Detecting Code-Switching in a Multilingual Alpine Heritage Corpus

Martin Volk and Simon Clematide
University of Zurich
Institute of Computational Linguistics
volk|siclemat@cl.uzh.ch

Abstract

This paper describes experiments in detecting and annotating code-switching in a large multilingual diachronic corpus of Swiss Alpine texts. The texts are in English, French, German, Italian, Romansh and Swiss German. Because of the multilingual authors (mountaineers, scientists) and the assumed multilingual readers, the texts contain numerous code-switching elements. When building and annotating the corpus, we faced issues of language identification on the sentence and subsentential level. We present our strategy for language identification and for the annotation of foreign language fragments within sentences. We report 78% precision on detecting a subset of code-switches with correct language labels and 92% unlabeled precision.

1 Introduction

In the Text+Berg project we have digitized the yearbooks of the Swiss Alpine Club (SAC) from its first edition in 1864 until today. They contain articles about mountain expeditions, the flora and fauna of the Alpes and other mountain regions, glacier and climate observations, geology and history papers, book reviews, accident and security reports, as well as the protocols of the annual club gatherings. The texts are in the four official languages of Switzerland French, German, Italian and Romansh [1] plus a few in English and Swiss German dialects.

Because of the multilinguality of the authors and readers, many articles are mixed-language texts with inter-sentential and intra-sentential code-switching. This poses a challenge for automatically processing the texts. When we apply Part-of-Speech (PoS) tagging, named entity recognition or parsing, our systems need to know the language that they are dealing with. Therefore we had used a language identifier from the start of the project to mark the language of each sentence. We report on our experiences with sentence-based language identification in section 3. Figure 1 shows an example of a French text with an English appendix title plus an English quote from this book.

Lately we discovered that our corpus also contains many intra-sentential code-switches. For example, we find sentences like

> ... und ich finde es ≪very nice and delightful≫ einen Vortrag halten zu dürfen. (Die Alpen, 1925) *(EN : ... and I find it very nice and delightful to be allowed to give a talk.)*

where the German sentence contains an English phrase in quotation marks. Obviously, a German PoS tagger will produce nonsense tags for the English phrase as the words will be unknown to it. PoS taggers are good at tagging single unknown words based on the surrounding context, but most taggers fail miserably when a sequence of two or more words is unknown. The upper half of figure 2 shows the PoS tagger output for the above example. The words *very, nice, delightful* are senselessly tagged as proper names (NE), only *and* is tagged as foreign word (FM).

Our goal is to detect all intra-sentential code-switches and to annotate them as exemplified in the lower half of figure 2. They shall be framed with the TEI-conformant tag `<foreign>` which also shall specify the language of the foreign language segment. All tokens in the segment shall be tagged as foreign words (e.g. FM in the German STTS tag set, ET in the French Le Monde tag set (Abeillé et al., 2003)), and each lemma shall get

1. Romansh is the 4th official language in Switzerland. It is spoken by around 25,000 people in the mountainous South-Eastern canton of Graubünden.

Proceedings of The First Workshop on Computational Approaches to Code Switching, pages 24–33,
October 25, 2014, Doha, Qatar.

the special symbol @fn@ to set it apart from lemmas of the surrounding sentence. In this paper we report on our experiments towards this goal and suggest an algorithm for detecting code-switching.

We adopt a wide definition of code-switching. We are interested in detecting all instances where a text is in a dominant language and contains words, phrases and sentences in another language. Though our definition is broad, it is clearly more restricted than others, as e.g. the definition by Kracht and Klein (2014) which includes special purpose codes like bank account numbers or shoe sizes.

In this paper we will give an overview of the language mix in the yearbooks of the Swiss Alpine Club over the 150 years, and we will illustrate how we identified inter-sentential and intra-sentential code-switching. We will give a quantitative overview of the number of code-switching candidates that we automatically located.

2 The Text+Berg Corpus

The Text+Berg corpus comprises the annual publications of the Swiss Alpine Club (SAC) from its first edition in 1864 until 2013. From the start until 1923 the official yearbook was called "Jahrbuch des Schweizer Alpen-Club" (EN : yearbook of the Swiss Alpine Club), and it typically consisted of 500 to 700 pages. The articles of these first 60 years were mostly in German (with 86% of the words), but some also in French (13% of the words) and few in Italian and Romansh (Volk et al., 2010).

Interestingly, the German articles contained passages in French and sometimes other languages (e.g. English, Swiss German, Latin) without translations, and vice versa. Obviously, the article authors and yearbook editors assumed that the readers of the yearbook were polyglott at least in English, French, German and Latin during that time. In fact, the members of the SAC in the 19th century came from an academic elite. Mountain exploration was a past-time of the rich and educated.

Still, during that same time the French-speaking sections of the Swiss Alpine Club published their own yearbook in parallel to the official yearbook and called it "Echo des Alpes". It started shortly after the official yearbook in the late 1860s and continued until 1923. Each "Echo des Alpes" yearbook contained between 300 to 600 pages adding up to a total of 22,582 pages with 7.4 million tokens, almost all in French with rare quotes in German.

As of 1925 the official SAC yearbook and the "Echo des Alpes" were merged into a new publication called "Die Alpen. Les Alpes. Le Alpi" (in German, French, Italian) which has been published ever since. Over the years it sometimes appeared as quarterly and sometimes as monthly magazine. Today it appears 12 times per year in magazine format. For the sake of simplicity we continue to call each annual volume a yearbook.

The merger in 1925 resulted in a higher percentage of French texts in the new yearbook. For example, the 1925 yearbook had around 143,000 words in German and 112,000 in French (56% to 44%). The ratio varied somewhat but was still at 64% to 36% in 1956.

From 1957 onwards, the SAC has published parallel (i.e. translated) French and German versions of the yearbooks. At the start of this new era only half of the articles were translated, the rest was printed in the original language in identical versions in the two language copies.

Over the next decade the number of translations increased and as of 1983 the yearbooks were completely translated between German and French. Few Italian articles were still published verbatim in both the French and German yearbooks. As of 2012 the SAC has launched an Italian language version of its monthly magazine so that now it produces French, German and Italian parallel texts.

In its latest release the Text+Berg corpus (comprising the SAC yearbooks, the ALPEN magazine and the Echo des Alpes) contains around 45.8 million tokens (after tokenization). French and German account for around 22 million tokens each, Italian accounts for 0.8 million tokens. The remainder goes to English, Latin, Romansh and Swiss German. The corpus is freely available for research purposes upon request.

3 Language Identification in the Text+Berg Corpus

We compiled the Text+Berg corpus by scanning all SAC yearbooks from 1864 until 2000 (around 100,000+ pages). Afterwards we employed commercial OCR software to convert the scan images into electronic text. We developed and applied techniques to automatically reduce the number of OCR errors (Volk et al., 2011).

We obtained the yearbooks from 2001 to

de la publication du *Mount Everest 1938* [2] de Tilman. L'*Appendix B*, intitulé: *Antropology or Zoology, with Particular Reference to the Abominable Snowman*, reprend la question ab *ovo* [3], en la soumettant, pp. 127–137, à une enquête strictement impartiale.

La conclusion de cette enquête est résumée dans les sept dernières lignes de la page 137:

«I merely affirm that traces for which no adequate explanation is forthcoming have been seen and will continue to be seen in various parts of the Himalaya, and until a worthier claimant is found we may as well attribute them to the ‚Abominable snowman'. *And I think he would be a bold and in some ways an impious sceptic who, after balancing the evidence, does not decide to give him the benefit of the doubt* [4].»

Conclusion: Devant l'opinion fortement documentée et motivée de Tilman, peut-être

FIGURE 1 – Example of an English title and an English quote in a French text (Die Alpen, 1955)

```
<w n="23-16-21" lemma="und"      pos="KON">und</w>
<w n="23-16-22" lemma="ich"      pos="PPER">ich</w>
<w n="23-16-23" lemma="finden"   pos="VVFIN">finde</w>
<w n="23-16-24" lemma="es"       pos="PPER">es</w>
<w n="23-16-25" lemma="«"        pos="$(">«</w>
<w n="23-16-26" lemma="unk"      pos="NE">very</w>
<w n="23-16-27" lemma="unk"      pos="NE">nice</w>
<w n="23-16-28" lemma="and"      pos="FM">and</w>
<w n="23-16-29" lemma="unk"      pos="NE">delightful</w>
<w n="23-16-30" lemma="»"        pos="$(">»</w>
<w n="23-16-31" lemma="ein"      pos="ART">einen</w>
<w n="23-16-32" lemma="Vortrag"  pos="NN">Vortrag</w>
<w n="23-16-33" lemma="halten"   pos="VVINF">halten</w>
<w n="23-16-34" lemma="zu"       pos="PTKZU">zu</w>
<w n="23-16-35" lemma="dürfen"   pos="VMINF">dürfen</w>
<w n="23-16-36" lemma="."        pos="$.">.</w>

============ after code-switch detection ============

<w n="23-16-21" lemma="und"      pos="KON">und</w>
<w n="23-16-22" lemma="ich"      pos="PPER">ich</w>
<w n="23-16-23" lemma="finden"   pos="VVFIN">finde</w>
<w n="23-16-24" lemma="es"       pos="PPER">es</w>
<foreign lang="en">
    <w n="23-16-25" lemma="«"     pos="$(">«</w>
    <w n="23-16-26" lemma="@fn@" pos="FM">very</w>
    <w n="23-16-27" lemma="@fn@" pos="FM">nice</w>
    <w n="23-16-28" lemma="@fn@" pos="FM">and</w>
    <w n="23-16-29" lemma="@fn@" pos="FM">delightful</w>
    <w n="23-16-30" lemma="»"     pos="$(">»</w>
</foreign>
<w n="23-16-31" lemma="ein"      pos="ART">einen</w>
<w n="23-16-32" lemma="Vortrag"  pos="NN">Vortrag</w>
<w n="23-16-33" lemma="halten"   pos="VVINF">halten</w>
<w n="23-16-34" lemma="zu"       pos="PTKZU">zu</w>
<w n="23-16-35" lemma="dürfen"   pos="VMINF">dürfen</w>
<w n="23-16-36" lemma="."        pos="$.">.</w>
```

FIGURE 2 – Example of an annotated German sentence with English segment, before and after code-switch detection (Die Alpen, 1925)

2009 as PDF documents which we automatically converted to text. The subsequent yearbooks from 2010 until 2013 we received as XML files from the SAC.

We have turned the whole corpus into a uniform XML format. For this, the OCR output texts as well as the texts converted from PDF and XML are structured and annotated by automatically marking article boundaries, by tokenization, language identification, Part-of-Speech tagging and lemmatization. Our processing pipeline also includes toponym recognition and geo-coding of mountains, glaciers, cabins, valleys, lakes and towns. Furthermore we recognize and co-reference person names (Ebling et al., 2011), and we annotate temporal expressions (date, time, duration and set) with a variant of HeidelTime (Rettich, 2013). Finally we analyze the parallel parts of our corpus and provide sentence alignment information that is computed via BLEUalign (Sennrich and Volk, 2011).

In order to process our texts with language-specific tools (e.g. PoS tagging and person name recognition) we employed automatic language identification on the sentence level. We used Lingua-Ident [2] (developed by Michael Piotrowski) to determine for each sentence in our corpus whether it is in English, French, German, Italian or Romansh. Lingua-Ident is a statistical language identifier based on letter n-gram frequencies. For long sentences it reliably distinguishes between the languages. Unfortunately it often misclassifies short sentences. Therefore we decided to use it only for sentences with more than 40 characters. Shorter sentences are assigned the language of the article. This can be problematic for mixed language articles. An alternative strategy would be to assign the language of the previous sentence to short sentences.

For sentences that Lingua-Ident judges as German we run a second classifier that distinguishes between Standard German and Swiss German dialect text. Since there are no writing rules for Swiss German dialects, they come in a variety of spellings. We have compiled a list of typical Swiss German words (e.g. Swiss-German : *chli, chlii, chlini, chline* = German : *klein, kleine* = English : *small*) that are not used in Standard German in order to identify Swiss German sentences. [3]

2. http ://search.cpan.org/dist/Lingua-Ident/

3. We are aware that the Text+Berg corpus contains also occasional sentences (or sentence fragments) in other German dialects (e.g. Austrian German, Bavarian German) and in old German spellings. Since these varieties are rare in the corpus, we do not deal with them explicitly.

Based on the language tag of each sentence we are able to investigate coarse-grained code-switching. Whenever the language of a sentence deviates from the language of the article, we have a candidate for code-switching. For example, in the yearbook 1867 we find a German text (describing the activities of the club) with a French quote :

> Der Berichterstatter bemerkt darüber : "On peut remarquer à cette occasion qu'il est rare que par un effort de l'esprit on puisse mettre du brouillard en bouteille, et ..." Die etwas ältere Sektion Diablerets, deren Steuer Herr August Bernus mit kundiger Hand ...

Most code-switching occurs with direct speech, quotes and book titles. The communicative goal is obviously to make the text more authentic.

4 Related Work on Detection of Code-Switching

Most previous work on automatically detecting code-switching focused on the switches between two known languages (whereas we have to deal with a mix of 6 languages).

Solorio and Liu (2008) worked on real-time prediction of code-switching points in Spanish-English conversations. This means that the judgement whether the current word is in a different language than the language of the matrix clause can only be based on the previous words. They use the PoS tag and its probability plus the lemma as provided by both the Spanish and the English TreeTagger as well as the position of the word in the Beginning-Inside-Outside scheme as features for making the decision. In order to keep the number of experiments manageable they restricted their history to one or two preceding words. As an interesting experiment they generated code-switching sentences Spanish-English based on their different predictors and asked human judges to rate the naturalness of the resulting sentences. This helped them to identify the most useful code-switching predictor.

Vu et al. (2013) and Adel et al. (2013) consider English-Mandarin code-switching in speech recognition. They investigate recurrent neural network language models and factored language models to the task in an attempt to integrate syntactic features. For the experiments they use SEAME,

the South East Asia Mandarin-English speech corpus compiled from Singaporean and Malaysian speakers. It consists of spontaneous interviews and conversations. The transcriptions were cleaned and each word was manually tagged as English, Mandarin or other. The data consists of an intensive mix of the two languages with the average duration of both English and Mandarin segments to be less than a second (!). In order to assign PoS tags to this mixed language corpus, the authors applied two monolingual taggers and combined the results.

Huang and Yates (2014) also work on the detection of English-Chinese code-switching but not on speech but rather on web forum texts produced by Chinese speakers living in the US. They use statistical word alignment and a Chinese language model to substitute English words in Chinese sentences with suitable Chinese words. Preparing the data in this way significantly improved Machine Translation quality. Their approach is limited to two known languages and to very short code-switching phrases (typically only one word).

Tim Baldwin and his group (Hughes et al., 2006) have surveyed the approaches to language identification at the time. They found a number of missing issues, such as language identification for minority languages, open class language identification (in contrast to identification within a fixed set of languages), sparse training data, varying encodings, and multilingual documents. Subsequently they (Lui and Baldwin, 2011) introduced a system for language identification of 97 languages trained on a mixture of corpora from different domains. They claim that their system Langid is particularly well suited for classifying short input strings (as in Twitter messages). We therefore tested Langid in our experiments for code-switching detection.

5 Exploratory Experiments with the SAC Yearbook 1925

In order to assess the performance of Langid for the detection of code-switching we performed an exploratory experiment with the SAC yearbook 1925. We extracted all word sequences between pairs of quotation marks where at least one token had been assigned the "unknown" lemma by our PoS tagger. The "unknown" lemma indicates that this word sequence may come from a different language.

The word sequence had to be at least 4 characters long, thus skipping single letters and abbreviations. In this way we obtained 333 word sequences that are potential candidates for intrasentential code-switching. We then ran these word sequences through the Langid language identification system with the restriction that we expect the word sequences only to be either English, French, German, Italian or Latin (Romansh and Swiss German are not included in Langid). For a given string Langid delivers the most likely language together with a confidence score.

We then compared the language predicted by the Langid system with the (automatically) computed language of the complete sentence. In 189 out of the 333 sentences the Langid output predicted a code-switch. We then manually graded all Langid judgements and found that 225 language judgements (67.5%) were correct. But only 89 of the 189 predicted code-switches came with the correct language. 40 of the 100 incorrect judgements were actually code-switches but with a different language. The remaining ones should have been classified with the same language as the surrounding sentence and are thus no examples of code-switching.

A closer inspection of the results revealed that the book contained not only code-switches in the expected 5 languages, but also into Romansh (6), Spanish (4) and Swiss-German (13). Obviously all of these were incorrectly classified. Most (8) of the Swiss-German word sequences were classified as German which could count as half correct, but the others were misclassified as English (among them a variant of the popular Swiss German farewell phrase *uf Wiederluege* spelled as *uf's Wiederluege*).

The Langid system has a tendency to classify word sequences as English. Many of the short, incorrectly classified word sequences were judged as English. It turns out that Langid judges even the empty string as English with a score of 9.06. Therefore all judgements with this score are dubious. We found that 56 short word sequences were classified as English with this score, out of which 35 were erroneously judged as English. Only strings with a length of 15 and more characters that are classified as English should be trusted. All others need to be discarded.

In general, if precision is the most important aspect, then Langid should only be used for strings

SAC yearbooks	candidates	predicted code-sw	correct	wrong lang	no code-sw
1868 to 1878	388	121	88	33	13
1926 to 1935	792	335	266	69	23
Total	1180	456	354	102	36

TABLE 1 – Recognition of code-switches in the Text+Berg corpus

with 20 or more characters. In our test set only 4 strings that were longer than 20 characters were incorrectly classified within the selected language set. Among the errors was the famous Latin phrase *conditio sine qua non* (length : 21 characters including blanks) which Langid incorrectly classified as Italian.

Another reason for the considerable number of misclassifications can be repeated occurrences of a word sequence. Our error count is a token-based count and thus prone to misclassified recurring phrases. In our experiment, Langid misclassified the French book name *Echo des Alpes* as Italian. Unfortunately this name occurs 18 times in our test set and thus accounts for 18 errors. We suspect that an *-o* at the end of a word is a strong indicator for Italian. In a short string like *Echo des Alpes* (14 characters), this can make the difference.

Another interesting observation is that hyphens speak for German. Our test set contains the hyphenated French string *vesse-de-neige* which Langid misclassifies as German with a clear margin over French. When the same string is analyzed without hyphens, then Langid correctly computes a preference for French over German. A similar observation comes from the Swiss German phrase *uf's Wiederluege* being classified as English when spelled with the apostrophe (which is less frequent in German than in English). Without the apostrophe Langid would count the string as German. With short strings like this, special symbols have a visible impact on the language identification.

We also observed that Langid is sensitive to all-caps capitalization. For example, *AUS DEM LEBEN DER GEBIRGSMUNDARTEN* (EN : The Lives of Mountain Dialects) is misclassified as English (with the default score) while *Aus dem Leben der Gebirgsmundarten* is correctly classified as German.

Overall, we found that code-switching within the same article rarely targets different languages. For example, if the article is in German and contains code-switches into English, then it hardly ever contains code-switches into other languages. In analogy to the one-sense-per-discourse hypothesis we might call this the one-code-switch-language-per-discourse hypothesis.

6 Detecting Intra-sentential Code-Switching

Based on exploratory studies and observations we decided on the following algorithm for detecting and annotating intra-sentential foreign language segments in the Text+Berg corpus. We search for sub-sentential token sequences (possibly of length 1) that are framed by a pair of quotation marks and that contain at least one "unknown" lemma. There must be at least two tokens outside of the quotation marks in the same sentence. As a compromise we restrict our detection to strings longer than 15 characters so that we get relatively reliable language judgements by Langid. The strings may consist of one token that is longer than 15 characters (e.g. *Matterhornhochtourist*) or a sequence of tokens whose sum of characters including blanks is more than 15. We feed these candidate strings to Langid for language identification and compare the output language with the language attribute of the surrounding sentence. If the languages are different, then we regard the token sequence as code-switch and mark it accordingly in XML as shown in figure 2.

In order to determine the **precision** of this algorithm, we checked 10 yearbooks from 1868 to 1878 (there was no yearbook in 1870) and from 1926 to 1935. The results are in table 1. From the 1180 code-switch candidates that we computed based on the above restrictions, Langid predicted 456 code-switches (39%). This means that in 39% of the cases Langid predicted a language that was different from the language of the surrounding sentence.

We manually evaluated all 456 predicted code-switches and found that 354 of them (78%) were correctly classified and labeled. These segments were indeed in a different language than the surrounding sentence and their language was correctly determined. For example, the French seg-

SAC yearbooks	> 15 characters without unknowns			≤ 15 characters		
	all	sample : TN/FN		all	sample : TN/FP	
1868 to 1878	322		20/1	404		15/8
1926 to 1935	1944		78/1	1136		54/23
Total	2266	(2%)	98/2	1540	(31%)	69/31

TABLE 2 – Estimation of the loss of recall due to the filtering approach based on a random sample of 100 quotations for each filtering category (TN : true negatives, FN : false negatives)

ment in the following German sentence is correctly detected and classified :

> Anschliessend führte Ambros dasselbe Bergsteigertrio ≪dans des circonstances très défavorables≫ auf den Monte Rosa ... (Die Alpen, 1935) *(EN : Afterwards Ambros led the same 3 mountaineers ≪under very unfavorable conditions≫ onto Monte Rosa.)*

Out of the 102 segments whose language was wrongly classified, only 36 were no code-switches. For example, the Latin segment *cum grano salis africani* is indeed a code-switch in a German sentence although Langid incorrectly classifies it as English. In fact, our evaluation showed that Langid is "reluctant" to classify strings as Latin. Latin strings are often misclassified as English or Italian.

Overall this means that only 8% of the predicted code-switches are no code-switches. Therefore we can safely add the module for code-switch detection into our processing and annotation pipeline.

In order to estimate the **recall** of our quotation filtering approach we manually evaluated a sample of the quotations that our algorithm excluded. Table 2 presents the numbers for the two time periods for two cases : first for sequences that are longer than 15 characters and contain only known lemmas, second for sequences that are shorter than 16 characters and contain at least one "unknown" lemma. For both cases we checked 100 instances.

The evaluation for the quotations with more than 15 characters but with all known lemmas (no "unknown" lemma) shows only 2 false negatives. Therefore, we can conclude safely that most of the code-switches with more than 15 characters were included in our candidate set.

Table 2 also shows that there were 1540 quotations with 15 or less characters. The manual inspection of 100 randomly selected quotations revealed that 31 indeed include foreign material. Some of these quotations are geographic names, e.g. the valley *Bergell* (EN/IT : Val Bregaglia), where it is difficult to decide whether this should be regarded as a code-switch. For this evaluation, we sticked to the principle that a foreign geographic name in quotation marks counts as a code-switch. The number of missed code-switches is high (31%). However, due to the limited precision of Langid (and other character-based language identifiers) for short character sequences, we still consider our length threshold appropriate. A different approach to language identification is needed to reliably classify these short quotes.

7 Discussion

The correctly marked code-switches in our test periods can be split by language of the matrix sentence and the language of the sub-sentential segment (= the code-switch segment). Table 3 gives an overview of the types of code-switches for the two periods under investigation. We see clearly that code-switches from German to English were rare in the 19th century (8 out of 89 = 9%) but became much more popular in the 1920s and 1930s (61 out of 265 = 23%). This came at the cost of French which lost ground from 54% (48 out of 89) to 40% (106 out of 265).

One can only compare the code-switch numbers from German with the corresponding numbers from French after normalizing the numbers in relation to the overall amount of text in German and French. During the first period (1868 to 1878) we count roughly 200,000 tokens in French and 1.4 million tokens in German, whereas in the second period (1926 to 1935) we have around 1 million tokens in French and again 1.4 million tokens in German. For the first period we find 87 code-switches (triggered by quotation marks) in the 1.4 million German tokens compared to 189 code-switches in the second period. The num-

sent lang	segm lang	1868 to 1878	1926 to 1935
de	en	8	61
de	fr	48	106
de	it	24	19
de	la	7	3
fr	de	2	35
fr	en	-	20
fr	it	-	11
fr	la	-	2
it	de	-	3
it	en	-	2
it	fr	-	3
Total		89	265

TABLE 3 – Correctly detected code-switches in the Text+Berg corpus

sent lang	segm lang	1868 to 1878	1926 to 1935
de	en	13	23
de	fr	9	5
de	it	8	12
de	la	2	1
fr	de	-	7
fr	en	1	10
fr	it	-	8
fr	la	-	1
it	en	-	2
Total		33	69

TABLE 4 – Incorrectly labeled code-switches in the Text+Berg corpus

ber of code-switches have clearly increased. For French we observe the same trend with 2 code-switches in 200'000 words in the first period compared to 68 code-switches in the 1 million tokens in the second period.

There is also a striking difference between French and German with many more code-switches in German than in French. For instance, for German we find 135 code-switches per 1 million tokens in the second period vs. 68 code-switches per 1 million tokens for French.

One surprising finding were the code-switches into Latin. We had not noticed them before, since our corpus does not contain longer passages of Latin text. But this study shows that code-switches into Latin persisted into the 1920s (3 out of German and 2 out of French).

correct segm	Langid prediction					
lang	en	it	fr	la	de	Total
la	15	12	3		1	31
de	7	5	5	1		18
fr	7	3				10
it	6					6
es	3	1		2		6
rm		1	2			3
ru	1					1
id	1					1
Total	40	22	10	3	1	76

TABLE 5 – Confusion matrix for incorrectly labeled code-switches in the periods 1868 to 1878 and 1926 to 1935

On the negative side (cf. table 4), misclassifying segments as English is the most frequent cause for a wrong language assignment in both periods. Table 5 shows the confusion matrix which contrasts the manually determined segment language with the incorrect language predicted by Langid. This confirms that Langid has a tendency to classify short text segments as English. But there are also a number of errors for Latin being mistaken for Italian, and German being mistaken for Italian or French.

As a general remark, it should be noted that an n-gram-based language identifier has advantages over a lexicon-based language identifier in the face of OCR errors. In the yearbook 1926 we observed the rare case of a whole English sentence having been contracted to one token *Ilovetobemothered*. Still, our code-switch detector recognizes this as an English string. [4]

8 Conclusions

We have described our efforts in language identification in a multilingual corpus of Alpine texts. As part of corpus annotation we have identified the language of each corpus sentence amongst English, French, Standard German, Swiss German,

4. The complete sentence is : *Un long Anglais, avec lequel, dans le hall familial, je m'essaie à échanger laborieusement quelques impressions à ce sujet, me dit : ≪I love to be mothered.≫*

Italian and Romansh. Furthermore we have developed an algorithm to identify intra-sentential code-switching by analyzing sentence parts in quotation marks that contain "unknown" lemmas.

We have shown that token sequences that amount to 15 or more characters can be judged by a state-of-the-art language identifier and will result in 78% correctly labeled code-switches. Another 14% are code-switches but with a language different from the auto-assigned language. Only 8% are not code-switches at all.

There are many ways to continue and extend this research. We have not included language identification for Swiss German nor for Romansh in the intra-sentential code-switch experiments reported in this paper. We will train language models for these two languages and add them to Langid to check the impact on the recognition accuracy. Since code-switches into Romansh are rare, and since Romansh can easily be confused with Italian, it is questionable whether the addition of this language model will have a positive influence.

We have used the "general-purpose" language identifier Langid in these experiments. It will be interesting to investigate language identifiers that are optimized for short text fragments as discussed by Vatanen et al. (2010). Given the relatively high number of short quotations (31%) that contain code-switches, recall could improve considerably.

In this paper we have focused solely on code-switching candidates that are triggered by pairs of quotation marks. In order to increase the recall we will certainly enlarge the set of triggers to other indicators such as parentheses or commas. We have briefly looked at parentheses as trigger symbols and found them clearly less productive than quotation marks. To also find code-switches that have no overt marker remains the ultimate goal.

Finally, we will exploit the parallel parts of our corpus. If a sentence in German contains a French segment, then it is likely that this French segment occurs verbatim in the parallel French sentence. Based on sentence and word alignment we will search for identical phrases in both language versions. We hope that this will lead to high accuracy code-switch data that we can use as training material for machine learning experiments.

Acknowledgments

We would like to thank Michi Amsler and Don Tuggener for useful comments on literature and tools for language identification and code-switching, as well as Patricia Scheurer for comments and suggestions on the language use in the SAC corpus. This research was supported by the Swiss National Science Foundation under grant CRSII2_147653/1 through the project "MODERN : Modelling discourse entities and relations for coherent machine translation".

References

Anne Abeillé, Lionel Clément, and Francois Toussenel. 2003. Building a Treebank for French. In Anne Abeillé, editor, *Building and Using Parsed Corpora*, volume 20 of *Text, Speech and Language Technology*, chapter 10, pages 165–187. Kluwer, Dordrecht.

Heike Adel, Ngoc Thang Vu, and Tanja Schultz. 2013. Combination of recurrent neural networks and factored language models for code-switching language modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia.

Sarah Ebling, Rico Sennrich, David Klaper, and Martin Volk. 2011. Digging for names in the mountains : Combined person name recognition and reference resolution for German alpine texts. In *Proceedings of The 5th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznan.

Fei Huang and Alexander Yates. 2014. Improving word alignment using linguistic code switching data. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–9, Göteborg.

Baden Hughes, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew Mackinlay. 2006. Reconsidering language identification for written language resources. In *Proceedings of LREC 2006*, pages 485–488, Genoa.

Marcus Kracht and Udo Klein. 2014. The grammar of code switching. *Journal of Logic, Language and Information*, 23(3) :313–329.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Katrin Rettich. 2013. Automatische Annotation von deutschen und französischen temporalen Ausdrücken im Text+Berg-Korpus. Master thesis, Universität Zürich, Institut für Computerlinguistik.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts.

In *Proceedings of The 18th International Nordic Conference of Computational Linguistics (Nodalida)*, Riga.

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

Tommi Vatanen, Jaakko J. Väyrynen, and Sami Virpioja. 2010. Language identification of short text segments with n-gram models. In *Proceedings of LREC*, pages 3423–3430, Malta.

Martin Volk, Noah Bubenhofer, Adrian Althaus, Maya Bangerter, Lenz Furrer, and Beni Ruef. 2010. Challenges in building a multilingual alpine heritage corpus. In *Proceedings of LREC*, Valletta, Malta.

Martin Volk, Lenz Furrer, and Rico Sennrich. 2011. Strategies for reducing and correcting OCR errors. In C. Sporleder, A. van den Bosch, and K. Zervanou, editors, *Language Technology for Cultural Heritage : Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing, pages 3–22. Springer-Verlag, Berlin.

Ngoc Thang Vu, Heike Adel, and Tanja Schultz. 2013. An investigation of code-switching attitude dependent language modeling. In *Statistical Language and Speech Processing*, pages 297–308. Springer.