Part of Speech Annotation of a Turkish-German Code-Switching Corpus

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

Çağrı Çöltekin
Department of Linguistics
University of Tübingen
Germany
ccoltekin@sfs.uni-tuebingen.de

Abstract

In this paper we describe our efforts on POS annotation of a code-switching corpus created from Turkish-German tweets. We use Universal Dependencies (UD) POS tags as our tag set. While the German parts of the corpus employ UD specifications, for the Turkish parts we propose annotation guidelines that adopt UD's language-general rules when it is applicable and adapt its principles to Turkish-specific phenomena when it is not. The resulting corpus has POS annotation of 1029 tweets, which is aligned with existing language identification annotation.

1 Introduction

Multilingual speakers cover a higher percentage of the world population than monolingual speakers (Tucker, 1999). Acting multilingual, that is, mixing languages is commonly observed among these multilingual speakers (Auer and Wei, 2007). The definition, types, and use of language mixing have long been studied by researchers, especially from a sociolinguistic perspective (Gumperz, 1964; Sankoff, 1968; Lipski, 1978). Some linguists make distinctions in the terminology according to the level of the language mixing, e.g. use *code-mixing* for sentence-internal alternations, some others use either *code-mixing* or *code-switching* for all types of mixing (Poplack, 1980; Myers-Scotton, 1997). In this paper we use code-switching (CS) as an umbrella term.

Unlike linguistic studies, computational research on code-switching has recently accelerated, although the first theoretical framework to parse code-switched sentences has been proposed by Joshi back in the 80s (Joshi, 1982). Several studies has emerged 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).

Computational approaches often need annotated data. The number of CS corpora annotated with language identification information has also increased proportional to the interest in the field (Nguyen and Doğruöz, 2013; Barman et al., 2014; Das and Gambäck, 2014; Maharjan et al., 2015).

Part of speech (POS) annotation of CS data, on the other hand, is not very common yet. To our knowledge, there are only three code-switching corpora with POS annotation:[1] one on Spanish-English (Solorio and Liu, 2008b) and two on Hindi-English (Vyas et al., 2014; Jamatia et al., 2015). These are valuable resources as part of speech tags can provide more insight on the nature of code-switching and pave the way for syntactic annotation.

Here in this work, we present a fourth CS corpus annotated with POS information. The corpus contains 1029 Turkish-German tweets, already annotated with language information (Çetinoğlu, 2016). We add the POS tag layer following Universal Dependencies (UD) (Nivre et al., 2016). German is one of the languages UD already covers. Turkish on the other hand is under development. Therefore, our work also contributes to the discussions on POS tagging and segmentation of Turkish in the UD framework.

The rest of the paper is as follows: We discuss previous annotation efforts in CS and POS annotation in social media in Section 2. The data is described in Section 3 and annotation decisions are explained in Section 4. Processing steps are given

[1] There are some POS-annotated corpora that contain CS instances although the intention of collection is different. For instance the KiezDeutsch corpus (Rehbein et al., 2014) has a small number of utterances with Turkish-German CS. Old German Reference Corpus (Dipper et al., 2004) has examples of mixing Old High German and Latin.

Proceedings of LAW X – The 10th Linguistic Annotation Workshop, pages 120–130,
Berlin, Germany, August 11, 2016. ©2016 Association for Computational Linguistics

in Section 5. We analyse the data and processing in Section 6 and conclude in Section 7.

2 Related Work

Corpora created for studying code-switching computationally mostly focus on data annotated with language information. Nguyen and Doğruöz (2013) collect Turkish-Dutch posts from an online discussion forum and annotate words as Turkish or Dutch. A small amount of English words are also annotated as Dutch. Punctuation, numbers, emoticons, links, chat language, meta forum tags, proper names are ignored during annotation. Barman et al. (2014) create a CS corpus of Bengali-Hindi-English from Facebook comments. They define English, Bengali, Hindi, Mixed tags, and annotate named entities, acronyms, and universal expressions such as symbols, numbers, emoticons as separate tags. The Shared Task on Language Identification in Code-Switched Data also uses social media, namely Twitter, as their main source in collecting code-switching data. They present corpora in pairs Spanish-English, Nepali-English, Mandarin-English, and Modern Standard Arabic-Egyptian Arabic (Maharjan et al., 2015).

POS tagged data sets are fewer as compared to ones annotated with language information. Solorio and Liu (2008b) are the first to annotate POS tags on code-switched data. They recorded conversations between bilingual speakers of Spanish and English. Then they transcribed this data and manually annotated with POS tags. They used a fine-grained tagset which is a combination of English and Spanish TreeTagger (Schmid, 1994) tags (a version of Penn Treebank tag set for English and 75 tags for Spanish). Out of the 922 sentences they collected, 576 are monolingual English. There are 239 switches throughout the conversations, 129 of them are intra-sentential.

Following studies on annotating code-switching data with POS tags come years later. Vyas et al. (2014) chose Facebook celebrity pages and BBC Hindi as their media and collected user posts mixed in English and Hindi. Their annotation is in multiple layers. First the posts were splitted into fragments so that they would have a unique matrix language English or Hindi. Each word in a matrix is identified as English, Hindi, or Other. The POS layer employs 12 Universal POS tags (Petrov et al., 2011) and three additional tags for named entities (people, location, organisation). They have a corpus of 381 posts which corresponds to 4135 words. 17.2% of these posts contains intersentential or intrasentential code-switching.

Jamatia et al. (2015) utilised both Facebook and Twitter in compiling their English-Hindi data. They divided posts and tweets into utterances and automatically tokenised them. Manual POS annotation uses a fine-grained tag set which could be mapped to a coarse-grained one. The fine-grained set combines tags developed for Indian languages with Twitter-specific tags from Gimpel et al. (2011). The coarse-grained version retains the Twitter-specific tags and maps the rest to Universal POS tags. The resulting corpus consists of 2583 utterances, with 68.2% being monolingual.

Efforts on POS annotation of social media started with using the Penn TreeBank tag set (Marcus et al., 1993) for English (Foster et al., 2011; Petrov and McDonald, 2012). Ritter et al. (2011) extended PTB tagset with Twitter-specific tags for retweets, usernames, hashtags, and URLs. Gimpel et al. (2011) designed a completely new set tailored to Twitter. For German, Neunerdt et al. (2013) use the standard STTS POS tag set (Schiller et al., 1995) to annotate web comments. Rehbein (2013) adopts the same tag set and introduces new tags for usernames, URLs, hashtags, and emoticons for POS tagging German tweets. Similarly for Turkish, Pamay et al. (2015) use the standard POS tag set of Oflazer (1994) and add tags for abbreviations, emoticons, mentions, hashtags, and URLs to cover the non-canonical content of a web treebank.

3 Data

We use the data that Çetinoğlu (2016) has collected on code-switching Turkish-German tweets. It consists of 1029 tweets, each having at least one code-switching point. Tweets are automatically collected and manually filtered. Before adding language identification annotation tokenisation and normalisation is applied based on Turkish and German orthography rules.

The tag set is based on the 2014 Shared Task on Language Identification in Code-Switched Data (Solorio et al., 2014; Maharjan et al., 2015): TR (Turkish), DE (German), LANG3 (third language), MIXED (intra-word CS), NE (named entity), AMBIG (words belong to both languages and cannot be disambiguated with the given context), OTHER (punctuation, numbers, URLs, emoticons, symbols, any other token that do not belong to previous classes). The Shared Task labels the tokens that belong to a third language as OTHER, Çetinoğlu (2016) introduces the LANG3 tag for them. Additionally, named entities are tagged both as NE as in the Shared Task, and with their language label TR, DE, or LANG3. MIXED tokens are also marked with the code-switching boundary, represented with the symbol '§'.

There are 16992 tokens in total, that corresponds to 16.51 tokens per tweet. Half of the tokens are Turkish, it is followed by OTHER and German, both being around 20%. In 790 tweets, there are more tokens labelled as TR than DE. Details of the data collection, correction, and annotation processes are explained in Çetinoğlu (2016).

4 Annotation Guidelines

The annotation process follows the Universal Dependencies (Nivre et al., 2016) conventions as much as possible.[2] We only use the POS tag labels from the UD inventory, and follow the general principles of UD as well as the available language-specific documentation for each language in the corpus. Although we do not explicity annotate in the syntactic level, we have to take into account UD syntax representation, especially for segmenting Turkish words.

Besides the recent popularity of the UD-based annotations, the major advantage of UD in our work is that the UD guidelines are intended to be as language-general as possible. For a multilingual corpus, such as ours, the importance of uniform annotations within the corpus cannot be overstated. The downsides, on the other hand, are potential confusion due to already established annotation conventions (such as STTS (Schiller et al., 1995) for German), and the fact that UD is an ongoing project, and parts of the formalism is still in development.

In this section we describe the annotation guidelines we follow briefly, focusing more on the aspects that differ from UD or the common conventions used in relevant monolingual corpora.

4.1 Segmentation

Following Universal Dependencies guidelines, we mark POS tags on *syntactic* words,[3] which results in segmenting some of the surface tokens in both German and Turkish. For German, the only case that require segmentation is the contraction of prepositions and definite articles. For example, the word *zur* 'to the' is tokenised into its parts as *zu* and *der*. The segmentation of Turkish syntactic words is more involved, and at present, the UD guidelines for Turkish tokenization are still a moving target. We describe the approach we employed for segmentation of Turkish below.

Turkish is a morphologically complex language. In addition to a large set of inflectional morphemes that can attach to verbal or nominal stems, some productive (derivational) morphemes may change the POS tag of an already inflected word. In Turkish NLP literature, this phenomenon is addressed with sub-word units that are often called *inflectional groups* (IGs) (Oflazer, 1999), which correspond to one or more morphemes grouped by derivational boundaries. In this work, we also follow the same convention, however, similar to Çöltekin (2016), we follow a more conservative approach to segmentation in comparison to most earlier work. Instead of segmenting a word into IGs after each derivation, we segment only before the morphemes that introduce a new syntactic word, such that parts of the word may carry conflicting morphological features, or participate in separate syntactic relations. In other words, we segment words to avoid potential ambiguous or conflicting morphosyntactic annotations.

An example of this is presented in (1) below,[4] which also coincides with an instance of word-internal code switching. As introduced earlier, the symbol '§' indicates the code switching boundary within a word. We mark inflectional group boundaries with the symbol '•' in the examples.

(1) sabah
morning.NOUN.Sg
Internetseite§-de•ki-ler-i
website.NOUN.Sg-Loc•ki.NOUN-Pl.Acc
ausdrucken ed-eceğ-im
print.VERB.Inf do.VERB-Fut-1Sg
'I will print the ones from the website in the morning'

The singular German noun *Internetseite* 'website' is inflected with the Turkish locative case marker *de*. This is the code-switching point. The

[2] More specifically we follow UD version 1.2.

[3] Segmentation is not in the morpheme level, yet words are not necessarily phonological or orthographic. http://universaldependencies.org/u/overview/tokenization.html

[4] Notation of examples and gloss descriptions are given in Appendix A.

rest of the word takes Turkish inflectional and derivational suffixes. The part *Internetseitede* 'on the website' functions as an adjective when it gets the derivational suffix *-ki* (e.g. *Internetseitedeki foto* 'the photo on the website'). With a zero derivation, the derived adjectival behaves as a noun, thus can bear a plural suffix and a case marker. In it is final form, the word *Internetseitedekileri* 'the ones on the website' refers to a set of objects (e.g., documents or pictures) on a website. Without segmentation, we cannot represent the fact that there is only one website but multiple items within the website. Similarly, the direct object of the predicate is the items on the website, not the website (which could have been a direct object of another predicate). As a result, annotations that allow correct interpretations of words like *Internetseitedekileri* above require further segmentation.

Besides the relativiser *-ki* discussed above, we mark the following suffixes which may introduce similar ambiguous or conflicting morphosyntactic annotations.[5]

- *-lH* deriving nouns and adjectives from a noun (N) with the meaning of 'with N' (*dondurma**lı*** '(the one) with ice cream', deriving adjectives and nouns from location names with the meaning 'from N' (*Berlin**li*** '(the person) from Berlin'

- *-sHz* deriving nouns and adjectives from a noun with the meaning of 'without N' (*eğitim**siz*** '(the person) **without** education')

- *-lHk* deriving nouns and adjectives from a noun with the meaning of 'fit/suitable for N' (*sen**lik*** '**fit for** you')

- *-CH* deriving nouns and adjectives from a noun with the meaning of 'preferring N' (*bira**cı*** '(the one) who prefers beer'), as well as mostly lexicalized use of deriving nouns referring to occupations (*fizik**çi*** 'physic**ist**')

- *-lAş* deriving verbs from nouns with the meaning of 'become N' (*özgür**leş**mek* 'to **become** free')

- Copular suffixes (*sizden**di*** '(he/she) **was** one of you')

Similar to *-ki*, the first four suffixes form either adjectives or nouns from nouns. In their adjectival use, segmentation is not strictly necessary as the adjectives in Turkish do not inflect. We segmented productive uses of these suffixes regardless of whether they derive nouns or adjectives for the sake of easier and more accurate annotation.

The last two examples in the above list form predicates form nouns and adjectives. When these suffixes are attached to simple nouns or adjectives, one may avoid segmentation. However, the copular suffixes may also attach to subordinate verbs, in which case, the same word carries two predicates with potentially conflicting sets of inflections and syntactic relations outside the word. For example, if we do not segment the copular part of *gördüğüyüz* in (2) below, we cannot identify the facts that the verb *gör* 'see' is inflected for past tense, while the copula is in present tense. Furthermore, the subject of the copula is *o* 'he/she', while the subject of the verb *gör* is *biz* 'we'.

(2) Biz o-nun
We.PRON he/she.PRON-Gen
rüya-sı-nda
dream.NOUN-P3S-Loc
gör-düğü•yüz
see.VERB-Past-3Sg•VERB-Cop-1Pl
'We are the ones that he/she saw in his/her dream'

We segment words before productive uses of all of the suffixes listed in this section. However, we do not segment words if they are lexicalised. For example the suffix *-siz* 'without' is segmented in ***arabasız*** *gidemeyiz* 'we cannot go (there) **without a car**', but not in ***evsiz**ler için yardım* 'help for the **homeless**'.

To decide if a word is lexicalised, we test if the parts of the segmented version can have syntactic dependencies. For instance, *futbolcu* 'footballer' is lexicalised although it is derived from *futbol* 'football' with the agentive suffix *-CH*. In the expression *Amerikan futbolcu*, *Amerikan* 'American' modifies the footballer. An expression where American modifies football requires a third word: *Amerikan futbolu oyuncusu* 'American football player'. In contrast, unless we introduce a new IG with the suffix *-CH*, *eski kitap•çı* have ambiguous interpretations 'old [book shop]' and '[old book] shop/seller'. In other words, parts of the word referring to the 'book' and the 'book shop'

[5] Capital letters in suffixes denote allomorphs. A = {a,e}, H = {ı,i,u,ü}, C = {c, ç}.

can participate in separate syntactic relations.

Another difference from the use of IGs in earlier Turkish NLP literature is that we do not admit 'zero derivations'. All tokens correspond to non-empty surface strings. This results in an inconsistency in the representation of copular suffixes, since a nominal/adjectival predicate in present tense with the third person singular subject does not have a corresponding surface suffix. As a result, the predicate in *Ben hasta-yım* 'I am sick' is segmented, while the predicate *o hasta* 'he is sick' is not segmented. This case poses no problem for our POS annotation purposes, although it would lead inconsistencies in syntactic representation.

4.2 POS Tagging

For both languages, we follow the Universal Dependencies POS tag scheme as closely as possible. UD defines a coarse set of 17 tags listed in Table 1. As in segmentation, the German POS tagging scheme is better defined and more standardised. Despite some existing work, Turkish POS tagging standards for UD is under development.[6] As a result, we focus more on some aspects of Turkish POS tagging in our work. Detailed POS tagging guidelines are included in the distribution of the corpus.

Special word and symbol sequences, such as mentions, hashtags and URLs, are also tagged using the UD POS tag set. We tag mentions (always coded as `@username`) as `PROPN`. The hashtags are tagged as usual when they are a single word with a clear POS tag. For example, `#Berlin` is tagged as `PROPN`, and `#happy` is tagged as `ADJ`. If the hashtag is a multi-word string that cannot be treated as a single word, e.g., `#GiveVoiceToCizre`, it is tagged as `X`. We keep multi-word hashtags intact as we prefer to retain their hashtag property.

Unintelligible alphanumeric sequences and words from other languages whose POS tag could not be determined by the annotators are also tagged as `X`. URLs, emoticons and non-alphanumeric tokens are tagged as `SYM` as per UD specification. We also use the tag `SYM` for the Twitter tags `RE`, `RT` and, the new line representation `<NL>`.

Tag	explanation
`ADJ`	adjective
`ADP`	adposition
`ADV`	adverb
`AUX`	auxiliary verb
`CONJ`	coordinating conjunction
`DET`	determiner
`INTJ`	interjection
`NOUN`	noun
`NUM`	numeral
`PART`	particle
`PRON`	pronoun
`PROPN`	proper noun
`PUNCT`	punctuation
`SCONJ`	subordinating conjunction
`SYM`	symbol
`VERB`	verb
`X`	other

Table 1: Universal dependencies tag set.

All forms of verbs, including verbs that are derived into other categories by subordinating suffixes are tagged as `VERB`. This is in line with the UD guidelines, but unlike most Turkish NLP work where subordinate word structures are typically segmented into multiple IGs, and the last IG (the head) is marked as `NOUN`, `ADJ` or `ADV` depending on whether the verbal form is a *verbal noun*, *participle*, or *converb* respectively.

Auxiliary verbs are tagged as `AUX`, and copulars as `VERB` for both Turkish and German. Similar to German verb *sein* 'to be', the Turkish copula *ol* 'to be/become' can act both as an auxiliary (`AUX`) or as a copula (`VERB`). Examples (3) and (4) show its verb and auxiliary uses respectively from the corpus we annotated.

(3) **Frau** **Geiger**§'i
Ms.`NOUN`.Sg Geiger.`PROPN`.Sg.Acc
gör-dü-m çok mutlu
see.`VERB`-Past-1Sg very.`ADV` happy.`ADJ`
ol-du-m
become.`VERB`-Past-1Sg
'I saw Ms Geiger I became very happy.'

(4) Osmanlı
Ottoman.`PROPN`.Sg
hayal-i
daydream.`NOUN`.Sg-P3S
kur-an-lar duvar-a
fancy.`VERB`.Part-3Pl wall.`NOUN`.Sg-Dat

[6] The UD version 1.3, which is released while the present paper was under review, contains a Turkish treebank. However, the treebank is still in development state, and the documentation is mainly based on Çöltekin (2015), which is not (yet) fully complient with the UD.

tosla-mış	ol-acak
bump.VERB-Evid-Past	be.AUX-Fut.3Sg

'The ones who daydream of Ottomans will have bumped the wall.'

Substantivised adjectives are marked as ADJ. In Turkish it is common to use an adjective as noun with the meaning of 'the object or person with the property described by the adjective'. We mark adjectives as ADJ regardless of their use. This contrasts with most Turkish NLP work to date, since these words are typically analyzed as two separate IGs one of which is introduced by a zero-derivation. In both languages, we also use the tag ADJ for adjectives that are used as predicates.

Multi-word named entities are annotated as normal linguistic units. That is, the words that form a multi-word named entity are not marked as PROPN but as the POS tags they would normally be assigned to. For example in (5) the German word *Aufbruch* and the Turkish word *Derneği* are marked as NOUN even though they are part of a multi-word named entity. The original annotations (Çetinoğlu, 2016) mark the named entities and language IDs as shown in the third row of (5).

(5)

Aufbruch		
Emergence.NOUN.Sg		
NE.DE		
Neukölln	Derneğ-i	
Neukölln.PROPN.Sg	Society.NOUN-P3S	
NE.DE	NE.TR	

'Emerging Neukölln Society'

Non-root inflectional groups in Turkish that are split off from the root part during the segmentation step are assigned POS tags that reflect their function. For example, the IG introduced by the suffix *-siz* in *eğitim-siz insan* 'uneducated person' is tagged as ADJ, while in *eğitim-sizler çoğunlukta* 'uneducated (people) are in majority' it is tagged as NOUN.

Particles of German separable verbs are, following the UD principle, tagged as ADP. This is in contrast with the most common tagging scheme, STTS, used in German NLP so far.

5 Processing

The team for segmentation and POS tagging consists of four annotators and two researchers. All annotators are Turkish-German bilingual undergraduate students. Three of them study computational linguistics, and one studies linguistics.

5.1 Segmentation

Before the task, the annotators were not familiar with the idea of segmenting Turkish words into sublexical units. Thus, the training included the concept of inflectional groups and the current take on segmentation through recent work (Nivre et al., 2016; Çöltekin, 2016). For the actual task, they have given segmentation guidelines. They are also told to oversegment rather than undersegment in case of doubt. Each tweet is segmented by two annotators, and then merged and corrected if necessary, by the researchers. Lexicalised derivations were the source of main conflicts or sometimes non-conflicting oversegmentation. This is expectable, as lexicalisation decisions are rather a continuum. The German side of the segmentation was straightforward and on few cases; annotators easily accomplished this part.

5.2 Restoring Language Identification

When the German and Turkish segmentation has altered, language identification assigned to each token should be altered too. We restored language information in a semi-automatic way. There are three possible scenarios of segmentation. First, when a token identified as German is segmented, all segments are German. Second, similarly, a segmented Turkish token has Turkish segments.

The third scenario is more complex. How the segments of a MIXED token are labelled depends on segmentation boundaries. In our corpus the mixed words to segment are all examples of German-Turkish code-switching (with a single English-Turkish code-switching example). If the segmentation boundary is after the code-switching boundary as in the earlier ***Internetseite****§de-kileri* 'the ones on the website' (1), repeated as (6) below as it is coded in the corpus, the first segment remains MIXED and the second segment is tagged as Turkish. If the segmentation boundary is also the code-switching boundary, then each part is annotated using the corresponding language tags, as in ***kreativ****§miş* 'she/he was creative' demonstrated in (7) below. The fact that these are examples of word-internal code-switching can still be recovered based on the symbols we use for marking code-switching boundaries (§) and non-root IGs (-).

(6)	Internetseite§de	MIXED	NOUN
	-kileri	TR	NOUN
(7)	kreativ§	DE	ADJ
	-miş	TR	VERB

We treated all scenarios automatically, and double-checked the third scenario manually.

5.3 POS Tagging

We started annotator training with existing guidelines and treebank demos from Universal Dependencies.[7] We employed two different training sets for POS tagging. As the first set, we gave annotators 20 tweets separate from the data set and ask them to annotate 10 of them to have double annotation for each. We used these annotations to discuss confusing points. As the second set we gave each annotator up to 15 phrases that are potentially hard to annotate, and ask them to label and add a source, e.g. one of the UD links, to make sure they are aware of multiple sources. Some of these phrases are later used as examples in annotation guidelines.

All tweets are annotated twice. Each annotator is assigned half of the corpus, and each half is annotated by two annotators. The inter-annotator agreement is calculated separately for each half, and then the researchers went through those tweets to resolve conflicts, correct mistakes, and ensure consistency.

6 Analysis

Our annotations are based on the twitter corpus of Çetinoğlu (2016). Originally, the corpus contains 1029 tweets, and 16922 tokens (See Section 3 for more details). After word segmentation, the number of tokens increase to 17274. All tokens are annotated with a POS tag from the Universal Dependencies POS tag inventory, as explained in Section 5. In this section, we provide statistics about the resulting corpora and present some preliminary analyses.

Majority of the segmented tokens are Turkish. In total, 226 Turkish words were segmented. Except three tokens that were tokenised as three IGs, all multi-IG words consist of two IGs. The resulting ratio of IGs per surface word is 1.02 (cf. 1.20 in METU-Sabancı Treebank (Oflazer et al., 2003)). Besides completely Turkish words, 18 mixed words are segmented into two tokens. 17 of these words are German stems with Turkish suffixes, and one is an English word with a Turkish suffix. On the German side, 31 contracted preposition+article combinations were segmented.

The overall inter annotator agreements (IAA) as measured by Cohen's kappa (Cohen, 1960) between two teams are 78.78 and 77.77 for the first and the second team respectively. The IAA per language differ. For Turkish, the agreement scores are lower, averaging 70.39 for both teams. The low score is partially due to the difficulty of the task in Turkish, which is also accented by the fact that our annotators have not received formal education in Turkish, but in German. However, the overall low score also has to do with the fact that non-linguistic tokens (e.g., punctuation, special Twitter symbols) are not included in this calculation. The common disagreements (that are resolved during correction phase) that stand out are, AUX–VERB, ADJ–ADV, DET–PRON, NOUN–PRON, NOUN–PROPN, INTJ–NOUN and between VERB and ADJ, ADV and NOUN (in subordinate structures). The IAA for German is higher, averaging at 74.24. The confusion in German POS tagging is almost exclusively between AUX–VERB, ADJ–ADV, NOUN–PROPN, and DET–PRON. The agreement is the lowest for language ID LANG3 (57.92), and highest for OTHER (86.33, non-linguistic tokens, and tokens whose language ID could not be determined).

The confusion between DET–PRON is common in both languages, since they share the same frequent word forms. The ADJ–ADV confusion seems to stem from the same reason. Again, AUX–VERB confusion is due to copular and auxiliary use of the same frequent tokens. Most NOUN–PROPN disagreements happen since, following UD, we tag parts of named entities as their respective POS tags, not as PROPN (for an example, see (5) in Section 4). Annotators tend to go against this guideline, and often tag parts of named entities as PROPN. Similarly, the guidelines require that parts of multi-word interjections should be tagged as their base POS tags. For example, the tokens in *Allaha şükür* 'Thank God' should be tagged PROPN and NOUN, while annotators may sometimes decide for INTJ for both.

Table 2 presents the distribution of POS tags

[7] http://universaldependencies.org/u/pos/index.html
http://universaldependencies.org/de/pos/index.html
http://bionlp-www.utu.fi/dep_search/

for each label used during language identification. Our total number of tokens per language is slightly different from Çetinoğlu (2016) due to segmentation. Majority of the tokens are Turkish. German follows Turkish after the label OTHER which includes all punctuation, symbols, numbers, URLs and Twitter-specific tokens.

One of the interesting observations in Table 2 is the high proportion of Turkish verbs (25% of all Turkish tokens) in comparison German verbs (15%). The reason for high rate of verbs are partially due to the fact that we mark all verbal forms, including all verbs derived into verbal nouns, participles, or converbs as VERB. However, this is true for both languages. The difference between the ratio of verbs in two languages has to do with the fact that most of the sentence are Turkish sentences. As a result, the predicates of main (and subordinate) clauses tend to be in Turkish, where German words are included in the (host) Turkish sentence. This is in line with the finding reported in Çetinoğlu (2016) that most tweets in this corpus have a majority of Turkish words. The ratio of nominals (NOUN, PRON and PROPN) are similar, having a distribution of 41% for German, and 40% for Turkish. POS tags with grammatical functions, such as ADP, AUX, DET and PART, are proportionally higher for German in comparison to Turkish. This is expected, since many of these grammatical functions are carried out as morphological processes in Turkish.

An interesting aspect of this corpus is rather high rate of MIXED tokens. Table 2 also shows that majority of the MIXED class involve PROPN and NOUNs, which is expected. In cases of mixed nouns or proper nouns, the mixed words are almost exclusively, DE or LANG3 (mostly English) words affixed by Turkish suffixes, e.g., (8) below. The mixed tokens that include verbs are predominantly German words with Turkish copular suffixes (9) or suffixes that derive verbs from nominals, as in *-len* suffix in (10). In some cases, German infinitives or participles are suffixed with Turkish nominal inflections (11). One last interesting case in (12) demonstrates that Turkish derivational suffixes that are normally attached to nouns or adjectives to form verbs may be attached to German (or, as in the example, English) verbs. In example (12), the suffix *-lu*[8] is attached to an English verb in a way to allow further verbal inflections.

(8) Bak şu benim
Look.VERB.Imp that.DET my.PRON
Lieblingsschwester§-im-a
favourite sister.NOUN.Sg-P1S-Dat
'Look at that favourite sister of mine'

(9) çok
very.ADV
kreativ§•miş
creative.ADJ•Cop.VERB.Evid.Past.3Sg
'he/she was very creative'

(10) **Kopie**§-len-ip
copy-Become.VERB-Sub
yapış-tır-ıl-mış
paste.VERB-Caus-Pass-Evid.Past.3Sg
'it was copied and (then) pasted'

(11) şu **kopieren**§-i
that.DET copy.VERB.Inf-Acc
icat ed-en
invention.NOUN.Sg do.VERB-Sub
'(the person) who invented (that) copying'

(12) Ben aslında
I.PRON in fact.ADV
FB§•lu-lar-ı
FB.PROPN.Sg•From.NOUN-Pl-Acc
follow§•lu-yor-du-m
follow.VERB•Derv-Prog-Past-1Sg
nur
only.ADV
'In fact, I was only following the ones of/from FB'

Turkish to German word-internal switches seem to predominantly involve introducing German nominals in Turkish host sentences. In 53% of the Turkish to German switches, the German word is NOUN, PROPN or PRON, in contrast to expected 41% in the complete corpus. The switches from German to Turkish does not have a clear pattern. For example, the ratio of Turkish nominals in German to Turkish switches amount to 40%, exactly as expected from the general corpus distribution.

7 Conclusion

In this work we present the POS annotation of a code-switching corpus created from Turkish-German tweets. The corpus has already been tokenised, normalised, and annotated with word-level language identification information.

[8] The original surface form of this suffix is -lA (-le/-la), it undergoes vowel harmony due to following suffix *-yor*.

Language	ADJ	ADP	ADV	AUX	CONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X	TOTAL
TR	767	289	1026	112	205	367	293	2563	52	41	691	428	3	40	1	2289	7	9174
DE	365	219	458	112	82	203	108	867	8	47	531	195	1	25	0	581	3	3805
LANG3	14	8	4	1	0	2	10	45	5	1	5	83	0	0	0	9	11	198
MIXED	10	0	1	0	1	0	0	97	0	0	1	73	0	0	0	6	1	190
AMBIG	4	0	1	0	0	0	7	18	0	0	0	11	0	0	0	1	0	42
OTHER	0	0	0	0	0	0	176	8	160	0	0	780	1820	0	820	0	101	3865
TOTAL	1160	516	1490	225	288	572	594	3598	225	89	1228	1570	1824	65	821	2886	123	17274

Table 2: Distribution of POS tag labels for each language identification label.

For POS annotation, we follow Universal Dependencies tokenisation and POS tagging policies as closely as possible. This requires revisiting tokenisation and aligning the language identification information with the new tokenisation as the first step.

Universal Dependencies is an evolving project. In its current version, German has a rather standardised tokenisation and less open questions regarding to POS and syntactic annotation as compared to Turkish. UD provides online documentation for German, the one for Turkish is work in progress. While we took the UD specifications as is for German, we developed our own annotation guidelines for Turkish, by adopting UD rules where applicable and by proposing our solutions to unresolved cases.

The resulting corpus contains 1029 tweets (17274 tokens) annotated with 7 different language IDs and 17 different POS tags. An obvious extension is to add morphological features as the next layer. This way we can better describe the distinctions among the words in the same category. For instance, it would be possible to distinguish Turkish verbal nouns, participles, and converbs that all have the `VERB` tag. We leave this finer-grained annotation as future work.

Another direction we want to pursue is experiments with automatic language identification and POS tagging. For other researcher who would like to conduct similar experiments, the corpus and the annotation guidelines are made available at `http://www.ims.uni-stuttgart.de/institut/mitarbeiter/ozlem/LAW2016.html`.[9]

Acknowledgments

We thank Sevde Ceylan, Hasret el Sanhoury, Esra Soydoğan, and Cansu Turgut for the annotation processes. 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*, pages 13–23, Doha, Qatar, October. 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.

Çağrı Çöltekin. 2015. A grammar-book treebank of Turkish. In Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk, and Adam Przepiórkowski, editors, *Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14)*, pages 35–49.

Çağrı Çöltekin. 2016. (When) do we need inflectional groups? In *The First International Conference on Turkic Computational Linguistics*, page (to appear).

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

Amitava Das and Björn Gambäck. 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.

Stefanie Dipper, Lukas Faulstich, Ulf Leser, and Anke Lüdeling. 2004. Challenges in modelling a richly annotated diachronic corpus of German. In *Workshop on XML-based richly annotated corpora, Lisbon, Portugal*, pages 21–29.

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.

[9] We follow the restrictions of Twitter's Terms of Service and distribute the tweet IDs instead of actual tweets. The scripts that combine downloaded tweets with annotations are also provided.

Jennifer Foster, Özlem Çetinoğlu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. 2011. From news to comment: Resources and benchmarks for parsing the language of web 2.0. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 893–901, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

John J. Gumperz. 1964. Linguistic and social interaction in two communities. *American Anthropologist*, 66(6):137–153.

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.

Aravind K Joshi. 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.

John Lipski. 1978. Code-switching and the problem of bilingual competence. *Aspects of bilingualism*, 250:264.

Suraj Maharjan, Elizabeth Blair, Steven Bethard, and Thamar Solorio. 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.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Melanie Neunerdt, Bianka Trevisan, Michael Reyer, and Rudolf Mathar. 2013. Part-of-speech tagging for social media texts. In *Language Processing and Knowledge in the Web*, pages 139–150. Springer.

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.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, page (accepted).

Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. 2003. Building a Turkish treebank. In Anne Abeille, editor, *Building and Exploiting Syntactically-annotated Corpora*. Kluwer Academic Publishers, Dordrecht.

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

Kemal Oflazer. 1999. Dependency parsing with an extended finite state approach. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 254–260. Association for Computational Linguistics.

Tuğba Pamay, Umut Sulubacak, Dilara Torunoğlu-Selamet, and Gülşen Eryiğit. 2015. The annotation process of the itu web treebank. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 95–101, Denver, Colorado, USA, June. Association for Computational Linguistics.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*, volume 59.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Shana Poplack. 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.

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.

Ines Rehbein. 2013. Fine-grained pos tagging of German tweets. In *Language Processing and Knowledge in the Web*, pages 162–175. Springer.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Gillian Sankoff. 1968. *Social aspects of multilingualism in New Guinea*. Montreal, McGill U.

Anne Schiller, Simone Teufel, and Christine Thielen. 1995. Guidelines für das tagging deutscher textcorpora mit stts. *Manuscript, Universities of Stuttgart and Tübingen*, 66.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer.

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

Thamar Solorio and Yang Liu. 2008b. Part-of-Speech tagging for English-Spanish code-switched text. In *Proceedings of the 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.

G Richard Tucker. 1999. A global perspective on bilingualism and bilingual education. *Georgetown University Round Table on Languages and Linguistics*, pages 332–340.

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.

Appendix A. Notation of Examples

German words are represented in bold and English words in italics in the examples. The POS tags in glosses correspond to the UD tags used in annotation. Gloss descriptions are given in Table 3.

Gloss	Explanation
Acc	Accusative case
Loc	Locative case
Dat	Dative case
Gen	Genitive case
Sg	Singular
Pl	Plural
1Sg	1st person singular
1Pl	1st person plural
3Pl	3rd person plural
P1S	1st person possessive
P3S	3rd person possessive
Past	Past tense
Fut	Future tense
Prog	Progressive tense
Caus	Causative
Pass	Passive
Imp	Imperative
Part	Participle
Inf	Infinitive
Evid	Evidentiality
Cop	Copular
Become	Derivational suffix with semantics 'become'
From	Derivational suffix with semantics 'of/from'
Sub	Subordinating derivational suffix
Derv	Derivational

Table 3: Gloss descriptions