A Code-Switching Corpus of Turkish-German Conversations

Özlem Çetinoğlu
IMS, University of Stuttgart
Germany
ozlem@ims.uni-stuttgart.de

Abstract

We present a code-switching corpus of Turkish-German that is collected by recording conversations of bilinguals. The recordings are then transcribed in two layers following speech and orthography conventions, and annotated with sentence boundaries and intersentential, intrasentential, and intra-word switch points. The total amount of data is 5 hours of speech which corresponds to 3614 sentences. The corpus aims at serving as a resource for speech or text analysis, as well as a collection for linguistic inquiries.

1 Introduction

Code-switching (CS) is mixing two (or more) languages in spoken and written communication (Myers-Scotton, 1993; Poplack, 2001; Toribio and Bullock, 2012) and is quite common in multilingual communities (Auer and Wei, 2007). With the increase in multilingual speakers worldwide, CS becomes more prominent.

In parallel, the interest in processing mixed language is on the rise in the Computational Linguistics community. Researchers work on core tasks such as normalisation, language identification, language modelling, part-of-speech tagging as well as downstream ones such as automatic speech recognition and sentiment analysis (Çetinoğlu et al., 2016). The majority of the corpora used in these tasks come from social media (Nguyen and Doğruöz, 2013; Barman et al., 2014; Vyas et al., 2014; Solorio et al., 2014; Choudhury et al., 2014; Jamatia et al., 2015; Samih and Maier, 2016; Vilares et al., 2016; Molina et al., 2016).

Social media has the advantage of containing vast amount of data and easy access. Depending on the medium, however, limitations might arise. For instance, Twitter, the most popular source so far, allows the distribution of tweet IDs rather than tweets themselves, which can be deleted. Hence it is hard to use the full resource, reproduce previous results or compare to them. Moreover the character limit and idiosyncratic language of social media bring extra challenges of processing in addition to challenges coming from code-switching.

Spoken data has also been a popular source in computational CS research (Solorio and Liu, 2008; Lyu and Lyu, 2008; Chan et al., 2009; Shen et al., 2011; Li et al., 2012; Lyu et al., 2015; Yilmaz et al., 2016). There are no limitations on the length of sentences, idiosyncrasies are less pronounced. Despite such advantages, it is almost solely used in speech analysis. To our knowledge, only Solorio and Liu (2008) have used transcriptions of CS speech in text analysis. One reason that researchers processing CS text prefer social media could be that it is already text-based, and it requires much less time and effort than speech collection transcription. For the existing speech corpora, discrepancies between the speech transcriptions and the input text processing tools expect could be a drawback. For instance the SEAME corpus (Lyu et al., 2015) does not use punctuation, capitalisation, or sentence boundaries in transcriptions, yet standard text processing tools (POS taggers, morphological analysers, parsers) are trained on edited text, hence make use of orthographic cues.

In this paper, we introduce a Turkish-German code-switching corpus of conversations and their two layers of transcriptions following speech and orthography conventions. The data is annotated with sentence boundaries and intersentential, intrasentential, and intra-word switch points. Our aim is to provide a resource that could be used by researchers from different backgrounds, e.g., for speech recognition and language identification in

Proceedings of the 11th Linguistic Annotation Workshop, pages 34–40,
Valencia, Spain, April 3, 2017. ©2017 Association for Computational Linguistics

speech, for language identification and predicting CS points in text, and as a corpus of empirical evidence for linguistically interesting structures.

2 Related Work

Creating code-switching corpora for speech analysis has started with reading designed text rather than spontaneous speech. Lyu and Lyu (2008) use a Mandarin-Taiwanese test set for their language identification system that consist of 4.8 hours of speech corresponding to 4600 utterances. The set is designed to have Mandarin as the main language with one or two Taiwanese words replaced with their Mandarin counterparts. Chan et al. (2009) introduce a Cantonese-English corpus of read speech of 3167 manually designed sentences. English is inserted into Cantonese as segments of one or more words. Another read speech corpus is created by Shen et al. (2011) for Mandarin-English and consists of 6650 utterances. Li et al. (2012) collected 5 hours of code-switched Mandarin-English speech from conversational and project meetings. Intersentential and intrasentential switches add up to 1068 in total.

Lyu et al. (2015) present the largest CS speech resource, the SEAME corpus, which has 192 hours of transcribed Mandarin-English interviews and conversations in the latest version.[1] The code-switching points naturally occur in the text, as both languages are written in their own scripts. A recent corpus of 18.5 hours is introduced by Yilmaz et al. (2016) on Frisian-Dutch broadcasts. CS points are marked in the transcriptions but not on the audio level.

Solorio and Liu (2008) recorded a conversation of 40 minutes among Spanish-English bilinguals. The transcribed speech contains 922 sentences with 239 switch points among them. The authors used this data to train machine learning algorithms that predict CS points of an incrementally given input.

Speech collections have always been the primary source in sociolinguistic and pyscholinguistic research. We list some of these spoken corpora that employ code-switching instances of Turkish and German, mixed with other languages or with each other. The “Emigranto” corpus (Eppler, 2003) documents conversations with Jewish refugees settled in London in 1930s, who mix Austrian German with British English. In this corpus, Eppler (2011) looks into mixed dependencies where a dependent and its head are from different languages. She observes that dependents with a mixed head have on average longer dependencies than ones with a monolingual head.

In a similar fashion, Tracy and Lattey (2009) present more than 50 hours of recordings of elderly German immigrants in the U.S. The data is fully transcribed and annotated, yet each session of recordings is transcribed as a single file with no alignment between transcript utterences and their corresponding audio parts, and annotations use Microsoft Word markings, e.g. bold, italic, underline, or different font sizes, thus require format conversions to be processed by automatic tools that accept text-based inputs.

Kallmeyer and Keim (2003) investigate the communication characteristics between young girls in Mannheim, mostly of Turkish origin, and show that with peers, they employ a mixed form of Turkish and German. Rehbein et al. (2009) and Herkenrath (2012) study the language acquisition of Turkish-German bilingual children. On the same data Özdil (2010) analyses reasons of code-switching decisions. The Kiezdeutsch corpus (Rehbein et al., 2014) consists of conversations among native German adolescents with a multiethnic background, including Turkish. As a result, it also contains a small number of Turkish-German mixed sentences.

3 Data

The data collection and annotation processes are handled by a team of five Computational Linguistics and Linguistics students. In the following sections we give the details of these processes.

3.1 Collection

The data collection is done by the annotators as conversation recordings. We asked the annotators to approach Turkish-German bilinguals from their circle for an informal setting, assuming this might increase the frequency of code-switching. Similarly we recommended the annotators to open topics that might induce code-switching, such as work and studies (typically German-speaking environments) if a dialogue started in Turkish, or Turkish food and holidays in Turkey (hence Turkish-specific words) in a German-dominated conversation.

[1] https://catalog.ldc.upenn.edu/LDC2015S04

28 participants (20 female, 8 male) took part in the recordings. The majority of the speakers are university students. Their ages range from 9 to 39, with an average of 24 and a mode of 26. We also asked the participants to assign a score from 1 to 10 for their proficiency in Turkish and German. 18 of the participants think their German is better, 5 of them think their Turkish is better, and the remaining 5 assigned an equal score. The average score for German is 8.2, and for Turkish 7.5.[2]

3.2 Annotation

The annotation and transcriptions are done using Praat.[3] We created six tiers for each audio file: `spk1_verbal`, `spk1_norm`, `spk2_verbal`, `spk2_norm`, `lang`, `codesw`. The first four tiers contain the verbal and normalised transcription of speakers 1 and 2. The tier `lang` corresponds to the language of intervals and can have TR for Turkish, DE for German, and LANG3 for utterances in other languages. The first five tiers are intervals, while the last one is a point tier that denotes sentence and code-switching boundaries. The labels on the boundaries are SB when both sides of the boundary are in the same language, SCS when the language changes from one sentence to the next (intersentential), WCS when the switch is between words within a sentence (intrasentential). Figure 1 shows a Praat screenshot that demonstrates the tiers and exemplifies SCS and WCS boundaries.

Since Turkish is agglutinative and case markers determine the function of NPs, non-Turkish common and proper nouns with Turkish suffixes are commonly observed in CS conversations. We mark such words in the `codesw` tier as a intra-word switch and use the symbol § following Çetinoğlu (2016). Example (1) depicts the representation of a mixed word where the German compound *Studentenwohnheim* 'student accommodation' is followed by the Turkish locative case marker *-da* (in bold).

(1) Studentenwohnheim § **da**
student accommodation § **Loc**
'in the student accommodation'

For many proper names, Turkish and German orthography are identical. Here, the speech data in parallel becomes an advantage, and the language is decided according to the pronunciation. If the proper name is pronounced in German, and followed by a Turkish suffix a § switch point is inserted. Otherwise it follows Turkish orthography.

3.3 Transcription

For speech analysis it is important to transcribe utterances close to how they are pronounced. In some transcription guidelines, capitalisation and punctuation are omitted (e.g. in the SEAME corpus (Lyu et al., 2015)[4]), in some others they are used to mark speech information (e.g. in the Kiezdeutsch corpus (Rehbein et al., 2014)[5]). Text analysis on the other hand generally relies on standard orthography. This raises a conflict between two tasks on how to transcribe speech. To avoid this problem, we introduced two tiers of transcription. The verbal tier follows the speech conventions. If a speaker uses a contraction, the word is transcribed as contracted. The acronyms are written as separate characters. Numbers are spelled out. Recurring characters are represented with the single character followed by a colon. The normalised tier follows the edited text conventions. Words obey the orthographic rules of standard Turkish and German, e.g. characters of acronyms are merged back. Punctuation is added to the text, obeying the tokenisation standards (i.e. separated from the preceding and following tokens with a space).

Example (2) gives a sentence showing the verbal and normalised tiers for a Turkish sentence. The *r* sound in the progressive tense suffix *-yor* is not pronounced, hence omitted in the verbal tier. The vowel of the interjection *ya* is extended during speech, and the colon representation is used to reflect it in the verbal tier, yet the normalised tier has the standard form. Also, the question mark is present in the normalised tier.

(2)				
`verbal:`	ne	diyosun	ya**:**	
`norm:`	Ne	diyo**r**sun	ya	?
	What	say.Prog.2PSg	Intj.	

'What do you say??'

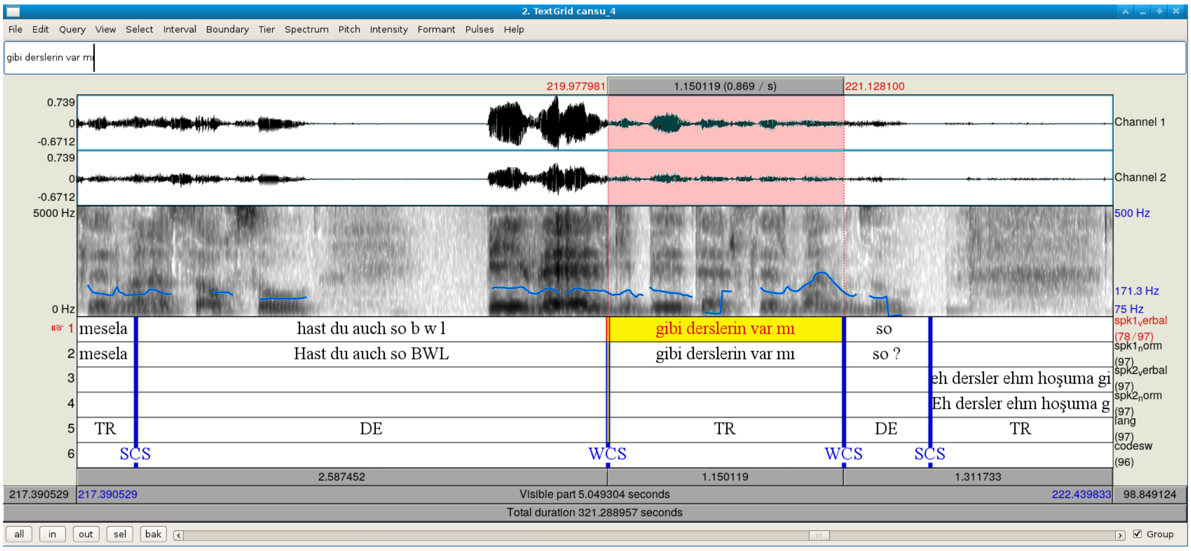
If a made-up word is uttered, it is preceded with an asterisk mark in the transcription. Note that dialectal pronunciation or using a valid word in

[2] The metadata is also available in the CMDI format at the IMS Clarin repository.

[3] www.fon.hum.uva.nl/praat

[4] `https://catalog.ldc.upenn.edu/docs/LDC2015S04/SEAME.V4.0.pdf`

[5] `http://www.kiezdeutschkorpus.de/files/kidko/downloads/KiDKo-Transkriptionsrichtlinien.pdf`



gibi derslerin var mı

1 mesela | hast du auch so b w l | gibi derslerin var mı | so

2 mesela | Hast du auch so BWL | gibi derslerin var mı | so ?

3 eh dersler ehm hoşuma gi

4 Eh dersler ehm hoşuma g

5 TR | DE | TR | DE | TR

6 SCS | WCS | WCS | SCS

Figure 1: A screenshot example from Praat annotations. It shows part of a Turkish sentence and a full mixed sentence from speaker 1, and part of a Turkish sentence from speaker 2.

wrong context is not considered within this class. Partial words are marked with two hyphens instead of the common use of one hyphen, as the latter is used in German to denote the initial part of a compound when two compounds share a common part and the first compound is written only as the unshared part (e.g. *Wohn- und Schlafzimmer* 'living room and bedroom').

We also marked `[silence]`, `[laugh]`, `[cough]`, `[breathe]`, `[noise]`, and put the remaining sounds into the `[other]` category. Overlaps occur usually when one speaker is talking and the other is uttering backchannel signals and words of acknowledgement. There are also cases both speakers tend to speak at the same time. In all such cases, both voices are transcribed, one speaker is chosen to be the main speaker, and an `[overlap]` marker is inserted to the secondary speaker's verbal and normalised tiers. The `codesw` and `lang` tiers are decided according to the main speaker's transcription.

3.4 Quality Control

Once the Praat annotation is completed its output files are converted to a simpler text format for easier access from existing tools and for easier human readability. [6] We ran simple quality control scripts that check if all the tiers are present and non-empty, if the `lang` and `codesw` tiers have values from their label sets, and if the `lang` and `codesw` labels are meaningful, for instance, if there are TR labels on both sides of a SCS (intersentential CS) boundary, either the boundary should be corrected to SB or one of the language labels should be DE or LANG3. Any mistakes are corrected by the annotators on a second pass.

For the quality control of the transcriptions we employed Turkish and German morphological analysers (Oflazer, 1994; Schmid et al., 2004) and analysed all the tokens in the normalised tier according to their languages. We then created a list of tokens unknown to the analysers, which are potentially mispelled words. The annotators went through the list and corrected their mistakes in both the verbal and normalised tiers. The remaining list also gives us the words unknown to the morphological analysers.

4 Statistics and Observations

The durations of recordings range from 20 seconds to 16 minutes. There are 47 transcribed files with a total of 5 hours. Each file is accompanied with a metadata file that contains speaker information, that could be used to filter the corpus according to age intervals, education levels, language proficiency etc.

Table 1 gives the basic statistics on the normalised version of the transcriptions. The token count includes punctuation and interjections, and excludes paralinguistic markers and overlaps.

[6] The format of the text files is given with an example in `http://www.ims.uni-stuttgart.de/institut/mitarbeiter/ozlem/LAW2017.html` The script that converts Praat .TextGrid files to that format is also provided.

sentences	3614
tokens	41056
average sent. length	11.36
sentence boundaries (SB)	2166
intersentential switches (SCS)	1448
intrasentential switches (WCS)	2113
intra-word switches (§)	257
switches in total	3818
sent. with at least one WCS	1108

Table 1: Basic statistics about the data.

Switch	Language Pair	#	%
SB	DE → DE	1356	62.60
	TR → TR	809	37.35
	LANG3 → LANG3	1	0.05
SCS	TR → DE	754	52.07
	DE → TR	671	46.34
	LANG3 → TR	7	0.48
	LANG3 → DE	6	0.41
	DE → LANG3	5	0.35
	TR → LANG3	5	0.35
WCS	TR → DE	1082	51.20
	DE → TR	914	43.26
	DE → LANG3	34	1.61
	TR → LANG3	31	1.47
	LANG3 → DE	28	1.33
	LANG3 → TR	24	1.14
§	DE → TR	246	95.72
	LANG3 → TR	11	4.28

Table 2: Breakdown of switches from one language to another, and their percentages within their switch type.

Switch points split mixed tokens into two in the transcriptions for representational purposes, but they are counted as one token in the statistics.

The majority of the switches are intrasentential and the language of the conversation changes when moving from one sentence to another in 40% of the time. They also correspond to the 55.3% of all switches. 38% of them happen between words, and the remaining 6.7% are within a word. Table 2 shows the breakdown of switches. There are 614 overlaps and 648 paralinguistic markers.[7]

We have observed that many CS instances fall into the categories mentioned in Çetinoğlu (2016), like German verbs coupled with Turkish light verbs *etmek* 'do' or *yapmak* 'make'; Turkish lexicalised expressions and vocatives in German sentences, and vice versa; subordinate clauses and conjuctions in the one language while the remaining of the sentence is in the other language. One category we have seen more prominent in speech data is non-standard syntactic constructions, perhaps due to spontaneity. For instance, Example (3), which is also given as Figure 1, is a question with two verbs (Turkish in bold). Both German *hast du* and Turkish *var mı* corresponds to 'do you have'.

(3)

Hast du	auch	so	BWL	**gibi**
Have you	also	like	business studies	**like**

derslerin	**var**	**mı**	so?
class.Poss2Sg	**exist**	**Ques**	like?

'Do you also have classes like business studies?'

5 Conclusion

We present a corpus collected from Turkish-German bilingual speakers, and annotated with sentence and code-switching boundaries in audio files and their corresponding transcriptions which are carried out as both verbal and normalised tiers. In total, it is 5 hours of speech and 3614 sentences.

Transcriptions are available for academic research purposes.[8] The licence agreement can be found at `http://www.ims.uni-stuttgart.de/institut/mitarbeiter/ozlem/LAW2017.html` along with transcription examples. Audio files will be manipulated before distribution in order to conceal speakers' identity, to comply with the German data privacy laws[9].

Acknowledgements

We thank Sevde Ceylan, Uğur Kostak, Hasret el Sanhoury, Esra Soydoğan, and Cansu Turgut for the data collection and annotation. We also thank anonymous reviewers for their insightful comments. This work was funded by the Deutsche Forschungsgemeinschaft (DFG) via SFB 732, project D2.

References

Peter Auer and Li Wei. 2007. *Handbook of multilingualism and multilingual communication*, volume 5. Walter de Gruyter.

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

[7] laugh: 279, noise: 148, silence: 113, breath: 74, other: 25, cough: 9.

[8] If parts of data contain information that can reveal the identity of a specific person, they are anonymised.

[9] National: `https://www.datenschutz-wiki.de/Bundesdatenschutzgesetz`, State: `https://www.baden-wuerttemberg.datenschutz.de/landesdatenschutzgesetz-inhaltsverzeichnis/`

pages 13–23, Doha, Qatar, October. Association for Computational Linguistics.

Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of computational processing of code-switching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 1–11, Austin, Texas, November. Association for Computational Linguistics.

Özlem Çetinoğlu. 2016. A Turkish-German code-switching corpus. In *The 10th International Conference on Language Resources and Evaluation (LREC-16)*, Portorož, Slovenia.

Joyce YC Chan, Houwei Cao, PC Ching, and Tan Lee. 2009. Automatic recognition of Cantonese-English code-mixing speech. *Computational Linguistics and Chinese Language Processing*, 14(3):281–304.

Monojit Choudhury, Gokul Chittaranjan, Parth Gupta, and Amitava Das. 2014. Overview of FIRE 2014 track on transliterated search. In *Forum for Information Retrieval Evaluation*, Bangalore,India, December.

Eva Eppler. 2003. Emigranto data: A dependency approach to code-mixing. In Pereiro Carmen C., Anxo M.L. Suarez, and XoÃąn P Rodriguez-Yanez, editors, *Bilingual communities and individuals.*, pages 652–63. Vigo: Servicio de Publicacions da Universidade de Vigo, 1.

Eva Duran Eppler. 2011. The dependency distance hypothesis for bilingual code-switching. In *Proceedings of the International Conference on Dependency Linguistics*.

Annette Herkenrath. 2012. Receptive multilingualism in an immigrant constellation: Examples from Turkish–German children's language. *International Journal of Bilingualism*, pages 287–314.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2015. Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 239–248, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Werner Kallmeyer and Inken Keim. 2003. Linguistic variation and the construction of social identity in a German-Turkish setting. *Discourse constructions of youth identities*, 110:29.

Ying Li, Yue Yu, and Pascale Fung. 2012. A Mandarin-English code-switching corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2515–2519, Istanbul, Turkey, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1573.

Dau-Cheng Lyu and Ren-Yuan Lyu. 2008. Language identification on code-switching utterances using multiple cues. In *Interspeech*, pages 711–714.

D.-C. Lyu, T.-P. Tan, E.-S. Chng, and H. Li. 2015. Mandarin–English code-switching speech corpus in South-East Asia: SEAME. *LRE*, 49(3):581–600.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, Texas, November. Association for Computational Linguistics.

C. Myers-Scotton. 1993. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Dong Nguyen and A. Seza Doğruöz. 2013. Word level language identification in online multilingual communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 857–862, Seattle, Washington, USA, October. Association for Computational Linguistics.

Kemal Oflazer. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2):137–148.

Erkan Özdil. 2010. *Codeswitching im zweisprachigen Handeln*. Waxmann Verlag.

Shana Poplack. 2001. Code-switching (linguistic). *International encyclopedia of the social and behavioral sciences*, pages 2062–2065.

Jochen Rehbein, Annette Herkenrath, and Birsel Karakoç. 2009. Turkish in Germany on contact-induced language change of an immigrant language in the multilingual landscape of europe. *STUF-Language Typology and Universals Sprachtypologie und Universalienforschung*, 62(3):171–204.

Ines Rehbein, Sören Schalowski, and Heike Wiese. 2014. The KiezDeutsch korpus (KiDKo) release 1.0. In *The 9th International Conference on Language Resources and Evaluation (LREC-14), Reykjavik, Iceland*.

Younes Samih and Wolfgang Maier. 2016. An Arabic-Moroccan Darija code-switched corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, may.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004*, pages 1263–1266.

Han-Ping Shen, Chung-Hsien Wu, Yan-Ting Yang, and Chun-Shan Hsu. 2011. Cecos: A chinese-english code-switching speech database. In *Speech Database and Assessments (Oriental COCOSDA), 2011 International Conference on*, pages 120–123. IEEE.

Thamar Solorio and Yang Liu. 2008. Part-of-speech tagging for English-Spanish code-switched text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060, Honolulu, Hawaii, October. Association for Computational Linguistics.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar, October. Association for Computational Linguistics.

Almeida Jacqueline Toribio and Barbara E Bullock. 2012. *The Cambridge handbook of linguistic code-switching*. Cambridge University Press.

Rosemarie Tracy and Elsa Lattey. 2009. It wasn't easy but irgendwie äh da hat sich's rentiert, net?: a linguistic profile. In Michaela Albl-Mikasa, Sabine Braun, and Sylvia Kalina, editors, *Dimensionen der Zweitsprachenforschung. Dimensions of Second Language Research*. Narr.

David Vilares, Miguel A. Alonso, and Carlos Gomez-Rodriguez. 2016. EN-ES-CS: An English-Spanish code-switching twitter corpus for multilingual sentiment analysis. 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).

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. Pos tagging of English-Hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979, Doha, Qatar, October. Association for Computational Linguistics.

Emre Yilmaz, Maaike Andringa, Sigrid Kingma, Jelske Dijkstra, Frits Van der Kuip, Hans Van de Velde, Frederik Kampstra, Jouke Algra, Henk van den Heuvel, and David van Leeuwen. 2016. A longitudinal bilingual Frisian-Dutch radio broadcast database designed for code-switching research. 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).