Mixed-Language and Code-Switching in the Canadian Hansard

Marine Carpuat
Multilingual Text Processing
National Research Council
Ottawa, ON K1A0R6, Canada
Marine.Carpuat@nrc.gc.ca

Abstract

While there has been lots of interest in code-switching in informal text such as tweets and online content, we ask whether code-switching occurs in the proceedings of multilingual institutions. We focus on the Canadian Hansard, and automatically detect mixed language segments based on simple corpus-based rules and an existing word-level language tagger.

Manual evaluation shows that the performance of automatic detection varies significantly depending on the primary language. While 95% precision can be achieved when the original language is French, common words generate many false positives which hurt precision in English. Furthermore, we found that code-switching does occur within the mixed languages examples detected in the Canadian Hansard, and it might be used differently by French and English speakers.

This analysis suggests that parallel corpora such as the Hansard can provide interesting test beds for studying multilingual practices, including code-switching and its translation, and encourages us to collect more gold annotations to improve the characterization and detection of mixed language and code-switching in parallel corpora.

1 Introduction

What can we learn from language choice patterns observed within multilingual organizations? While this question has been addressed, for instance, by conducting fieldwork in European Union institutions (Wodak et al., 2012), we aim to use natural language processing tools to study language choice directly from text, leveraging the publicly available proceedings of multilingual institutions, which are already widely used for machine translation. Early work on statistical approaches to machine translation (Brown et al., 1990) was made possible by the availability of the bilingual Canadian Hansard in electronic form[1]. Today, translated texts from the Hong Kong Legislative Council, the United Nations, the European Union are routinely used to build machine translation systems for many languages in addition to English and French (Wu, 1994; Koehn, 2005; Eisele and Chen, 2010, inter alia), and to port linguistic annotation from resource-rich to resource-poor languages (Yarowsky et al., 2001; Das and Petrov, 2011, among many others).

As a first step, we focus on detecting code-switching between English and French in the Canadian Hansard corpus, drawn from the proceedings of the Canadian House of Commons. Code-switching occurs when a speaker alternates between the two languages in the context of a single conversation. Since interactions at the House of Commons are public and formal, we suspect that code-switching does not occur as frequently in the Hansard corpus as in other recently studied datasets. For instance, Solorio and Liu (2008) used transcriptions of spoken language conversation, while others focused on informal written genres, such as microblogs and other types of online content (Elfardy et al., 2013; Cotterell et al., 2014). At the same time, the House of Commons is a "bilingual operation where French-speaking and English-speaking staff work together at every level" (Hicks, 2007), so it is not unreasonable to assume that code-switching should occur. In addition, according to the "Canadian Candidate Survey", in 2004, the percentage of candidates for the House of Commons who considered themselves bilingual ranged from 34% in the Conservative

[1] See http://cs.jhu.edu/~post/bitext/ for a historical perspective

Proceedings of The First Workshop on Computational Approaches to Code Switching, pages 107–115,
October 25, 2014, Doha, Qatar.

party to 86% in the Bloc Québecois. The study also shows that candidates have a wide range of attitudes towards bilingualism and the importance of language to their sense of identity (Hicks, 2007). This suggests that code-switching, and more generally language choice, might reveal an interesting range of multilingual practices in the Hansard.

In this paper, we adopt a straightforward strategy to detect mixed language in the Canadian Hansard, using (1) constraints based on the parallel nature of the corpus and (2) a state-of-the-art language detection technique (King and Abney, 2013). Based on this automatic annotation, we conduct a detailed analysis of results to address the following questions:

- How hard is it to detect mixed language in the Canadian Hansard? What are the challenges raised by the Hansard domain for state-of-the-art models?
- Within these mixed language occurrences, does code-switching occur? What kind of patterns emerge from the code-switched text collected?

After introducing the Canadian Hansard corpus (Section 2), we describe our strategy for automatically detecting mixed language use (Section 3). We will see that it is a challenging task: precision varies varies significantly depending on the primary language, and recall is much lower than precision for both languages. Finally, we will focus on the patterns of mixed language use (Section 4): they suggest that code-switching does occur within the mixed language examples detected in the Canadian Hansard, and that it might be used differently by French and English speakers.

2 The Canadian Hansard Corpus

According to Canada's Constitution, "either the English or French language may be used by any person in the debates of the Houses of the Parliament."[2] As a result, speaker interventions can be in French or English, and a single speaker can in principle switch between the two languages.

Our corpus consists of manual transcriptions and translations of meetings of Canada's House of Commons and its committees from 2001 to 2009. Discussions cover a wide variety of topics, and speaking styles range from prepared speeches by a single speaker to more interactive discussions. The part of the corpus drawn from meetings of the House of Commons, is often also called *Hansard*, while *committees* refers to the transcriptions of committee meetings.

This corpus is well-suited to the study of multilingual interactions and their translation for two main reasons. First, the transcriptions are annotated with the original language for each intervention. Second, the translations are high quality direct translations between French and English. In contrast, a French-English sentence pair in the European Parliament corpus (Koehn, 2005) could have been generated from an original sentence in German that was translated into English, and then in turn from English into French. Direct translation eliminates the propagation of "translationese" effects (Volansky et al., 2013), and avoids losing track of code-switching examples by translation into a second or third language.

One potential drawback of working with transcribed text is that the transcription process might remove pauses, repetitions and other disfluencies. However, it is unclear whether this affects mixed language utterances differently than single language ones.

2.1 Corpus Structure and Processing

The raw corpus consists of one file per meeting. The file starts with a header containing meta information about the meeting (event name, type, time and date, etc.), followed by a sequence of "fragments". Each "fragment" corresponds to a short segment of transcribed speech by a single speaker, usually several paragraphs. Fragments are the unit of text that translators work on, so the original language of the fragment is tagged in the corpus, as it determines whether the content should be translated into French or into English. We use the original language tagged as a gold label to define the primary language of the speaker in our study of code-switching.

The raw data was processed using the standard procedure for machine translation data. Processing steps included sentence segmentation and sentence alignment within each fragment, as well as tokenization of French and English. This process yields a total of 8,194,055 parallel sentences. We exclude subsets reserved for the evaluation of machine translation systems, and work with the re-

[2] *Constitution Act*, 1867, formerly the *British North America Act*, 1867, "Appendices", *Revised Statuses of Canada* (RS 1985), s.133.

Data origin	# English segments	# French segments
Committees	4,316,239	915,354
Hansard	2,189,792	738,967
Total	6,506,031	1,654,321

Table 1: Language use by segment

Data origin	# English speakers	# French speakers	# Bilingual speakers
Committees	8787	888	3496
Hansard	198	61	327
Total	8985	949	3823

Table 2: Language use by speaker

maining 8,160,352 parallel segments.[3]

2.2 Corpus-level Language Patterns

English is used more frequently than French: it accounts for 80% of segments, as can be seen in Table 1. The French to English ratio is significantly higher in the Hansard than in the Committees section of the corpus. But how often are both languages used in a single meeting? We use the "DocumentTitle" tags marked in the metadata in order to segment our corpus into meetings. Both French and English segments are found in the resulting 4740 meetings in the committees subcorpus and 927 meetings in the Hansard subcorpus.

How many speakers are bilingual? Table 2 describes language use per speaker per subcorpus. Here, we define a speaker as bilingual if their name is associated with both French and English fragments. Note that this method might overestimate the number of biilingual speakers, as it does not allow us to distinguish between two different individuals with the same name. Overall 22% of speakers are bilingual. The percentage of bilingual speakers in the Hansard (56%) is more than twice that in the Committees (26.5%), reflecting the fact that Hansard speakers are primarily Members of Parliament and Ministers, while speakers that address the Committees represent a much wider sample of Canadian society.

[3]The raw and processed versions of the corpus are both available on request.

3 Automatic Detection of Mixed Language

3.1 Task Definition

We aim to detect code-switching between English and French only. While we found anecdotal evidence of other languages such as Spanish and Italian in the corpus[4], these occurrences seem extremely rare and detecting them is beyond the scope of this study.

We define mixed-language segments as segments which contain words in the language other than their "original language". Recall that the original language is the manually assigned language of the fragment which the segment is part of (Section 2). We want to automatically (1) detect mixed-language segments, and (2) label the French and English words that compose them, in order to enable further processing. These two goals can be accomplished simultaneously by a word-level language tagger.

In a second stage, the automatically detected mixed language segments are used to manually study code-switching, since our mixed language tagger does not yet distinguish between code-switching and other types of mixed language (e.g., borrowings).

3.2 Challenges

When the identity of the languages mixed is known, the state-of-the-art approach to word-level language identification is the weakly supervised approach proposed by King and Abney (2013). They frame the task as a sequence labeling problem with monolingual text samples for training data. A Conditional Random Field (CRF) trained with generalized expectation criteria performs best, when evaluated on a corpus comprising 30 languages, including many low resources languages such as Azerbaijani or Ojibwa.

In our case, there are only two high-resource languages involved, which could make the language detection task easier. However, the Hansard domain also presents many challenges: English and French are closely related languages and share many words; the Hansard corpus contains many occurrences of proper names from various origins which can confuse the language detector; the corpus is very large and unbalanced as we expect the vast majority of segments to be monolingual.

[4]e.g., "merci beaucoup, thank you very much, grazie mille"

To address these challenges, we settled on a two pass approach: (1) select sentences that are likely to contain mixed language, and (2) apply CRF-based word-level language tagging to the selected sentences.

3.3 Method: Candidate Sentence Selection

We select candidates for mixed language tagging using two complementary sources of information:

- frequent words in each language: a mixed-language segment is likely to contain words that are known to be frequent in the second language. For instance, if a segment produced by a French speaker contains the string "of", which is frequent in English, then it is likely to be a mixed language utterance.
- parallel nature of corpus: if a French speaker uses English in a predominantly French segment, the English words used are likely to be found verbatim in the English translation. As a result, overlap[5] between a segment and its translation can signal mixed language.

We devise a straightforward strategy for selecting segments for word-level language tagging:

1. identify the top 1000 most frequent words on each side of the parallel Hansard corpus.
2. exclude words that occur both in the French and English list (e.g., the string "on" can be both an English preposition and a French pronoun)
3. select originally French sentences where (a) at least one word from the English list occurs, and (b) at least two words from the French sentence overlap with the English translation
4. select originally English sentences in the same manner.

3.4 Method: Word-level Language Tagging

The selected segments are then tagged using the CRF-based model proposed by King and Abney (2013). It requires samples of a few thousand words of French and English for training. How can we select samples of English and French that are strictly monolingual?

We solve this problem by leveraging the parallel nature of our corpus again: We assume that a segment is strictly monolingual if there is no overlap in vocabulary between a segment and its translation. Using this approach, we randomly select a sample of 1000 monolingual French segments and 1000 monolingual English segments. This yields about 21k/4k word tokens/types for English, and 24k/4.6k for French. Using these samples, we apply the CRF approach on each candidate sentence selected during the previous step. For the low resource languages used by King and Abney (2013), the training samples were much smaller (in the order of hundreds of words per language), and learning curves suggest that the accuracy reaches a plateau very quickly. However, we decide to use larger samples since they are very easy to construct in our large data setting.

[5]Except for numbers, punctuation marks and acronyms.

fr mixed in en	gold pos.	gold neg.	total
predicted pos.	21	8	29
predicted neg.	1	109	110
total	22	117	139

Table 4: Confusion matrix for detecting segments containing French words when English is the original language. It yields a Precision of 95.4% and a Recall of 72.4%

en mixed in fr	gold pos.	gold neg.	total
predicted pos.	3	1	4
predicted neg.	13	105	118
total	16	106	122

Table 5: Confusion matrix for detecting segments containing English words when French is the original language. It yields a Precision of 75% and a Recall of 18.75%

3.5 Evaluation

At this stage, we do not have any gold annotation for code-switching or word-level language identification on the Hansard corpus. We therefore ask a bilingual human annotator to evaluate the precision of the approach for detecting mixed language segments on a small sample of 100 segments for each original language. The annotator tagged each example with the following information: (1) does the segment actually contain mixed language? (2) are the language boundaries correctly detected? (3) what does the second language express? (e.g., organization name, idiomatic expression, quote, etc. The annotator was not given predefined categories) . Table 3 provides annotation examples.

Tagged Lang.	**[FR** Et le premier ministre nous répond que**] [EN** a farmer is a farmer a Canadian is a Canadian**] [FR** d' un bout à l' autre du Canada**]**
Gold Lang.	**[FR** Et le premier ministre nous répond que**] [EN** a farmer is a farmer a Canadian is a Canadian**] [FR** d' un bout à l' autre du Canada**]**
Evaluation	Mixed-language segment? **yes** Are boundaries correct? **yes** What is the L2 content? **quote**
Tagged Lang.	**[FR** Autrement**] [EN** dit they are getting out of the closet**] [FR** parce que cela leur donne le droit d avoir deux enfants**]**
Gold Lang.	**[FR** Autrement dit**] [EN** they are getting out of the closet**] [FR** parce que cela leur donne le droit d avoir deux enfants**]**
Evaluation	Mixed-language segment? **yes** Are boundaries correct? **no** What is the L2 content? **idiom**

Table 3: Example of manual evaluation: the human annotator answers three questions for each tagged example, based on their knowledge of what the gold language tags should be.

2-step	*committees*		*Hansard*	
detection	en	fr	en	fr
Selection	62,069	13,278	42,180	13,558
Tagger	7,713	317	3,993	164

Table 6: Number of mixed-language segments detected by each automatic tagging stage, as described in Section 3.

Based on this gold standard, we can first evaluate the performance of the segment-level mixed language detector (Task (1) as defined in Section 3.1). Confusion matrices for English and French sentences are given in Tables 5 and 4 respectively. The gold label counts confirm that the classes are very unbalanced, as expected.

The comparison of the predictions with the gold labels yields quite different results for the two languages. On English sentences, the mixed language tagger achieves a high precision (95.4%) at a reasonable level of recall (72.4%) , which is encouraging. However, on French sentences, the mixed language tagger achieves a slightly lower precision (75%) with an extremely low recall (18.75%). These scores are computed based on a very small number of positive predictions by the tagger (4 only) on the sample of 100+ sentences. Nevertheless, these results suggest that, while we might miss positive examples due to the low recall, the precision of the mixed language detector is sufficiently high to warrant a more detailed study of the examples of mixed language detected.

lang	corpus	detection precision	segmentation precision
en	committees	72.6%	44.4%
	Hansard	45.9%	28.6%
fr	committees	98.4%	67.7%
	Hansard	96.8%	75.4%

Table 7: Evaluation of positive predictions: precision of mixed language detection at the segment level, and precision of the language segmentation (binary judgment on accuracy of predicted language boundaries for each segment.)

4 Patterns of Mixed Language Use

Discovering patterns of mixed language use, including code-switching, requires a large sample of mixed language segments. Since the gold standard constructed for the above evaluation (Section 3) only provides few positive examples, we ask the human annotator to apply the annotation procedure illustrated in Table 3 to a sample of positive predictions: French segments where the tagger found English words, and vice versa.

The number of positive examples detected can be found in Table 6. Only a small percentage of the original corpus is tagged as positive, but given that our corpus is quite large, we already have more than 10,000 examples to learn from.

The human annotator annotated a random sample of 60+ examples for each original language and corpus partition. The resulting precision

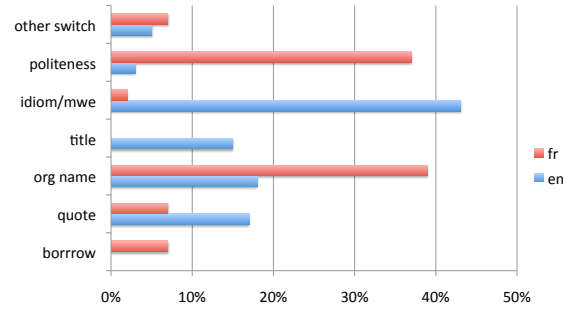

Figure 1: Categories of mixed language use observed depending on the original language of the segment, in the *committees* data

scores, both for mixed language detection at the segment level, and for accurately tagging words with French and English, are given in Table 7. For segment level detection, the precision is much higher on French than on English, as observed previously. On English data, the annotation reveals that most false positives are due to frequent words that occur both in languages (e.g., "province","Premier","plus"), and are incorrectly tagged as French in our English segment. The boundaries of French and English segments are correctly detected for up to 75% of French segments, but only for 44% at best in English segments. More work is therefore needed to accurately tag languages at the word-level. Some of the second language words are usually detected, but the boundaries are often wrong, especially at code-switching points.

In addition to correctness, the annotator was asked to identify the kind of information conveyed by the second language, and they came up with categories that reflected the patterns that emerged from the examples. Examples of these categories are given for each language in Table 8, and the percentage of examples observed per category for each language are plotted in Figure 1.

While many correctly detected mixed language segments are due to borrowings, use of organization names or titles in the other language, we do find examples of code switching such as:

- quotes
- multiword expressions or idioms,
- politeness formulas and formality.

The distribution of code-switching across these categories is very different for French and English as original languages. Multiword expressions and idioms account for more than 40% of English use in French segments, while there are no examples of French idioms in English segments. Conversely, while politeness formulas in French account for more than 30% of correctly detected mixed language use in English segments, there are only fewer than 5% such instances in French. This might suggest that French speakers who code-switch are more proficient in English than English speakers in French, or that code-switching is used for different purposes by English and French speakers in the Hansard context.

While more analysis is definitely needed to better understand code-switching patterns and their use, we have established that code-switching occurs in the Hansard corpus, and that it might be used differently by French and English speakers.

In the parallel corpus, different types of mixed language are handled differently by human translators, which suggests that machine translation of code-switched data requires specific strategies: while English idioms, quotes or named entities in a French segment might be directly copied to the output when translating into English, other categories should be handled diffferently. For instance, mixed language that discusses translation of terms might require to *avoid translating* the original French terms in order not to lose the original meaning in translation. When English is used in politeness, the reference translations often perform a *normalization* of titles and capitalization. In that case, copying the English segments in the French sentence to the MT output would produce translations that are understandable, but would not match the conventions used in the reference.

5 Related Work

To the best of our knowledge, this is the first study of mixed language and code-switching in the Canadian Hansard parallel corpus, a very large parallel corpus commonly used to build generic machine translation systems.

Previous work at the intersection of *machine translation* and *mixed languages* has focused on specific application scenarios: word translation disambiguation for mixed language queries (Fung et al., 1999), or building applications to help second language learners, such as translating of short L1 phrases in sentences that are predominantly

	Use of English in primarily French segments
Quote	**[FR** C' est écrit**]** "**[EN** will have full access**]** "
	[FR Vous avez dit et je vous cite**]** " **[EN** we do not have to change the definition of marriage to protect equality rights**]** "
Translation	**[FR** On parle en anglais de**] [EN** carrots and sticks**]**
	[FR Milliard correspond à**] [EN** billion**] [FR** en anglais**]**
Politeness	**[FR** Nous accueillons ce matin M Brulé M Baines M McDougall et M Mann**]**
	[EN Welcome to all of you**]**
	[EN Thank you Mr Chair**] [FR** Merci beaucoup**]**
Idioms/MWEs	**[FR** Le contraire ne m avait jamais été dit**] [EN** by the way**]**
	[FR Oui en français**] [EN** as well**]**
Title	**[FR** Je cite l auteur israélien Simha Flapan dans l ouvrage**] [EN** The Birth of Israel**]**
	[FR Des courts métrages présents dans la compétition officielle**] [EN** The stone of folly**] [FR** a nettement été le film préféré du public**]**
Organization	**[FR** La**] [EN** Western Canadian Wheat Growers Association**] [FR** est une association de producteurs**]**
	[FR M Thomas Axworthy l ancien président du**] [EN** Centre for the Study of Democracy**] [FR** s y trouvait aussi**]**
Other	**[FR** Alors en ce moment le comité est maître de sa propre procédure pour étudier cette question importante**] [EN** this breach of its own privileges which appears to have taken place**]**
	[FR Merci aux collègues**] [EN** who gave me this opportunity**]**
	Use of French in primarily English segments
Quote	**[EN** The great French philosopher Blaise Pascal spoke of the essence of human life as a gamble**] [FR** un pari **] [EN** and so it is in political life**]**
	[EN You mentioned**] [FR** les fusions**] [EN** but I gather that**] [FR** les défusions**]**
	[EN is now the order of the day in Quebec**]**
Translation	**[EN** The French text had a small error in that it used the word**] [FR** aux**] [EN** where the word**] [Fr** des**] [EN** should have been used**]**
	[EN Mr Speaker to teach is to open doors to a better world in French**] [FR** enseigner ouvre les portes vers un monde meilleur**]**
Politeness	**[EN** Thank you Mr Chairman**] [FR** monsieur le président**] [EN** honourable members**] [FR** mesdames et messieurs**]**
	[EN On this important traditional Chinese holiday**] [FR** bonne année à toute la communauté canadienne**] [EN** I wish all Canadians health happiness and prosperity in the year of the ox**]**
Idioms/MWEs	**[EN** We were the first ones to start to ask about it and we are following**] [FR** à la lettre**] [EN** as we say in French
	[EN So that s just to**][FR** entrer en matière**]**
Borrowing	**[EN** We think it fundamentally adjusts the loss of culture and language which was the**] [FR** raison d'être**] [EN** of the residential school program**]**
	[EN Everything is a**] [FR** fait accompli**]**
Organization	**[EN** That s a fair question and I d like to thank Mr Blaney for participating in the**] [FR** Forum socioéconomique des Premières Nations**]**
	[EN If the **[EN** Bloc Québécois**] [EN** brings forward a witness you may want to go to them first**]**
Other	**[EN** The same committee rejected an amendment**] [FR** proposé par le Bloc québécois proposé par moi pour le NPD**]**
	[EN This is not the current government**] [FR** C est la même chose**] [EN** it doesn t matter which one is in power**]**

Table 8: Examples of mixed language segments

L2[6] (van Gompel and van den Bosch, 2014), or on detecting code-mixing to let an email translation system handle words created on the fly by bilingual English-Spanish speakers (Manandise and Gdaniec, 2011). While code-switched data is traditionally viewed as noise when training machine translation systems, Huang and Yates (2014) showed that appropriately detecting code-switching can help inform word alignment and improve machine translation quality.

There has been renewed interest on the study of mixed language recently, focusing on detecting code-switching points (Solorio and Liu, 2008; Elfardy et al., 2013) and more generally detecting mixed language documents. Lui et al. (2014) use a generative mixture model reminiscent of Latent Dirichlet Allocation to detect mixed language documents and the languages inside them. Unlike the CRF-based approach of King and Abney (2013), the languages involved do not need to be known ahead of time. In contrast with all these approaches, we work with parallel data with unbalanced original languages.

6 Conclusion

We investigated whether code-switching occurs in the Canadian Hansard parallel corpus.

We automatically detected mixed language segments using a two-step approach: (1) candidate sentence selection based on frequent words in each language and overlap between the two side of the parallel corpus, and (2) tag each word in the segment as French or English using the CRF-based approach of King and Abney (2013).

Manual evaluation showed that automatic detection can be done with high precision when the original language is French, but common words generate many false positives which hurt precision in English. More research is needed to improve recall, which is lower than precision in both languages, and particularly low when the original language is French. Further analysis reveals that code-switching does occur within the mixed language examples detected in the Canadian Hansard, and suggests that it is used differently by French and English speakers.

While much work is still needed to construct larger evaluation suites with gold annotations, and improving the detection and tagging of mixed language sentences, this work suggests that the proceedings of multilingual organizations such as the Canadian Hansard can provide interesting test beds for (1) corpus-based study of language choice and code-switching, which can complement the direct observation of meetings, as conducted by Wodak et al. (2012), and (2) investigating the interactions of code-switching and machine translation. Furthermore, it would be interesting to study how code-switching in the Hansard differs from code-switching in more informal settings.

[6] http://alt.qcri.org/semeval2014/task5/

References

Peter Brown, John Cocke, Stephen Della Pietra, Vincent Della Pietra, Frederik Jelinek, John Lafferty, Robert Mercer, and Paul Rossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16:79–85.

Ryan Cotterell, Adithya Renduchintala, Naomi Saphra, and Chris Callison-Burch. 2014. An Algerian Arabic-French code-switched corpus. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*. May.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 600–609, Stroudsburg, PA, USA.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from United Nation documents. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 2868–2872, 5.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2013. Code switch point detection in Arabic. In *Natural Language Processing and Information Systems*, pages 412–416. Springer.

Pascale Fung, Xiaohu Liu, and Chi Shun Cheung. 1999. Mixed Language Query Disambiguation. In *Proceedings of ACL'99*, Maryland, June.

Bruce Hicks. 2007. Bilingualism and the Canadian house of commons 20 years after B and B. In *Parliamentary Pespectives*. Canadian Study of Parliament Group.

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, Gothenburg, Sweden, April.

Ben King and Steven Abney. 2013. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of*

the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1110–1119, Atlanta, Georgia, June.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit X*, Phuket, Thailand, September.

Marco Lui, Jey Han Lau, and Timothy Baldwin. 2014. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*.

Esmé Manandise and Claudia Gdaniec. 2011. Morphology to the rescue redux: Resolving borrowings and code-mixing in machine translation. *Systems and Frameworks for Computational Morphology*, pages 86–97.

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

Maarten van Gompel and Antal van den Bosch. 2014. Translation assistance by translation of L1 fragments in an L2 context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 871–880, Baltimore, Maryland, June.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Literary and Linguistic Computing*.

Ruth Wodak, Michal Krzyzanowski, and Bernhard Forchtner. 2012. The interplay of language ideologies and contextual clues in multilingual interactions: Language choice and code-switching in European Union institutions. *Language in Society*, 41:157–186.

Dekai Wu. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 80–87, Stroudsburg, PA, USA.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, pages 1–8, Stroudsburg, PA, USA.