A Turkish-German Code-Switching Corpus

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

Abstract

Bilingual communities often alternate between languages both in spoken and written communication. One such community, Germany residents of Turkish origin produce Turkish-German code-switching, by heavily mixing two languages at discourse, sentence, or word level. Code-switching in general, and Turkish-German code-switching in particular, has been studied for a long time from a linguistic perspective. Yet resources to study them from a more computational perspective are limited due to either small size or licence issues. In this work we contribute the solution of this problem with a corpus. We present a Turkish-German code-switching corpus which consists of 1029 tweets, with a majority of intra-sentential switches. We share different type of code-switching we have observed in our collection and describe our processing steps. The first step is data collection and filtering. This is followed by manual tokenisation and normalisation. And finally, we annotate data with word-level language identification information. The resulting corpus is available for research purposes.

Keywords: code-switching, Turkish, German

1. Introduction

Code-switching (CS) is a natural extension of the language used among immigrant communities (Toribio and Bullock, 2012). Bilingual speakers fluently switch between the language of their background culture and the language of the country they live in, by alternating inter-sentence, intra-sentence, or even intra-word. This is especially easy to observe in daily life where one immigrant group constitutes a large group of minority such as the Turkish community in Germany. Example 1 gives such a code-switching instance:[1]

(1) **seitdem ich keine schule hab** cok **faul** oldum ben ya:D
‘**since I have no school** I became very **lazy** :D’

Due to the large number of Turkish immigrants in Germany, Turkish-German CS is studied by several researchers, mostly on the grounds of sociolinguistics and language acquisition. Kallmeyer and Keim (2003) investigate the character of communication between young girls in Mannheim, mostly of Turkish origin and show, among other things, that they employ a mixed form of Turkish and German with peers. Androutsopoulos and Hinnenkamp (2001) and Hinnenkamp (2008) look into chat rooms of immigrants of Greek and Turkish origin to observe the multilingual behaviour.

The SKOBI Corpus collects the recordings of Turkish-German bilingual children (Rehbein et al., 2009a). Rehbein et. al (2009b) and Herkenrath (2012), who primarily study the development of the complex language these children employ, report on CS in their language. Kiezdeutsch corpus (Rehbein et al., 2014) consists of conversations among native German adolescents from a multiethnic urban area. Since the majority of the participants have Turkish background, about 400 Turkish sentences found place in the corpus, 40 of which employ intra-sentential CS.

Although code-switching is more observed in spoken data, written data also exhibits CS, especially in the informal, conversation-like environment of social media. User generated content is as interesting as data collected from field work. However, its large size raises the need of computational approaches to code-switching. Researchers who want to apply such automatic approaches to Turkish-German CS encounter two obstacles regarding the lack of resources: Either the collected corpus does not contain large amounts of CS data, as the research purposes of the mentioned corpus are broader than CS, or the corpus is not available to other researchers due to licence issues.

In this paper, we take an initial step to fill a gap in resources and present the creation of a corpus of Turkish-German code-switching, with a focus on intra-sentence alternations. We chose Twitter as a natural medium of CS with the advantage of large amount of content and easy collection. We manually tokenised and normalised the tweets we collected, and annotated them with language identification tags. To our best knowledge the presented work is the first Turkish-German CS data collected and annotated for computational objectives.

The rest of the paper is as follows: We discuss related work and clarify terminology we employ in this paper in Section 2. We describe the data collection and present our observations on the data in Section 3. Data processing and annotation is explained in Section 4. We give corpus and annotation analysis in Section 5. and conclude in Section 6.

2. Background

Researchers have used different terminology when defining utterences that alter between two languages. Some researchers define alternations on the intra-sentential level as code-mixing while others do not make such a distinction and use the term code-switching to cover both inter-sentential and intra-sentential alternations (Poplack, 1980; Myers-Scotton, 1997). In this paper we follow the latter

[1] German parts are in bold.

definition. Another discussion point is the distinction between code-switching and borrowing a lexical item. Both Poplack (2001) and Myers-Scotton (1997) argue that although the two phenomena quite differ in definition, it is not as easy to distinguish them in use. In our corpus we mark single-word language alternations without a further distinction between code-switches and loanwords. Either way they are very interesting from our point of view as they pose a challenge to computational systems built for monolingual purposes.

Despite the long practice of linguistic research on code-switching, computational studies have been sparse after Joshi's (1982) theoretical framework to parse code-switched sentences. This picture has been changing in the past few years, with research focusing on word-level language identification (Nguyen and Doğruöz, 2013; Das and Gambäck,2014; cf. Solorio et al., 2014), predicting code-switching points(Solorio and Liu, 2008a; Elfardy et al., 2013), and POS-tagging(Solorio and Liu, 2008b; Vyas et al., 2014; Jamatia et al., 2015).

Together with the computational approaches, several studies provide code-switching corpora. One commonly used source of data is social media, mainly forums, Facebook, and Twitter. Nguyen and Doğruöz (2013) create their corpus from Turkish-Dutch posts in an online discussion forum. Barman et al. (2014) collect their Bengali-Hindi-English dataset from Facebook comments. Das and Gambäck (2014) also use Facebook for English-Hindi and English-Penjabi corpora. Vyas et al. (2014) choose Facebook celebrity pages and BBC Hindi as their media and collected user posts to these sites. Jamaita et al. (2015) utilise both Facebook and Twitter in compiling their English-Hindi data. Twitter is also the main source in creating corpora for the Shared Task on Language Identification in Code-Switched Data, in pairs Spanish-English, Nepali-English, Mandarin-English, and Modern Standard Arabic-Egyptian Arabic (Maharjan et al., 2015).

3. Data

We have 1029 tweets in our corpus; in 917 of them Turkish and German are switched intra-sententially and in the remaining 112 inter-sententially. The following sections explain the data collection and our observations on the data.

3.1. Collection

We are interested in a very specific type of tweets and with the features Twitter provides it seems straightforward to fetch such data through the Twitter API. After all, it gives geolocation and language features and users who tweet in Turkish from German-speaking countries would be the most prominent candidates in tweeting also code-switching tweets. There are, however, two obstacles to this idea. First, few German users geo-tag their tweets (Scheffler, 2014). Second, the language feature is automatically assigned by Twitter, and does not always reflect the actual language of a tweet. We tried limiting ourselves to Turkish tweets from Germany at first, but tweets we collected per day was so few that we found it slow for our purposes.

Instead we employed two other strategies in collecting our data. For the first set, we downloaded 10 million tweets in September 2015 that are marked as Turkish by the Twitter API. We also downloaded 1 million German, 1 million English, and 1 million Turkish tweets to create frequency dictionaries out of them. Since language-specific tweets can have many tokens that do not belong to the language, we filtered the Turkish and German dictionaries through morphological analysers (Oflazer, 1994; Schmid et al., 2004) and created *pure* dictionaries that contain valid Turkish and German words respectively. Then we used these five dictionaries to decide if a token of a tweet is Turkish, German, or neither. We looked at frequencies if a word belongs to more than one language. The language assignment gave us the potential set of 8000 code-switched tweets.[2] We then manually went through the tweets and selected the tweets with intra-sentential CS. This resulted in 680 tweets.

In our experience, due to our strict policy, the automatic filtering did give Turkish tweets with German words, but we eliminated tweets when German words are only proper names,[3] or only check-ins (e.g. *@München Hauptbahnhof* '@Munich Main Train Station'). Also we did not take into account newspaper tweets where the same headline is given both in Turkish and in German. There were also some false positives where a word is actually in two languages (e.g. *Kombi* 'combi boiler' in Turkish and 'estate car' in German), but according to the morphological analysers it is only German.

For the second set we used an in-house collection of tweets that are crawled between April 2009-April 2010 and then between April-July 2011, which have Germany as their geolocation. We selected 39 million German and 80.000 Turkish tweets according to their Twitter language feature. For the Turkish tweets we applied the same procedure as the first set, and also made sure they are still existing tweets. After the manual inspection, we had a set of 53 tweets. From the German tweets, we selected potential CS tweets by regular expressions that search for frequent Turkish words, and then manually eliminated the tweets that are fully Turkish. After the availability check for tweets, this gave us a list of 296 tweets.

3.2. Observations

From the context of the tweets in our corpus, it seems young people, often students, mix German and Turkish in their posts. In addition, we observed some recurring patterns in switching between languages. In this section we list examples of patterns by giving the actual tweet and its English translation. German words are marked in bold in both the original and translated versions. The Turkish part of the intra-word CS is marked in italics in the translations.

It is common to couple the infinitive form of a German verb with a Turkish light verb *etmek* 'do' or *yapmak* 'make'. The constructions **verb** *yapmak* or **verb** *etmek* mean 'to verb', that is, the meaning does not change. The advantage is to avoid the inflection on the foreign verb and mark it on the light verb.

[2] The code for the automatic part of the data collection is available at `https://github.com/EggplantElf/creepy`

[3] We keep German proper names if they are inflected with Turkish suffixes.

(2) @username ben seyi **komisch finden** ettim türkce hic yazmamislar hic mi türk tanidiklari yok? asiri sacma
‘@username I **find** this thing **funny**, haven’t they ever written in Turkish, don’t they have any Turkish acquaintance? Beyond ridiculous.’

When the Turkish and German orthography of some words are very similar, the German spelling is used instead of the Turkish one. **energie**-*enerji*, **ethik**-*etik*, **negativ**-*negatif* are such examples.

(3) Bu nevi **ethik** anlayışı kaç kişide, kaç kurumda var bu cografyada? [url]
‘How many people, how many institutions in this region have such type of **ethics** mentality? [url]’

Intra-word CS occurs due to the agglutinative nature of Turkish. German words are declined sometimes by directly concatenating the suffix, sometimes by using an apostrophe between the word and suffix, but more often using a space between the word and suffix.

(4) @username no way senyor ya :D **menschen** beni **verrückt** çok **traurig***im* hacı ya :D aahah iki dilide katlettim eğlencek şey arıyorum :(
‘@username no way mister, **people** make me **crazy**, *I am* very **sad** :D I slew both languages, I look for something for fun :(’

(5) tok karınla ne **diät**’*ler* planlanır.
‘One plans so many **diet***s* with a full stomach’

In Example 6, the locative suffix *de* is normally written attached to the preceding word, but the user has chosen to write it separately from the German word.

(6) **Lehrerzimmer2** *de* **schokolade** dağıtıyorlar acil RT :)
‘They are giving away **chocolate** *in* **classroom 2**, urgent RT :)’

Often times, there are lexicalised expressions such as greetings and best wishes in Turkish, in otherwise German tweets. The opposite is also observed.

(7) @username **nicht schlecht** afiyet olsun :)
‘@username **not bad** enjoy your meal :)’

German vocatives in Turkish text (e.g. **Bruder** ‘brother’, **Schatz** ‘sweetheart’) or Turkish vocatives in German text (e.g. *lan, oğlum* ‘son’) is also popular. Sometimes German vocatives bear the Turkish 1st person possessive marker as it is common in Turkish vocatives.

(8) @username aynen **Bruder** saol ah
‘@username exactly **bro** thanks ah’

There are several examples where the subordinate clause of a sentence is in one language and the main clause is in another language. Same goes with clauses connected with conjunctions.

(9) @username **ich habe keine singstimme** ama cok güzel siir okurum. Ormantink :)
‘@username **I have no singing voice** but I read poems very well. Romantic :)’

It is also frequent to alternate back and forth between languages in inter-sentential CS.

(10) tmm cnm benim.. **sag mir was du haben möchtest, welche farbe?** ben hemen yaparim.. gelirkende getiririm
‘OK my dear. **Tell me what you want to have, which colour?** I will do immediately.. and will bring when I come.’

4. Data Processing and Annotation

The processing and annotation of the data is conducted with a team of three annotators and one researcher. The annotators are Turkish-German bilingual computational linguistics students. At all stages, each tweet is processed by two annotators. The annotators then compared their results with the second annotator to catch and correct their own mistakes. Remaining conflicts are resolved by the researcher. Further changes (e.g. overlooked tokenisation or normalisation by both annotators) are done on the resolved version.

4.1. Tokenisation and Normalisation

The user generated nature of Twitter exhibits many tokenisation mistakes. One common Turkish mistake is the use of *de, ki, mi* as suffixes or clitics. Suffixes are attached to the preceding word, whereas clitics are written separately, and users often mix between two. Another pattern widely seen in Turkish social media is omitting the use of apostrophes when adding suffixes to proper names. The rest of frequent non-standard orthography is common also in other languages and could be grouped as ignoring capitalisation, repeated characters especially in interjections, omitting vowels within words.

Guidelines we used are from Turkish Language Association (TDK).[4] In addition we normalised interjections, restored capital letters where necessary, inserted apostrophes before suffixes attached to proper names, normalised interjections and unvocalised tokens, as well as correcting all sorts of typos. We also replaced usernames with *@username* and URLs with *[url]*. A tweet can consist of multiple lines. For the sake of simplicity in processing we replaced newlines with <*NL*>.

At this point, as we manually went through all the tokens, we also marked the switching point of mixed tokens. We used ‘§’, an otherwise unused character in the corpus, to denote the boundary. For instance *Terminde* ‘at the appointment’ consists of the German word *Termin* ‘appointment’ and the Turkish locative case suffix *de*. We represent this word in our corpus as *Termin§de*. If a German proper name takes a Turkish suffix, then the boundary marker is placed before the apostrophe as in *Türkei§’da* ‘in Turkey’.

Since it is not possible to calculate inter-annotator agreement on tokenisation we cannot give any quantitative measures. But comparison observations showed that, majority

[4] `http://www.tdk.gov.tr/index.php?option=com_content&view=category&id=50`

of the disagreement came from not identifying all *de, ki, mi* mistakes correctly. This is understandable, as the annotators never had a formal education of written Turkish before, and learned the rules from the tokenisation guidelines.

4.2. Language Identification

The second step is to annotate the words in the corpus for language identification. We mainly follow the annotation scheme from the 2014 Shared Task on Language Identification in Code-Switched Data (Solorio et al., 2014; Maharjan et al., 2015) with one extra label. The original tag set has two language labels, any third language is considered `OTHER`. We add `LANG3` to identify them.

We go through the labels by giving examples from Section 3.2. where possible:

- **`TR:`** Turkish, e.g., *ben* 'I' from (2).
- **`DE:`** German, e,g., **komisch** 'funny' from (2).
- **`LANG3:`** Third language, e.g., 'no way' from (4).
- **`MIXED:`** Intra-word CS, e.g., **traurig***im* 'I am sad' from (4).
- **`NE:`** Named entity, e.g., Bern, Ankara, DW (German international broadcaster), Kanal D (Turkish TV channel).
- **`AMBIGuous:`** Words that exist in both languages and cannot be disambiguated by the given context.
- **`OTHER:`** Punctuation, numbers, emoticons, symbols, and any token that cannot be classified with previous labels, e.g., 'RT' from (6).

As our annotation guidelines, we followed the appendix given in (Maharjan et al., 2015). We used Hovy et al.'s (2014) annotation tool for language identification and modified it to suit our needs. Similar to the the original tool, we pre-tagged certain token classes: emoticons, punctuation, Twitter-specific tokens, most frequent Turkish and German words. The annotators could decide to keep the tags or modify them.

4.3. Extension to the Tag Set

Discussion meetings with the annotators showed that most of confusion arises from annotating named entities. This finding is supported with the confusion matrix we have derived during the inter-annotator agreement calculations.

We calculated inter-annotator agreement after the first pass on language identification annotation, before one annotator compared her tags to the second annotator. We measured an agreement of 93.95% and Cohen's kappa(Cohen, 1960) is 0.91. The agreement score is quite high, showing an overall success in annotation. Tokens annotated as `DE`, `TR`, and `OTHER` by both annotators are very high and the number of disagreements are between 50-60 for these tag pairs (less than 1% of all tokens). It is more common to see a token that is assigned a `NE` tag by one annotator and a language tag (`DE`, `TR`, or `LANG3`) by another annotator. This is usually observed when the word itself is not a proper name but still part of a named entity, e.g. association, bridge, museum. This problem could be solved by more annotator training or more descriptive annotation guidelines but there is another aspect of assigning a `NE` tag to named entities. Some named entities are language-specific, such as 'Munich' in English vs. 'München' in German vs. 'Münih' in Turkish. Assigning only a `NE` tag loses this distinction. Moreover a named entity itself might be an instance of code-switching. An example from our corpus is *Aufbruch Neukölln Derneği* 'Emerging Neukölln Society'. The first two words are in German and the last one is in Turkish. It is not possible to recognise such mixed named entities with dictionaries or gazetteers for instance.

In order to mark the language information on named entities, we replaced the `NE` tag with `NE.X` tags, where `X` could be any other member of the tag set. This way we accommodate a fine-grained tag set (NE.fine) without introducing another annotation layer. According to this new extension, the representation of 'Emerging Neukölln Society' is given in Example 11.

(11) **Aufbruch Neukölln** Derneği
NE.DE NE.DE NE.TR

When a named entity is represented the same among languages, we decide the NE.fine tag according to context, and if it not possible we assign `NE.AMBIG`. Example 12 shows two different annotations of 'Stuttgart' in two phrases taken from the same tweet.

(12) @username ben Stuttgart'tayım . (...)
OTHER TR NE.TR OTHER (...)
Ökumenisches Zentrum - **Uni**
NE.DE NE.DE OTHER NE.DE
Stuttgart
NE.DE
'@username I am in Stuttgart. (...) Ecumenical Centre - University of Stuttgart'

Due to time limitations, each NE is annotated by NE.fine tags only once by an annotator, later the researcher has gone through the annotations for quality and consistency check.

5. Corpus and Annotation Analysis

Our corpus has 1029 tweets and 16992 tokens with an average of 16.51 tokens per tweet after tokenisation and normalisation. These processing steps are done manually, but we also created the edit transcripts based on Levenshtein distance to simulate the transformation. It needs 4917 substitutions, 4079 insertions, and 800 deletions to go from the original tweets to the edited versions.

Table 5. shows the breakdown of tokens according to language identification tags. The first column is the tokens labelled with the original tag set. The second column gives only the distribution of fine-grained (NE.fine) tags of named entities. The last column adds each NE.fine tag to the respective original tag to have a comparative insight on the corpus when named entity information is not taken into account. Turkish tokens constitute half of the tweets, when the Turkish named entities are added to this amount, it goes up to 52.51%. The `OTHER` tag is the second biggest set in overall, but they have a few occurances in named entities, which correspond to numbers and the & symbol in some brand names. German tokens are less than half of Turkish tokens in general, but note that in the named entities they are on par. Mixed tokens are a small percentage in overall but has a higher weight in the name entities. Both German and mixed tokens show that named entities are commonly used in code-switching. LANG3 is mostly English and we have observed a few tokens of Dutch, Arabic, and Italian.

Tag	Original	NE.fine	NE.fine dist.
TR	8556 (50.35%)	367	8923 (52.51 %)
OTHER	3843 (22.62 %)	24	3867 (22.76) %
DE	3450 (20.30 %)	321	3771 (22.19 %)
NE	913 (5.37 %)	-	-
MIXED	109 (0.64 %)	88	197 (1.16 %)
LANG3	92 (0.54 %)	100	192 (1.13 %)
AMBIG	29 (0.17 %)	13	42 (0.25 %)

Table 1: Breakdown of language identification tags. The columns give the original tag set, the tag distribution on named entities only, and the addition of NE.fine distribution to the respective tags.

When we look at the distribution of tags at the tweet level, 790 of the cases tokens labelled with TR are more than tokens labelled with DE. For the remaining 239 tweets the number of German tokens are more than or equal to the number Turkish tokens. There are no tweets without a Turkish token. However, there are 47 tweets with no German tokens; in 44 of these cases there are Mixed tokens and for 3 cases there are Ambigious tokens.

6. Conclusion

In this paper, we present a new collection of 1029 Turkish-German tweets that serve as a code-switching corpus. The corpus is manually tokenised and normalised, and annotated with word-level language identification information.

We compiled our data with a combination of automatic and manual filtering that is applied to a large amount of Twitter harvest. Daily Twitter traffic is high, and its API is easy-to-use, which can lead heaps of data quickly. Yet, in cases where specific types of tweets are pursued, Twitter could be costly timewise. We partially solved this problem by using an existing in-house tweet collection.

We also had the chance to use existing guidelines both for tokenisation and normalisation, and for language identification. Since the guidelines we used for tokenisation are not designed for social media, we added new instructions where necessary. We followed the guidelines for 2014 Shared Task on Language Identification in Code-Switched Data (Solorio et al., 2014; Maharjan et al., 2015) quite closely, with two alternations. First, we introduced a separate tag for words in a third language. Second, we annoted also named entities with language information instead of marking them only as NE.

Looking at examples from the data showed that common code-switching types are German verb + Turkish light verb constructions, writing in German orthography instead of Turkish when words are very similar, using German vocatives, or lexicalised expressions in Turkish text or vice versa, clause level alternations, and inflecting German words with Turkish suffixes.

The final corpus and our internal annotation guidelines are available to researchers at http://www.ims.uni-stuttgart.de/institut/mitarbeiter/ozlem/cetinogluLREC2016.html.[5]

7. Acknowledgements

We thank Xiang Yu for tweet collection and automatic filtering and André Blessing for providing the in-house German tweets. We thank Leyla Adıgüzel, Sevde Ceylan, and Hasret el Sanhoury for the annotation processes. This work was funded by the Deutsche Forschungsgemeinschaft (DFG) via SFB 732, project D2.

8. Bibliographical References

Androutsopoulos, J. and Hinnenkamp, V. (2001). Code-switching in der bilingualen chat-kommunikation: ein explorativer blick auf #hellas und #turks. In Michael Beisswenger, editor, *Chat-Kommunikation. Sprache, Interaktion, Sozialität & Identität in synchroner computer-vermittelter Kommunikation Perspektiven auf ein interdisziplinäres Forschungsfeld*. Ibidem Verlag.

Barman, U., Das, A., Wagner, J., and Foster, J. (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*, pages 13–23, Doha, Qatar, October. Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Das, A. and Gambäck, B. (2014). Identifying languages at the word level in code-mixed indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing, Goa, India*, pages 169–178.

Elfardy, H., Al-Badrashiny, M., and Diab, M. (2013). Code switch point detection in Arabic. In *Natural Language Processing and Information Systems*, pages 412–416. Springer.

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

Hinnenkamp, V. (2008). Deutsch, doyc or doitsch? chatters as languagers–the case of a German–Turkish chat room. *International Journal of Multilingualism*, 5(3):253–275.

Hovy, D., Plank, B., and Søgaard, A. (2014). Experiments with crowdsourced re-annotation of a pos tagging data set. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 377–382, Baltimore, Maryland, June. Association for Computational Linguistics.

Jamatia, A., Gambäck, B., and Das, A. (2015). Part-of-speech tagging for code-mixed english-hindi twitter and

[5] Following the restrictions of Twitter's Terms of Service, we distribute the tweet IDs instead of actual tweets. We also distribute the edit transcript that converts the original tweets to edited versions. We provide scripts to generate the edited tweets from their original versions and to merge them with annotations.

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.

Joshi, A. K. (1982). Processing of sentences with intra-sentential code-switching. In *Proceedings of the 9th conference on Computational linguistics-Volume 1*, pages 145–150. Academia Praha.

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

Maharjan, S., Blair, E., Bethard, S., and Solorio, T. (2015). Developing language-tagged corpora for code-switching tweets. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 72–84, Denver, Colorado, USA, June. Association for Computational Linguistics.

Myers-Scotton, C. (1997). *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Nguyen, D. and Doğruöz, A. S. (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.

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

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en Espanol: toward a typology of code-switching. *Linguistics*, 18(7-8):581–618.

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

Rehbein, J., Herkenrath, A., and Karakoc, B. (2009a). Corpus Rehbein-SKOBI (sprachliche konnektivität bei bilingual türkisch-deutsch aufwachsenden kindern und jugendlichen).

Rehbein, J., Herkenrath, A., and Karakoç, B. (2009b). 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.

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

Scheffler, T. (2014). A German twitter snapshot. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2284–2289.

Schmid, H., Fitschen, A., and Heid, U. (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.

Solorio, T. and Liu, Y. (2008a). Learning to predict code-switching points. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 973–981, Stroudsburg, PA, USA. Association for Computational Linguistics.

Solorio, T. and Liu, Y. (2008b). 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.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., and Fung, P. (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.

Toribio, A. J. and Bullock, B. E. (2012). *The Cambridge handbook of linguistic code-switching*. Cambridge University Press.

Vyas, Y., Gella, S., Sharma, J., Bali, K., and Choudhury, M. (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.