Code-Switching Ubique Est - Language Identification and Part-of-Speech Tagging for Historical Mixed Text

Sarah Schulz
Institute for Natural Language Processing (IMS)
University of Stuttgart
70569 Stuttgart, Germany
schulzsh@ims.uni-stuttgart.de

Mareike Keller
Institute for English Linguistics
University of Mannheim
60131 Mannheim, Germany
markelle@mail.uni-mannheim.de

Abstract

In this paper, we describe the development of a language identification system and a part-of-speech tagger for Latin-Middle English mixed text. To this end, we annotate data with language IDs and Universal POS tags (Petrov et al., 2012). As a classifier, we train a conditional random field classifier for both sub-tasks, including features generated by the TreeTagger models of both languages. The focus lies on both a general and a task-specific evaluation. Moreover, we describe our effort concerning beyond proof-of-concept implementation of tools and towards a more task-oriented approach, showing how to apply our techniques in the context of Humanities research.

1 Introduction

Code-switching is often described as a phenomenon highly frequent in spoken language. In today's multi-cultural society, addressing mixed language in natural language processing appears to be inevitable, as the development of methods close to real-world data touches a nerve in recent computational linguistics. Especially social media as a form of written language close to spontaneous speech has recently been focused on code-switching research (e.g. Das and Gambäck (2013)).

However, code-switching is not just a recent phenomenon but can already be observed in medieval writing. As has been pointed out in several studies (Wenzel, 1994; Schendl and Wright, 2012; Jefferson et al., 2013), historical mixed text is an interesting, yet still widely unexplored, source of information concerning language use in multilingual societies of Medieval Europe. Even though some studies use text corpora in order to qualitatively describe the phenomenon (cf. Nurmi and Pahta (2013)), a deeper analysis of the underlying structures has not been carried out due to the lack of adequate resources.

In order to pave the way for an in-depth corpus-based analysis, we promote the systematic annotation of resources and concentrate on developing and implementing automatic processing tools. To this end, combining forces from Humanities and Computer Science seems promising for both sides. As an additional challenge, joint work in this context and with a specific purpose in mind does not just require the developing proof-of-concept tools. We need to tackle the issue of how to make tools available to Humanities scholars. Consequently, we do not just focus on developing techniques for automatic processing but also take into consideration how to share tools and make them useful for interpreting and analyzing data.

For the project presented in this study, we annotate Macaronic sermons (Horner, 2006)[1] with language information and part-of-speech (POS), respectively and use this resource to develop tools for automatic language identification (LID) on the word level and POS tagging of mixed Latin-Middle English text. The resulting tools allow for the automatic annotation of larger quantities of text and thus for the investigation of code-switching constraints within specific syntactic constructions on a larger scale. In particular, we aim at an analysis of code-switching rules within nominal phrases.

In the following example, determiner and modifier (*þe briʒt / the bright*) are written in Middle English whereas the head of the noun

[1]We are greatly endebted to the Pontifical Institute of Mediaeval Studies (PIMS), Toronto, for their support and kind permission to use a searchable PDF version of the sermon transcripts.

Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 43–51,
Berlin, Germany, August 11, 2016. ©2016 Association for Computational Linguistics

phrase (*sol / sun*) is written in Latin. Keller (2016) provides an analysis of adjectival modifiers in the framework of the Matrix Language Frame model introduced by Myers-Scotton (2001).

þe	briȝt	sol	sapiencie	subtrahit	lumen	suum
the	bright	sun	wisdom	withdraws	light	its
eng.	eng.	lat.	lat.	lat.	lat.	lat.

The focus of our work lies on the extraction of such phrases with the help of POS patterns along with the language information for all words of each phrase.

The body of this paper is organized as follows. Section 2 gives an overview of work that has been done in the context of code-switching. In Section 3, we describe the data set that serves as a basis for the experiments described in Sections 4 and 5. Section 6 concludes with an outline of how our tools will be made available for wider use by the academic community.

2 Related Work

Previous work on automatic processing of mixed text can be divided into two main areas: research on LID and work on POS tagging.

LID for written as well as for spoken code-switching has been tackled for a wide range of language pairs and with different methods. Lyu and Lyu (2008) investigate Mandarin-Taiwanese utterances from a corpus of spoken language. They propose a word-based lexical model for LID integrating acoustic, phonetic and lexical cues. Solorio and Liu (2008a) predict potential code-switching points in Spanish-English mixed data. Different learning algorithms are applied to transcriptions of code-switched discourse. Jain and Bhat (2014) present a system on using conditional posterior probabilities for the individual words along with other linguistically motivated language-specific as well as generic features. They experiment with a variety of language pairs, e.g. Nepali-English, Mandarin-English or Spanish-English. Yeong and Tan (2011) use morphological structure and sequence of syllables in Malay-English sentences to identify language. Barman et al. (2014) investigate mixed text including three languages: Bengali, English and Hindi. They experiment with word-level LID, applying a simple unsupervised dictionary-based approach, supervised word-level classification with and without contextual clues, and sequence labeling using CRFs.

So far, not much work has been published on POS tagging of code-switching text. Solorio and Liu (2008b) present results on POS tagging Spanish-English code-switched discourse. They investigate methods ranging from simple heuristics to an algorithm combining features from the output of an English and a Spanish POS tagger. Rodrigues and Kübler (2013) show POS tagging for speech transcripts containing multilingual intra-sentinal code-mixing. They compare a tagging model trained on a heterogeneous-language data set to a model that switches between two homogeneous-language tagging models dynamically using word-by-word LID. Jamatia et al. (2015) use both a coarse-grained and a fine-grained POS tag set for tagging English-Hindi Twitter and Facebook chat messages. They compare performance of a combination of language specific taggers to that of applying four machine learning algorithms using a range of different features.

Considering the rather limited number of automatic processing tools for our languages at hand, we focus on those methods suggesting the application of shallow features for written language. Thus, we renounce morphological processing as described in Yeong and Tan (2011) and prosodic features since we are working with written text.

3 Data

The texts addressed in the following are so-called Macaronic sermons (Horner, 2006), a text genre containing diverse code-switching structures of Middle English and Latin which is thus highly informative both for historical multilingualism research and for computational linguistics. Our aim is to investigate phrase-internal code-switching. This requires language information on the token level on one hand and a basic understanding of the syntax of a sentence on the other. We aim at POS tagging as a basis for a pattern-extraction-based approach. In particular, we are interested in extracting mixed-language nominal phrases with a focus on determiners, attributive adjectives and adjective phrases as adnominals.

Since we are often dealing with a critically low data situation in Digital Humanities focusing on historical topics, we experiment with a data set which can realistically be acquired with just a few hours of annotation effort. This implies that our approach is easily applicable to language pairs for

label	explanation	%
l	Latin	60.5
e	Middle English	24.6
a	word in both languages	1.8
n	Named Entity	1.0
p	punctuation	12.1

Table 1: Labels annotated for LID along an explanation for each label and the occurrence in percent.

which there is only a limited amount of annotated data. Our annotated corpus comprises about 3000 tokens.

In a first step, we annotate the tokens for the following language information, mostly Latin and Middle English. The two languages share a small part of their vocabulary. Those words can e.g. be simple function words like *in*. For these items the attribution to one or the other language is not possible. We label these words with a separate tag to preserve the information that no decision on language could be made. Moreover, we mark named entities since they are often not part of the vocabulary of a language, as well as punctuation. Just about 25% of the tokens are Middle English compared to more than 60% of Latin words (cp. Table 1). Our data set comprises 159 sentences with an average length of 19.4 tokens. Overall we observe 316 switch points, which means an average number of two code-switching points per sentence.

In a second step, we annotate coarse-grained POS using the Universal Tagset (UT) suggested by Petrov et al. (2012). This choice facilitates a consistent annotation across languages since language specificities are conflated into more comprehensive categories. Nouns constitute by far the most frequent POS (cp. Table 2), which makes our data set a promising source for the investigation of nominal phrases.

4 Automated Processing of Mixed Text

We model LID and POS tagging as both two subsequent tasks in which POS tagging builds upon the results of the LID and two independent tasks where POS tagging and LID do not inform each other. LID can be understood as a step to facilitate POS tagging and any further processing of mixed text. In order to be used as a feature for POS tagging, it needs to be solved with a high accuracy to avoid error percolation through the entire processing pipeline.

label	explanation	%
ADJ	adjective	8.0
ADP	adposition (pre- and post)	7.9
ADV	adverb	6.0
CONJ	conjunction	7.9
DET	determiner	6.8
NOUN	noun (common and proper)	29.1
NUM	cardinal number	0.03
PRON	pronoun	4.3
PRT	particle or other function word	3.2
VERB	verb (all tenses and modes)	14.4
X	foreign word, typo, abbrev.	0.06
.	punctuation	12.3

Table 2: Labels annotated for POS tagging along with the explanation for each label and the occurrence in percent.

4.1 Language Identification

We use an approach similar to the one described by Solorio and Liu (2008a). Since there is no available lemmatizer for Middle English, in contrast to Solorio and Liu (2008b) we cannot add lemma information to our training. To compensate for the lack of lemmas, we include POS informed word lists for both languages extracted from manually annotated corpora. Following the POS introduced by the universal dependency initiative (Nivre et al., 2016), we extract lists for the following POS: adjectives, adverbs, prepositions, proper nouns, nouns, determiners, interjections, pronouns, verbs, auxilary verbs and conjunctions. For Middle English we extract these lists from the Penn Parsed Corpora of Historical English (Kroch and Taylor, 2000). For Latin, we revert to the Latin corpora included in the Universal Dependency treebank namely Latin Dependency Treebank 2.0 (LDT) (Bamman and Crane, 2011), Latin-PROIEL UD treebank (Haug and Jøhndal, 2008) and the Latin-ITTB UD treebank (McGillivray et al., 2009). In case a word is found in one of the lists, we add its POS as a feature.

CRF classifiers are known to be successful for sequence labeling tasks. Based on features extracted from the results given by monolingual taggers for our data, we train a CRF classifier (Lafferty et al., 2001) combining those features with

several other features. The features we implement are the following:

1 surface form
2 POS tag TreeTagger Latin
3 TreeTagger confidence Latin
4 POS tag TreeTagger Middle English
5 TreeTagger confidence Middle English
6 POS from Middle English word list
7 POS from Latin word list
8 character-unigrams prefix
9 character-bigrams prefix
10 character-trigrams prefix
11 character-unigram suffix
12 character-bigram suffix
13 character-trigram suffix

Features 2-5 are generated by the Latin and Middle English TreeTagger (Schmid, 1995), respectively. This means that this method is only an option for languages for which a TreeTagger model is available or can be trained[2]. We include character-n-gram affixes from length 1-3 to account for the fact that Latin is characterized by a relatively restricted suffix assignment. In addition, we use a context window of 5 tokens on all features.

4.2 Part-of-speech Tagging

For POS tagging, we use the same features as described in Section 4.1 (CRF$_{base}$). In order to investigate the influence of LID as a feature on POS Tagging, we also train the CRF classifier (CRF$_{predLID}$) using information generated by the LID system (feature 14.a). Since we cannot assume perfect LID, we evaluate the performance of a CRF classifier (CRF$_{goldLID}$) having the gold standard LID (feature 14.b) at its disposal. In this way, we can investigate to which degree differences in the quality of LID influence the POS tagging quality.

14.a LID label predicted by the system described in Section 4.1

14.b gold LID label manually annotated for our corpus

[2]We want to thank Achim Stein, University of Stuttgart, for providing the parameter file for Middle English.

	label	l	e	a	n	p	all
P	BL	68.9	0.0	0.0	0.0	100	33.8
	CRF	93.1	93.9	45.5	0.0	98.7	66.0
R	BL	100	0.0	0.0	0.0	99.4	40.0
	CRF	97.6	92.1	7.1	0.0	98.9	59.2
F	BL	81.6	0.0	0.0	0.0	100	36.3
	CRF	95.3	93.0	14.9	0.0	99.3	59.9

Table 3: Performance of the CRF system for langouage identification compared to the baseline (BL). Precision, recall and F-score per class and macro-average of all classes.

5 Results

We evaluate our systems in a 10-fold cross-validation setting using 80% for training, and 10% each for development and testing. We tune the hyper-parameter settings of our learning algorithm on our development set by testing different manually chosen parameter settings. The CRF classifier is trained with the CRF++ toolkit (Lafferty et al., 2001) using L2-regularization and a c-value of 1000. We report average results over all sets.

5.1 Language Identification

Since the sermons are primarily written in Latin featuring Middle English insertions, we use a combination of Latin and perfect punctuation labeling as a majority baseline (BL) for our LID system. We report per class precision, recall and F-score along with macro-averages for the overall system. We do not report accuracy since the number of instances per class highly varies.

As was to be expected, our system reliably finds the right label for Latin text and just a little less so for English. We attribute the poor performance for named entities and words appearing in both languages to the low number of training instances in

label	% err	% l	% e	% a	% n	% p
l	2.4	-	84.1	6.8	0.0	9.1
e	7.9	95.0	-	3.3	0.0	1.7
a	92.9	90.4	9.6	-	0.0	0.0
n	100	90	10.	0.0	-	0.0
p	0.5	100	0.0	0.0	0.0	-

Table 4: Percentage of incorrectly labeled tokens per class along with the distribution of incorrect labels among the other labels.

	label	ADJ	ADP	ADV	CONJ	DET	NOUN	NUM	PRON	PRT	VERB	X	.	all
	BL1	43.3	92.0	72.9	85.1	25.0	71.1	0.0	30.5	0.0	55.8	5.1	100	48.4
	BL2	55.7	83.1	68.6	87.2	37.5	82.5	0.0	34.5	23.2	78.2	7.1	100	54.8
P	CRF_{base}	68.1	92.0	81.2	88.8	79.3	85.2	0.0	82.2	71.4	85.9	0.0	98.2	69.4
	$\text{CRF}_{predLID}$	69.2	92.8	79.5	89.7	78.9	85.3	0.0	82.2	72.5	86.2	0.0	98.2	69.5
	$\text{CRF}_{goldLID}$	69.4	92.4	80.0	90.4	77.8	85.6	0.0	82.2	72.5	86.4	0.0	98.4	69.6
	BL1	51.0	80.6	56.8	63.1	3.3	79.4	0.0	45.1	0.0	76.5	1.0	98.4	46.3
	BL2	51.8	89.7	68.6	81.1	8.6	90.6	0.0	53.4	23.2	84.4	100	98.4	65.8
R	CRF_{base}	60.0	86.0	67.6	88.1	82.3	95.3	0.0	66.2	60.6	86.9	0.0	98.7	66.0
	$\text{CRF}_{predLID}$	60.4	85.5	69.2	88.9	82.3	95.4	0.0	66.2	58.6	87.6	0.0	98.4	66.0
	$\text{CRF}_{goldLID}$	65.1	89.1	74.2	89.4	80.0	90.3	0.0	73.3	64.8	87.0	0.0	98.7	66.2
	BL1	46.9	85.9	63.8	72.5	5.9	75.0	0.0	36.4	0.0	64.5	9.8	99.2	46.7
	BL2	53.7	86.3	68.8	84.1	14.0	86.4	0.0	41.9	36.5	81.2	13.3	99.2	55.5
F	CRF_{base}	**63.8**	88.9	73.7	88.5	**80.8**	**90.0**	0.0	73.3	65.6	86.4	0.0	98.4	67.4
	$\text{CRF}_{predLID}$	**64.5**	89.0	74.0	89.3	**80.6**	**90.1**	0.0	73.3	64.8	86.9	0.0	98.3	67.6
	$\text{CRF}_{goldLID}$	**65.1**	89.1	74.2	89.4	**80.0**	**90.3**	0.0	73.3	64.8	87.0	0.0	98.7	67.7

Table 5: Performance of the CRF systems for POS tagging compared to the majority baseline (BL1), the confidence baseline (BL2). CRF_{base}: system with the 13 basic features, $\text{CRF}_{predLID}$: system with predicted LID as an additional feature, $\text{CRF}_{goldLID}$: system with gold-standard LID as an additional feature. Precision (P), Recall (R) and F-score (F) per class and macro-average of all classes are given. The task-relevant results are emphasized in bold.

our corpus.

In order to investigate the primary sources of errors, we inspect the incorrectly labeled tokens per class. Table 4 shows that all but 2.4% of the Latin tokens are labeled correctly. The erroneous labels can be attributed to about 84% to English, 7% to the class that can appear in both languages. The remaining 9% contain wrong labels for punctuation. The performance for English tokens is slightly lower with a error rate of 7.9% incorrect labels which are almost all tagged as Latin. This can be due to the fact that our data contains more Latin tokens overall. The same effect is observable for the labels *a* (word in both languages) and *n* (named entities). Since the corpus contains just a few instances with those labels, they get incorrectly assigned to Latin. The small error in classifying punctuation appears in one of our cross-validation sets where colons are not part of the training but the test set.

5.2 Part-of-speech Tagging

For the evaluation of our POS tagger, we use two baselines. We compare the output of our systems to the output of the monolingual Latin tagger after mapping the Latin tagset to the UT. Moreover, we add a strong baseline, drawing on the confidence feature of the monolingual TreeTagger models. We choose the POS label of the monolingual tagger with a higher level of confidence. In case the label indicates that a word is a foreign word, we choose the label from the other language (in our case Middle English). We map all POS tags to the UT. Per-class results along with macro-F-score are shown in Table 5.

All our systems beat the baseline systems for almost all classes (except for BL2 adverb and verb) (cf. Table 5). With overall F-scores between 67.4 and 67.7 our systems achieve better F-scores than the baseline systems with an F-score of 46.7 and 55.5, respectively. In the further analysis we leave the results for NUM and X aside cause they appear just once and three times in the entire corpus, respectively. Even though the average scores for all classes combined range just between about 60 and 90, we achieve good results for classes with a high number of tokens in our corpus (e.g. nouns and verbs), and also for adpositions and conjunctions. Since macro-F-score gives equal weight to all classes the numbers might be misleading, depending on the purpose of the system. Given that we built the POS tagger with a specific task in mind, namely the extraction of nominal phrases, we calculate the F-score for the POS classes relevant to this task (determiners, adjectives and nouns). This gives a task-specific macro F-score of 78.2 (CRF_{base}), 78.4 ($\text{CRF}_{predLID}$) and 74.5 ($\text{CRF}_{goldLID}$), respectively. Those F-scores are noticeably above the average F-scores for the overall systems and also beat the task-specific F-

scores of BL1 (42.6) and BL2 (51.4). The relatively high average recall of almost 80 for these three labels combined for all three systems is important for the task whereas precision has lower priority, since the extracted phrases are manually inspected afterwards. Since our LID system performs well, the system with automatically predicted labels shows a slight increase in performance compared to the system without LID information. The system with manually annotated LID information yields the best performance. However, according to McNemar's test the differences are not statistically significant.

The analysis of the incorrectly labeled tokens shows which POS tags are difficult to distinguish (cf. Table 6). Since we are especially interested in adjectives, an error rate of 40% is rather high. Out of these, about 63% have been incorrectly labeled as nouns, which has considerable negative effect on our objective, especially since most of the incorrectly labeled nouns are labeled as adjectives. Almost 70% of the adjectives that are incorrectly labeled as nouns are Latin. This can be explained by the morphology of adjectives in Latin. As Latin adjectives and nouns have often similar, if not the same suffixes of case marking, the two classes cannot be distinguished using the suffix as a defining feature. These difficulties are also observed by vor der Brück and Mehler (2016) who present a morphological tagger for Latin.

	þis	made	hom	to	lede
	this	made	them	to	lead
lang.	eng.	eng.	eng.	eng.	eng.
gold	PRON	VERB	PRON	PRT	VERB
pred	PRON	VERB	PRON	PRT	VERB

	super	terram	celestem	conuersacionem
	on	earth	heavenly	regime
lang.	lat.	lat.	lat.	lat.
gold	ADP	NOUN	ADJ	NOUN
pred	ADP	DET	NOUN	NOUN

The first half of the sentence [3] is written in Middle English. The assigned POS tags are correct and also the first Latin word after the code-switching point is labeled correctly. The phrase *terram clestem conuersacionem* is tagged in the pattern of a noun phrase with a determiner and a compounded noun instead of a prepositional phrase *super terram* (Engl.: on earth) and a noun phrase (Engl.: heavenly behavior) consisting of an adjective and a noun. The similar syntactic function of pronouns (in case of possessive pronouns and demonstrative pronouns) and determiners leads to a source of error.[4]

[3]Translation by Horner (2006): *this made them lead on earth a heavenly regime*.

	LID			POS		
size	pre	rec	f-score	pre	rec	f-score
800	56.3	56.8	56.5	60.8.1	54.6	56.8
1600	56.6.0	57.8	57.2	66.7	63.0	64.6
2400	66.0	59.2	59.9.3	69.5	66.0	67.6

Table 7: Different portions of the training set along with precision, recall and F-score for LID and POS tagging.

	In	isto	non	est	fiducia
	In	this	not	is	confidence
lang.	lat.	lat.	lat	lat.	lat.
gold	ADP	PRON	PRT	VERB	NOUN
pred	ADP	DET	PRT	VERB	NOUN

On closer inspection, we find that many of the incorrectly tagged words appear in POS sequences which are either rarely or not at all contained in the training data. We predict that adding more training data will significantly decrease errors of this kind. Since data sparsity in general is an issue dealing with historical text, we investigate how different sizes of the training set influence the results. We compare results for 800 tokens, 1600 tokens, and for the complete training set (around 2400 tokens).

With an increase of training instances, the results improve for both tasks (cf. Table 7). The increase from 800 to 1600 is higher than from 1600 to 2400. This suggests that the F-score might grow logarithmically with increasing training size.

6 Tools for Digital Humanities

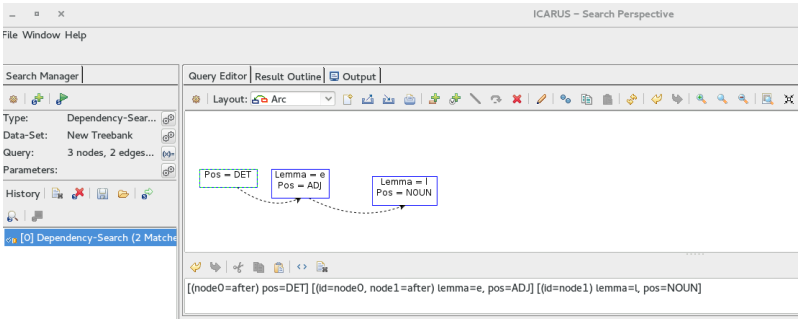
Since the aim of our project is not only to build a proof-of-concept system but to enable Humanities scholars to automatically process their data with the help of our tools, we implement a simple web service in Java to offer an easily accessible interface to our tool.[5]. The data is returned in a format compatible with ICARUS, a search and visualization tool which primarily targets dependency trees (Gärtner et al., 2013). Despite the present lack of a dependency-parsed syntax layer, ICARUS offers the opportunity to inspect the data and pose complex search requests, combining the three layers

[4]Translation by Horner (2006): *in it there is no confidence*.

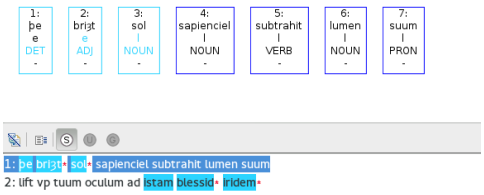
[5]The web service is hosted at `https://clarin09.ims.uni-stuttgart.de/normalisierung/mixed-pos.html` For access, please contact the author.

label	% err	ADJ	ADP	ADV	CONJ	DET	NOUN	PRON	PRT	VERB	.
ADJ	39.6	-	2.1	3.1	0.0	9.3	**62.9**	0.0	1.0	20.6	1.0
ADP	14.6	11.4	-	8.6	6.5	5.7	11.4	0.0	**37.1**	**14.3**	2.9
ADV	30.8	19.3	5.3	-	10.5	5.3	**33.3**	7.0	1.8	14.0	0.0
CONJ	11.1	0.0	0.0	**37.0**	-	11.1	7.4	22.2	11.1	7.4	3.7
DET	17.7	16.2	10.8	10.8	2.7	-	**32.4**	10.8	8.1	8.1	0.0
NOUN	4.6	**56.1**	0.0	9.8	0.0	0.0	-	2.4	0.0	26.8	4.9
PRON	33.8	8.8	0.0	2.2	15.5	**31.1**	20.0	-	2.2	17.8	2.2
PRT	41.4	4.9	12.2	14.6	17.1	**22.0**	14.6	2.4	-	12.2	0.0
VERB	12.4	25.5	3.6	1.8	0.0	7.3	**54.5**	5.5	0.0	-	1.8
.	1.6	33.3	0.0	0.0	16.7	0.0	**50.0**	0.0	0.0	0.0	-

Table 6: Percentage of incorrectly labeled tokens per class along with the distribution of incorrect labels among the other labels for the CRF$_{predLID}$ system.



(a) Formulation of a search query in ICARUS.



(b) Results shown by ICARUS

Figure 1: Search interface of ICARUS returning results on a query for an English adjective followed by a Latin noun within the next 3 tokens.

of token, language information and POS tag. Figure 1 shows a query that extracts all sequences of a determiner in either of both languages followed by a Middle English adjective followed by a Latin noun. ICARUS shows the results within the sentence of origin. ICARUS also allows searches including gaps. This is helpful, since nominal phrases vary according to the number of adjectives and as to whether or not they contain an overt determiner. Thus, flexibility in formulating the search query facilitates an in-depth search of all possible constructions.

Our method can easily be adapted to other languages by inserting the fitting monolingual taggers (TreeTagger) and POS related word lists (if available). For this purpose, the code is publicly available on Github[6].

7 Conclusion and future work

We show the implementation and application of two systems developed for a specific purpose. We get reasonable results given the very low number of annotated training instances. Considering the detailed error analysis for our system, we can purposefully extend our training data in order to correct the sources of error in the future by for example adding monolingual data from the Penn-Helsinki Parsed Corpus of Middle English (Kroch and Taylor, 2000).

Subsequently, we will look into the possibility of jointly modeling LID and POS tagging. Eventually, we aim at a dependency parser for mixed

[6] https://github.com/sarschu/CodeSwitching

text in order to get deeper insights into the constraints on intra-sentinal code-switching.

We aim to show that not just the development of tools but also the support with respect to applying them constitutes an important component of successful collaboration between Humanities and Computer Science. In return, a task-oriented tool development along with immediate feedback on the performance and analysis of error from the Humanities side facilitate the implementation of systems that do not only serve the proof of a concept but are applied to real-world data. We believe that this kind of collaboration is the way to give Computer Science the chance to support other fields in their research and find new and interesting challenges throughout this work.

Acknowledgments

We want to thank André Blessing for his support with setting up the web service and Achim Stein for providing us with a TreeTagger model for Middle English. We are greatly endebted to the Pontifical Institute of Mediaeval Studies (PIMS), Toronto, for their support and kind permission to use a searchable PDF version of the sermon transcripts. This research has been performed within the CRETA project which is funded by the German Ministry for Education and Research (BMBF).

References

David Bamman and Gregory Crane. 2011. The ancient greek and latin dependency treebanks. In Caroline Sporleder, Antal Bosch, and Kalliopi Zervanou, editors, *Language Technology for Cultural Heritage*, Theory and Applications of Natural Language Processing, pages 79–98. Springer Berlin Heidelberg. 10.1007/978-3-642-20227-8_5.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code-mixing: A challenge for language identification in the language of social media. In *In Proceedings of the First Workshop on Computational Approaches to Code-Switching*.

Amitava Das and Björn Gambäck. 2013. Code-mixing in social media text. the last language identification frontier? *TAL*, 54(3):41–64.

Markus Gärtner, Gregor Thiele, Wolfgang Seeker, Anders Björkelund, and Jonas Kuhn. 2013. Icarus – an extensible graphical search tool for dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Sofia, Bulgaria, August. Association for Computational Linguistics.

Dag T. T. Haug and Marius L. Jøhndal. 2008. Creating a parallel treebank of the old indo-european bible translations. In Caroline Sporleder and Kiril Ribarov, editors, *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34.

Patrick J. Horner. 2006. *A Macaronic Sermon Collection from Late Medieval England: Oxford, MS Bodley 649*. Pontifical Institute of Mediaeval Studies Toronto: Studies and texts. Pontifical Institute of Mediaeval Studies.

Naman Jain and Riyaz Ahmad Bhat. 2014. Language identification in code-switching scenario. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 87–93, Doha, Qatar.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2015. Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages. In *Recent Advances in Natural Language Processing, RANLP 2015, 7-9 September, 2015, Hissar, Bulgaria*, pages 239–248.

Judith A. Jefferson, Ad Putter, and Amanda Hopkins. 2013. *Multilingualism in Medieval Britain (c. 1066-1520): Sources and Analysis*. Medieval texts and cultures of Northern Europe. Brepols.

Mareike Keller. 2016. Code-switched adjectives and adverbs in macaronic sermons. In Elise Louviot and Catherine Delesse, editors, *Proceedings of the Biennial Conference on the diachrony of English (CBDA4)*. forthcoming.

Anthony Kroch and Ann Taylor. 2000. The penn-helsinki parsed corpus of middle english (ppcme2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 4 (http://www.ling.upenn.edu/ppche-release-2016/PPCME2-RELEASE-4).

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Dau-Cheng Lyu and Ren-Yuan Lyu. 2008. Language identification on code-switching utterances using multiple cues. In *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, pages 711–714.

Barbara McGillivray, Marco Passarotti, and Paolo Ruffolo. 2009. The index thomisticus treebank project: Annotation, parsing and valency lexicon. *TAL*, 50:103–127.

Carol Myers-Scotton. 2001. The matrix language frame model: Development and responses. *Codeswitching Worldwide II*, pages 23–58.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Arja Nurmi and Päivi Pahta. 2013. Multilingual practices in the language of the law: Evidence from the lampeter corpus. In Olga Timofeeva Jukka Tyrkkö and Maria Salenius, editors, *Ex Philologia Lux: Essays in Honour of Leena Kahlas-Tarkka (Mémoires de la Société Nophilologique de Helsinki XC)*, pages 187–205. Société Nophilologique.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Paul Rodrigues and Sandra Kübler. 2013. Part of speech tagging bilingual speech transcripts with intrasentential model switching. In *Analyzing Microtext, Papers from the 2013 AAAI Spring Symposium, Palo Alto, California, USA, March 25-27, 2013*.

Herbert Schendl and Laura Wright. 2012. *Code-Switching in Early English*. Topics in English Linguistics [TiEL]. De Gruyter.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Thamar Solorio and Yang Liu. 2008a. Learning to predict code-switching points. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 973–981, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thamar Solorio and Yang Liu. 2008b. Part-of-speech tagging for english-spanish code-switched text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 1051–1060, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tim vor der Brück and Alexander Mehler. 2016. TLT-CRF: A lexicon-supported morphological tagger for Latin based on conditional random fields. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, LREC 2016. Accepted.

S. Wenzel. 1994. *Macaronic sermons: bilingualism and preaching in late-medieval England*. Recentiores : Later Latin Texts and Contexts. University of Michigan Press.

Yin-Lai Yeong and Tien-Ping Tan. 2011. Applying grapheme, word, and syllable information for language identification in code switching sentences. In *International Conference on Asian Language Processing, IALP 2011, Penang, Malaysia, 15-17 November, 2011*, pages 111–114.