Section 2.2 and outputs one of the following labels: *allLowercase*, *allCaps*, *firstCap*, *capPeriod*, *onlyDigits*, *containsDigitAndAlpha*, *allCapsAndPunct*, *firstCapAlphaAndPunct*, *alphaAndPunct*, *onlyPunct*, *mixedCase* or *other*. It is included to force feature conjunction of orthographic information.
3. **affix features**
   - prefixes consisting of the first three and four characters of the token
   - suffixes consisting of the last three and four characters of the token

6. http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt

Training CRF++ on the corpora described above resulted in a Named Entity Recognition model for each of the four supported languages. Note that the training data and thus the NER models make use of an annotation format for Dutch and French that is different from that for English and German. In Dutch and French, the IOB2 format is used, which labels the first token of a named entity as *B-XXX* and the other tokens of the NE as *I-XXX*, with *XXX* being the type of the named entity (e.g. *ORG* for organization names). The English and German Named Entity Recognizers employ the IOB1 annotation scheme and always tag tokens contained in a named entity as *I-XXX*, unless confusion with other NEs occurs because they appear next to each other (Tjong Kim Sang 2002a). To attain consistency between the different languages, the named entity tags assigned by the English and German models are mapped to the IOB2 annotation format.

3. Multi-domain evaluation

3.1 Experimental setup

To evaluate the LeTs Preprocess components described above, we first performed ten-fold cross-validation on the training corpora. In real-world scenarios however, preprocessing tools are often applied to domains and text types not covered by the training data. In order to assess the performance of the linguistic preprocessing modules on different domains, we collected user-specific texts for Dutch, English, French and German from companies covering various domains, namely manuals from an automotive company (PSA Peugeot Citroën), annual reports from a dredging company (Jan De Nul) and manuals for human resources software (SDWorx)[7]. Furthermore, we added financial news articles for Dutch (from De Tijd) and English (from the Financial Times) to the multi-domain evaluation corpus. Note that, unlike the data collected for the other domains, these are not parallel texts. An overview of the corpus size per domain per language can be found in Table 5. Each text was first converted to UTF-8, split into sentences and tokenized using the LeTs components described in Section 2.

	Dutch	English	French	German
dredging	8,893	9,761	10,750	8,707
automotive	4,304	4,247	4,637	3,725
human resources	7,624	7,657	8,331	N/A
financial news	5,053	5,108	N/A	N/A
Total	25,874	26,773	23,718	12,432

Table 5: Number of tokens in the multi-domain evaluation corpus

This multi-domain corpus was enriched with gold standard Part-of-Speech tags, lemmas and named entity tags. Each text in the corpus was processed with the three LeTs Preprocess modules and subsequently the output was manually corrected based on the guidelines used to create the training data discussed in Section 2. To assess the performance of the LeTs components, their output is compared to the manually verified gold standard. For Part-of Speech Tagging and Lemmatization, we also compare the results of the LeTs Preprocess Toolkit to the performance of two frequently used systems, namely Frog (van den Bosch et al. 2007) (version 0.12.13) for Dutch and TreeTagger (Schmid 1994, Schmid 1995) (version 3.2) for English, French and German. Although TreeTagger also supports Dutch, we chose to use Frog for this language because it also uses the D-Coi/CGN PoS tagset, which is considered the standard for Dutch.

7. The corpus we received from SDWorx was not translated into German.

3.2 Results and discussion

This section presents the ten-fold cross-validation results for all three LeTs preprocessing components on the training data and gives a detailed overview of their performance on the multi-domain evaluation corpus. For Part-of-Speech Tagging and Lemmatization, a comparison is made with the output of TreeTagger[8] and Frog. Note that the output of the Frog Part-of-Speech Tagger was first mapped to the reduced Dutch tagset used by the LeTs system (see Section 2.2). For our experiments, we made use of the standard preprocessing evaluation metrics, namely accuracy for Part-of-Speech Tagging and Lemmatization, and precision, recall and F-score for Named Entity Recognition. In order to calculate the latter three measures, we used the standard evaluation script of Tjong Kim Sang and De Meulder (2003), which has been used in two shared tasks for Named Entity Recognition (Tjong Kim Sang 2002a, Tjong Kim Sang and De Meulder 2003) and makes use of the FB1 score formula as defined by van Rijsbergen (1975).

Table 6 lists the accuracy scores for Part-of-Speech Tagging, while Table 7 gives an overview of the Lemmatization results. For both tasks, the rules adopted by TreeTagger and Frog sometimes differ from the conventions used in the LeTs training data. In French, for example, TreeTagger assigns ordinal numbers (e.g. *deuxième*, meaning *second*) the PoS tag *NUM* for numerals. In the training corpus of the LeTs Part-of-Speech Tagger for French however, words of this category are labeled *ADJ*, since they are used as adjectives. An example of a Part-of-Speech category for which different Lemmatization rules are applied is the German *ART* category for articles. LeTs always reduces tokens of this PoS class to the singular male form, whereas TreeTagger assigns the singular female form as the lemma. In these cases, the output of both systems can be considered correct, depending on the Part-of-Speech Tagging/Lemmatization conventions preferred by the user. To enable a fair comparison of LeTs and TreeTagger/Frog, strict as well as relaxed accuracy scores are calculated for each system. Relaxed accuracy is calculated by allowing several PoS tags or lemmas in cases where the two systems adopt different rules. In Tables 6 and 7 below, the relaxed accuracy results can be found between brackets. Note that for Lemmatization, ten-fold cross-validation was performed on the training data for Lemmatization through suffix removal/addition and prefix/infix removal. These numbers give no indication of the performance of the rule-based components used to change the casing of the tokens or add/remove punctuation marks and diacritics if necessary.

	10-fold CV on training data	dredging year reports	automotive manuals	human resources software manuals	financial news
Dutch					
LeTs	96.59	97.27 (97.44)	93.38 (93.38)	96.77 (96.83)	96.75 (97.13)
Frog	N/A	91.70 (94.87)	90.75 (90.75)	95.80 (96.07)	89.91 (95.44)
English					
LeTs	97.35	96.30	95.20	94.82	96.79
TreeTagger	N/A	94.15	90.18	92.67	95.61
French					
LeTs	96.84	97.57 (97.58)	95.79 (95.79)	97.34 (97.35)	N/A
TreeTagger	N/A	96.08 (96.34)	85.08 (85.10)	88.34 (88.46)	
German					
LeTs	97.64	97.45	95.73	N/A	N/A
TreeTagger	N/A	96.19	92.56		

Table 6: Part-of-Speech accuracy scores (in %) for LeTs (Dutch, English, French and German), Frog (Dutch) and Treetagger (English, French and German) on the multi-domain test corpus, preceded by 10-fold cross-validation figures for the LeTs system on the training data

8. For tokens assigned multiple lemmas by TreeTagger, only the first one was retained for the evaluation experiments.

	10-fold CV on training data	dredging year reports	automotive manuals	human resources software manuals	financial news
Dutch					
LeTs	95.21	98.53 (98.53)	97.96 (97.96)	97.85 (97.85)	98.59 (98.59)
Frog	N/A	95.74 (96.63)	96.93 (97.00)	97.14 (97.89)	96.02 (97.21)
English					
LeTs	97.58	98.46 (98.47)	97.79 (97.98)	96.49 (96.49)	98.69 (98.73)
TreeTagger	N/A	97.11 (97.49)	96.40 (96.47)	94.92 (94.98)	98.24 (98.88)
French					
LeTs	97.67	98.82 (98.82)	97.99 (97.99)	98.60 (98.60)	N/A
TreeTagger	N/A	98.03 (98.10)	97.41 (97.41)	93.19 (93.19)	
German					
LeTs	94.65	98.66 (98.74)	98.20 (98.28)	N/A	N/A
TreeTagger	N/A	82.66 (97.26)	74.01 (94.50)		

Table 7: Lemmatization accuracy scores (in %) for LeTs (Dutch, English, French and German), Frog (for Dutch) and Treetagger (English, French and German) on the multi-domain test corpus, preceded by 10-fold cross-validation figures for the LeTs system on the training data

When applying the LeTs Part-of-Speech Taggers and Lemmatizers to the multi-domain gold standard corpus, we see that the achieved results range from 93.38 to 97.57 for PoS Tagging and from 96.49 to even 98.82 for Lemmatization. Furthermore, Frog and TreeTagger are outperformed for almost every domain in each of the four languages, even when only taking into account the relaxed accuracy figures. Only for the Dutch human resources dataset and the English financial news corpus, slightly higher relaxed Lemmatization scores are obtained by resp. Frog and TreeTagger. We also observe that the Part-of-Speech Tagging scores for French are comparable to those of the three other languages, even though a much smaller amount of training data was used (see Section 2.2). A possible explanation can be the fact that a smaller tagset was used, which makes it easier for the classifier to learn and predict the correct Part-of-Speech tag for unseen instances.

Overall, the best performance is obtained on the financial news articles and the dredging corpus. A possible explanation for this is that the text types in these corpora (viz. newswire text from De Tijd and the Financial Times, and annual reports of a Belgian dredging company) are most similar to the LeTs training data. The LeTs tools yield the lowest scores when applied to the automotive corpus and, for English, the human resources dataset. Both corpora contain instructive text: the former is made up of text strings used to compile user manuals for the automotive domain, while the latter contains instructions related to human resources software. The sentence structures used in these kinds of texts are often very different from the (mostly) general-domain text in the training data; this decreases performance for Part-of-Speech Tagging in particular. The following example is a Dutch text string taken from the automotive corpus: *binnenpaneel dakversteviging achter* (*rear roof arch lining*). The tokens in this compound are to be labeled as resp. a noun, a noun, and a preposition. However, contrary to English, sequences of nouns are not common in Dutch (they are usually composed as one orthographic unit) and by consequence rarely occur in the training data for Part-of-Speech Tagging. As the LeTs tagger expects a different type of construction, it labels the tokens in the compound as resp. an adjective, noun and preposition. Erroneous tagging caused by uncommon compound, phrase and sentence structures is one of the most frequent errors made by the LeTs Part-of-Speech Taggers in each of the four languages. Another problem is the incorrect labeling of capitalized tokens as proper nouns, e.g. when they appear in an uppercased title. These errors occur frequently in the human resources dataset. Finally, a big challenge for Part-of-Speech Tagging in English is distinguishing between gerunds (*VBG*) and past participles (*VBN*) on the one hand, and gerunds and past participles used as adjectives (*JJ*) or sometimes nouns (*NN*) on the

other hand. In the dredging corpus, for example, the gerund *dredging* is often used as a noun, but incorrectly labeled as a gerund by the LeTs tagger. The distinction between verbs and verbs used as adjectives or nouns is difficult to make even for humans, as we see that the annotation of such tokens in the DPC and the Penn Treebank has not always been performed consistently either. For this reason, the Penn Treebank guidelines allow annotators to assign multiple Part-of-Speech tags to one token if necessary. A similar problem is identified in French, where LeTs can often not correctly distinguish between past and present participles (resp. *VER:pper* and *VER:ppre*) and participles used as adjectives (*ADJ*). These types of errors are a possible explanation for the fact that the LeTs Part-of-Speech Tagger for Dutch obtains scores close to those of the English and French taggers, even though the tagset is much more fine-grained. In the D-Coi tagset, past and present participles for example are always labeled as verbs, even when used as nouns or adjectives. This makes for more consistent PoS annotations, from which the classification algorithm can learn more easily.

Qualitative analysis of the output of the LeTs Lemmatizers results in the following findings. Most Lemmatization mistakes are the result of erroneous Part-of-Speech Tagging. In English, for example, the comparative adjective *lower* (tagged as *JJR*) is to be assigned the lemma *low*. However, when this word is assigned the incorrect PoS tag *VB* for base verb forms, it is lemmatized as *lower* by LeTs. When comparing the performance of the LeTs Part-of-Speech Taggers and Lemmatizers on the multi-domain corpus, we see that a higher Part-of-Speech accuracy is often associated with a higher lemma accuracy. Another observation we can make is that Lemmatization scores could be improved by using more fine-grained PoS tags as a feature. We think this would benefit the Lemmatization of German nouns in particular. *Montage*, for example, can be a singular noun (meaning *assembly*) as well as a plural noun (meaning *Mondays*). However, the Part-of-Speech tag used for nouns (*NN*) does not specify their number. Lacking this information, the LeTs Lemmatizer cannot disambiguate between the lemma of the singular noun (= *Montage*) and that of the plural noun (= *Montag*). Finally, we observe that even for French and English, suffix addition and removal is not always sufficient for accurate Lemmatization. The plural French noun *navires-citernes* (*tankers*), for example, should be lemmatized as *navire-citerne*, but is assigned the lemma *navires-citerne* in the output of the LeTs Lemmatizer. Such errors occurred for only a small number of tokens.

As we mentioned above, LeTs outperforms both TreeTagger and Frog for almost every domain in each of the four languages. For Part-of-Speech Tagging, the types of mistakes discussed above are also made by these two systems. LeTs obtains better results when tagging capitalized tokens which are not proper nouns (e.g. in uppercase titles). This might be due to the sentence-level features we use for Part-of-Speech Tagging (see Section 2.2), which specify whether every token in a sentence is capitalized and thus help predict whether a sentence is a type of title. LeTs also achieves better results than TreeTagger for certain structures typical to instructive texts (e.g. the use of imperative verb forms). One could argue that this is because the Part-of-Speech training data also contains these types of texts (see Section 2.2). This is true for English and French, however not for German. On the other hand, some errors typical of instructive text (e.g. the incorrect PoS Tagging of uncommon compound structures in Dutch discussed above) are made more frequently by the LeTs system. A possible explanation for the higher Lemmatization scores of LeTs is that Frog and TreeTagger do not always seem to disambiguate between multiple possible lemmas of a token based on its Part-of-Speech tag. The Dutch simple past verb form *verbonden* (*connected*), for example, is lemmatized by Frog as *verbond*, which is the lemma of the plural noun *verbonden* (*pacts*, *treaties*). Note that the large difference between the strict accuracy scores of Frog and LeTs for Part-of-Speech Tagging of the Dutch dredging corpus and the financial news dataset is caused by the fact that Frog always assigns proper nouns the tag *SPEC(deeleigen)*. According to the CGN guidelines for PoS Tagging however, this tag should only be used for proper nouns that are part of a multi-word named entity. This influences the strict accuracy scores for Frog in particular for the dredging and financial news corpora because these texts contain a large amount of proper nouns. The low strict accuracy scores of TreeTagger for German can be explained by the fact that for this language, the difference between

the Lemmatization conventions used by LeTs and TreeTagger is substantial.

In Table 8, the precision, recall and F-score figures for the LeTs Named Entity Recognition modules are presented. We also calculated relaxed scores for Named Entity Recognition, meaning that we did not take into account the types assigned to the named entities by the LeTs NER modules. In doing so, we can measure the capability of the system to correctly identify NEs, irrespective of its performance in classifying these NEs. In the table below, the relaxed NER scores can be found between brackets.

	10-fold CV on training data	dredging year reports	automotive manuals	human resources software manuals	financial news
Dutch					
precision	88.54 (95.38)	58.02 (94.44)		8.44 (13.64)	71.84 (96.12)
recall	87.63 (94.39)	59.49 (96.84)	N/A	46.43 (75.00)	70.81 (94.74)
F-score	88.08 (94.88)	58.75 (95.63)		14.29 (23.08)	71.33 (95.42)
English					
precision	91.52 (95.75)	59.33 (86.25)		7.45 (13.66)	63.74 (88.28)
recall	90.40 (94.58)	62.92 (91.46)	N/A	44.44 (81.48)	61.27 (84.86)
F-score	90.96 (95.16)	61.07 (88.78)		12.77 (23.40)	62.48 (86.54)
French					
precision	87.78 (94.11)	56.52 (88.56)		7.69 (20.00)	
recall	85.05 (91.18)	53.58 (83.95)	N/A	32.26 (83.87)	N/A
F-score	86.39 (92.62)	55.01 (86.19)		12.42 (32.30)	
German					
precision	85.68 (89.97)	60.53 (89.71)	6.67 (6.67)		
recall	74.90 (78.64)	54.18 (80.30)	100.00 (100.00)	N/A	N/A
F-score	79.93 (83.93)	57.18 (84.75)	12.50 (12.50)		

Table 8: Precision, recall and F-score figures (in %) for LeTs for all four supported languages (Dutch, English, French and German), preceded by 10-fold cross-validation figures on the training data

The results of the multi-domain evaluation for Named Entity Recognition show that LeTs produces varying results depending on the domain and language. The highest scores are obtained when processing the financial news text corpora. These are the datasets most similar to the corpora used to train the Named Entity Recognition models. The very low scores for the human resources and automotive corpus can partly be explained by the fact that these texts contain only a very small number of named entities. Even if only a few tokens are incorrectly labeled as being part of a named entity, the precision decreases significantly. The automotive corpus, for example, does not even contain a single named entity for Dutch, English and French; this automatically results in recall and precision scores of 0.00. For these datasets, the metrics precision and recall give no indication of the performance of the LeTs Named Entity Recognition modules. The scores were therefore replaced with the label *N/A*.

For the other datasets, most errors made by LeTs involve the assignment of a named entity to the wrong category. In the dredging corpus, for example, an overview is given of the activities conducted by the dredging company in question in the past year. Many of the dredging vessels used by the company are named after explorers, geographers, engineers and other persons (e.g. *Vasco da Gama*) and are labeled by LeTs as named entities of the category *PER* (for persons) instead of *MISC* (for named entities of the miscellaneous category such as boat names). Even though the LeTs system does not always correctly classify these NEs, it is able to correctly identify them, as can be seen from the high relaxed scores in Table 8. Other frequently made errors involve the labeling of capitalized titles or abbreviations as named entities (e.g. *ID*).

4. Conclusions and future work

In this paper we presented the LeTs Preprocess Toolkit, a suite of linguistic preprocessing modules. The toolkit includes Part-of-Speech Taggers, Lemmatizers and Named Entity Recognizers for four languages, viz. Dutch, English, French and German. These modules were developed by means of a supervised machine learning method using the Conditional Random Fields implementation CRF++. We discussed the data and methods used to train each component and presented ten-fold cross-validation results on the training data.

To assess the performance of the LeTs tools on different domains, we constructed a gold standard evaluation corpus for Part-of-Speech Tagging, Lemmatization and Named Entity Recognition. This corpus covers various domains, a.o. dredging, automotive and human resources. The LeTs system was applied to this corpus and its performance was compared to that of Frog (for Dutch) and TreeTagger (for English, French and German). We showed that the LeTs Part-of-Speech Taggers and Lemmatizers obtain state-of-the-art results and can be successfully applied to different domains. Named Entity Recognition scores are lower, but a large part of the errors made by LeTs involve only the incorrect classification of correctly identified named entities.

Both the LeTs Preprocess Toolkit and the annotated multi-domain corpus described in Section 3.1 will be made available for academic use.

The LeTs Preprocess Toolkit has been developed in the framework of the TExSIS system for terminology extraction based on linguistic and statistical information, but it can also be used for other types of linguistic processing. In future work, we will investigate the impact of preprocessing quality on the performance of the TExSIS terminology extraction system. For this purpose, domain-specific terminology was manually annotated in the multi-domain evaluation corpus discussed above. The availability of gold standard preprocessing as well as term annotations for the same corpus, allows us to measure the influence of the performance of each preprocessing step on the output of the terminology extractor.

Acknowledgements

This research was conducted in the framework of the IWT-TETRA funded TExSIS project (Terminology Extraction for Semantic Interoperability and Standardization, IWT code 100144). Part of the experimental work for this project was carried out using the STEVIN Supercomputer Infrastructure at Ghent University, funded by Ghent University, the Flemish Supercomputer Center (VSC), the Hercules Foundation and the Flemish Government - department EWI. We would like to thank the members of the TExSIS user committee for providing us with data for the multi-domain evaluation corpus. We also want to thank the reviewers for their valuable comments and suggestions.

References

Abeillé, A., L. Clément, and F. Toussenel (2003), Building a treebank for French, *Treebanks: Building and Using Parsed Corpora*, Kluwer, Dordrecht, pp. 165–188.

Baayen, R.H., R. Piepenbrock, and L. Gulikers (1995), The CELEX lexical database (CD-ROM).

Brants, S., S. Dipper, S. Hansen, W. Lezius, and G. Smith (2002), The TIGER treebank, *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria, pp. 24–41.

Chrupała, G. (2006), Simple data-driven context-sensitive lemmatization, *Proceedings of the Sociedad Española para el Procesamiento del Lenguaje Natural, volume 37*, Zaragoza, Spain, pp. 121–130.

Crysmann, B., S. Hansen-Schirra, G. Smith, and D. Ziegler-Eisele (2005), TIGER Morphologie-Annotationsschema, *Technical report*, Universität des Saarlandes - Computerlinguistik, Univer-

sität Stuttgart - Institut für Maschinelle Sprachverarbeitung, Universität Potsdam - Institut für Germanistik.

Daelemans, W., J. Zavrel, P. Berck, and S. Gillis (1996), MBT: A memory-based part of speech tagger-generator, *Proceedings of the Fourth Workshop on Very Large Corpora*, Copenhagen, Denmark, pp. 14–27.

Desmet, B. and V. Hoste (2010a), Dutch named entity recognition using classifier ensembles, *Computational Linguistics in the Netherlands 2010: selected papers from the twentieth CLIN meeting*, Netherlands Graduate School of Linguistics, pp. 29–41.

Desmet, B. and V. Hoste (2010b), Towards a balanced named entity corpus for Dutch, *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, pp. 535–541.

Giménez, J. and L. Màrquez (2004), SVMTool: A general POS tagger generator based on Support Vector Machines, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, pp. 43–46.

Hemphill, C.T., J.J. Godfrey, and G.R. Doddington (1990), The ATIS Spoken Language Systems pilot corpus, *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 96–101.

Isozaki, H. and H. Kazawa (2002), Efficient Support Vector Classifiers for named entity recognition, *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, Taipei, Taiwan, pp. 1–7.

Koehn, P. (2005), Europarl: A parallel corpus for statistical machine translation, *Proceedings of the tenth Machine Translation Summit*, Phuket, Thailand, pp. 79–86.

Kupiec, J. (1992), Robust part-of-speech tagging using a hidden Markov model, *Computer Speech and Language* **6** (3), pp. 225–242.

Lafferty, J., A. McCallum, and F. Pereira (2001), Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proceedings of the eighteenth International Conference on Machine Learning (ICML'01)*, Williamstown, Massachusetts, USA, pp. 282–289.

Macken, L. and W. Daelemans (2010), A chunk-driven bootstrapping approach to extracting translation patterns, *Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics (Iaşi, Romania)*, Vol. 6008 of *Lecture Notes in Computer Science*, Springer-Verlag, Berlin Heidelberg, pp. 394–405.

Macken, L., E. Lefever, and V. Hoste (2013), TExSIS: bilingual terminology extraction from parallel corpora using chunk-based alignment, *Terminology* **19** (1), pp. 1–30, John Benjamins Publishing Company.

Marcus, M.P., B. Santorini, and M.A. Marcinkiewicz (1993), Building a large annotated corpus of English: The Penn Treebank, *Computational Linguistics* **19** (2), pp. 313–330.

McCallum, A. and W. Li (2003), Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, Morristown, New Jersey, USA, pp. 188–191.

Paroubek, P. (2000), Language resources as by-product of evaluation: the MULTITAG example, *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, pp. 151–154.

Paulussen, H., L. Macken, W. Vandeweghe, and P. Desmet (2013), Dutch Parallel Corpus: a balanced parallel corpus for Dutch-English and Dutch-French, *Essential Speech and Language Technology for Dutch* pp. 185–199, Springer.

Romary, L., S. Salmon-Alt, and G. Francopoulo (2004), Standards going concrete: from LMF to Morphalou, *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries - COLING-2004*, Geneva, Switzerland, pp. 22–28.

Sagot, B. (2010), The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French, *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Santorini, B. (1990), Part-of-speech tagging guidelines for the Penn Treebank Project, *Technical report*, University of Pennsylvania, Department of Computer and Information Science, Philadelphia, Pennsylvania, USA.

Schiller, A., S. Teufel, C. Stöckert, and C. Thielen (1999), Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset), *Technical report*, Universität Stuttgart - Institut für maschinelle Sprachverarbeitung, Universität Tübingen - Seminar für Sprachwissenschaft.

Schmid, H. (1994), Probabilistic part-of-speech tagging using decision trees, *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, pp. 44–49.

Schmid, H. (1995), Improvements in part-of-speech tagging with an application to German, *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland, pp. 47–50.

Schuurman, I., V. Hoste, and P. Monachesi (2009), Cultivating trees: Adding several semantic layers to the Lassy treebank in SoNaR, *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories*, Groningen, The Netherlands, pp. 135–146.

Skut, W., T. Brants, B. Krenn, and H. Uszkoreit (1998), A linguistically interpreted corpus of German newspaper text, *Proceedings of the Tenth European Summer School in Logic, Language and Information (ESSLLI'98). Workshop on Recent Advances in Corpus Annotation*, Saarbrücken, Germany, pp. 705–711.

Smith, G. (2003), A brief introduction to the TIGER treebank, version 1, *Technical report*, Universität Potsdam.

Telljohann, H., E. Hinrichs, and S. Kübler (2004), The Tüba-D/Z treebank: annotating German with a context-free backbone, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, pp. 2229–2235.

Tjong Kim Sang, E.F. (2002a), Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition, *Proceedings of the 6th Conference on Natural Language Learning (COLING'02)*, Taipei, Taiwan, pp. 155–158.

Tjong Kim Sang, E.F. (2002b), Memory-based named entity recognition, *Proceedings of CoNLL-2002*, Taipei, Taiwan, pp. 203–206.

Tjong Kim Sang, E.F. and F. De Meulder (2003), Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, *Proceedings of CoNLL-2003*, Edmonton, Canada, pp. 142–147.

van den Bosch, A. and W. Daelemans (1999), Memory-based morphological analysis, *Proceedings of the 37th annual meeting of the Association for Computational Linguistics (ACL'99)*, University of Maryland, USA, pp. 285–292.

van den Bosch, A., B. Busser, W. Daelemans, and S. Canisius (2007), An efficient memory-based morphosyntactic tagger and parser for Dutch, *Computational Linguistics in the Netherlands 2006*, Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting, Leuven, Belgium, pp. 99–114.

Van Eynde, F. (2005), Part of speech tagging en lemmatisering van het D-Coi corpus, *Technical report*, Centrum voor Computerlinguïstiek, KU Leuven.

van Gompel, M. (2012), FoLiA: Format for Linguistic Annotation. Documentation. ILK Technical Report 12-03, *Technical report*, Center for Language Studies, Radboud University Nijmegen. http://ilk.uvt.nl/downloads/pub/papers/ilk.1203.pdf.

van Noord, G., G. Bouma, F. Van Eynde, D. de Kok, J. van der Linde, I. Schuurman, E. Tjong Kim Sang, and V. Vandeghinste (2009), Large scale syntactic annotation of written Dutch: Lassy, *Essential Speech and Language Technology for Dutch*, Springer Berlin Heidelberg, pp. 147–164.

van Rijsbergen, C.J. (1975), *Information Retrieval*, Buttersworth, London.

Zhou, GD. and J. Su (2002), Named entity recognition using an HMM-based Chunk Tagger, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*, Philadelphia, USA, pp. 473–480.