The Denglisch Corpus of German-English Code-Switching

Doreen Osmelak
Language Science and Technology
Saarland University, Germany
s9doosme@stud.uni-saarland.de

Shuly Wintner
Department of Computer Science
University of Haifa, Israel
shuly@cs.haifa.ac.il

Abstract

When multilingual speakers involve in a conversation they inevitably introduce *code-switching (CS)*, i.e., mixing of more than one language between and within utterances. CS is still an understudied phenomenon, especially in the written medium, and relatively few computational resources for studying it are available.

We describe a corpus of German-English code-switching in social media interactions. We focus on some challenges in annotating CS, especially due to words whose language ID cannot be easily determined. We introduce a novel schema for such word-level annotation, with which we manually annotated a subset of the corpus. We then trained classifiers to predict and identify switches, and applied them to the remainder of the corpus. Thereby, we created a large-scale corpus of German-English mixed utterances with precise indications of CS points.

1 Introduction

Multilinguality is becoming more and more ubiquitous, to the extent that psycholinguists increasingly acknowledge that bilingualism is the rule and not the exception (Harris and McGhee Nelson, 1992). Grosjean (2010, p. 16) stated that "bilingualism is a worldwide phenomenon, found on all continents and in the majority of the countries of the world" and Grosjean and Li (2013) assessed that more than half the world's population is multilingual.

Multilingual speakers have two or more language systems active in their minds, and they tend to use them interchangeably, especially when communicating with other multilinguals. This process of mixing two or more languages within a discourse or even within a single utterance is called *code-switching (CS)*. In order to understand and produce natural language, NLP systems need to cope with this phenomenon, but today's language technology still cannot efficiently process CS, partly due to lacunae in our understanding of the factors driving CS, and partly due to lack of resources.

We introduce a corpus of German-English CS in spontaneous written communication.[1] We discuss challenges in determining the language ID of tokens in multilingual texts in Section 4, and present our novel annotation scheme in Section 5. We describe the corpus in Section 6, and then describe classifiers (Section 7) that accurately identify the language ID of tokens in the corpus, thereby allowing us to effectively identify switch points in unseen texts. We conclude with suggestions for future research.

2 Background and Related Work

The Phenomenon of CS Code-switching is the process of mixing two or more languages within a discourse or even within a single utterance, where the mixed words or fragments do not suffer any syntactic or phonological alternation. CS can happen on various linguistic levels (phonological, morphological, lexical, syntactic), and can be *intra-sentential* (the switch occurs within the boundaries of a sentence or utterance), or *inter-sentential* (the switch occurs between two sentences or utterances). There are two competing theories on how this process works: as a symmetric relation or as an asymmetric relation. In the *symmetric approach* both languages are equally dominant, and any lexical items from either language can be replaced by the corresponding items of the other language, as long as the switch happens at syntactic boundaries that are shared by both languages. The monolingual fragments conform to the grammar of the corresponding language they are taken from (Poplack, 1980). In the *asymmetric approach* one of the languages is more dominant than the other, and only content morphemes can be taken from both languages, whereas late system morphemes indicating grammatical relations can only be taken from the subordinate language. The dominant language

[1] All the data and code developed in this work are available at https://github.com/HaifaCLG/Denglisch.

from which the grammatical framework is taken is called the *Matrix Language*, and the subordinate language that is mixed into it is called the *Embedded Language* (Joshi, 1982).

Oral CS CS in oral communication has been studied extensively. It interacts with speakers' proficiency as well as style and content of the utterances, serving several, partly contradicting, purposes, such as compensating for words the speaker does not know in one language or expressing nuanced meanings that cannot be expressed precisely with the other language (Gardner-Chloros, 2009). But CS can also serve sociolinguistic purposes such as conveying identity, interpersonal relations and formality. Conclusions from past research have differed greatly in whether CS is a strategy used by highly adaptive speakers to convey very subtle meaning differences between words of different languages (Kootstra et al., 2012), or a strategy used by speakers less familiar with one of the languages to overcome lexical deficiencies (Poulisse, 1990).

Written CS CS in written communication has not drawn much attention in research so far. Written communication differs significantly from spoken interaction, especially in formality and spontaneity: e.g., literary texts undergo an inherent process of conscious reflection, correction, editing and review. Findings thus far have differed on whether oral and written CS behave in the same manner and serve the same purposes. Written CS in literary texts does partially serve the same purposes as in spoken CS (Gardner-Chloros and Weston, 2015), but there are additional functions and purposes that are not found in spontaneous oral speech, such as serving as a poetic device (Chan, 2009).

Online Forums With the increasing ubiquity of online discussion platforms, there are large amounts of written communication reflecting more spontaneous speech productions than classical written texts, thereby constituting a hybrid between speech and formal writing. Research on CS in online forums has so far mainly focused on computational challenges for NLP algorithms (Çetinoğlu et al., 2016). Sociolinguistic aspects of the communicative purpose of CS in these settings are severely understudied. Most sociolinguistic works mainly focused on very limited data of a small number of language-pairs or authors (Sebba et al., 2011).

Rabinovich et al. (2019) developed a large-scale corpus of written CS data from Reddit posts containing various languages switched with English, but not including the German-English pair that we focus on here. They compared monolingual and code-switched posts, finding that there are topical and stylistic distinctions, as well as a difference in the proficiency of speakers. Shehadi and Wintner (2022) compiled an Arabizi corpus from Twitter and Reddit posts which contains CS between Arabic, English and French, and trained classifiers to identify switches.

Annotating CS Language annotation of bilingual data is not always trivial (Clyne, 2003; Alvarez-Mellado and Lignos, 2022), especially when borrowings and named entities are involved. Borrowing is a continuous process, with different stages, where a word is first introduced as a completely foreign sounding word and is then phonologically and morphologically adapted to the borrowing language, until it becomes a common word of the language's lexicon. Clear cuts on when a word is still to be considered a foreign word or already a common word of the language are hard to make. Due to the geographical and phylogenetic closeness of German and English and their common cultural and religious roots, it is often hard to determine whether a word is borrowed, adapted, foreign or native to the language. Alvarez-Mellado and Lignos (2022) added a "language" tag, BOR, to indicate recent borrowings, in addition to a tag for named entities. Shehadi and Wintner (2022) proposed the use of a *shared* category for words that can be used in both languages. We further refine their annotation scheme and the definition of the *shared* category. For a different approach to language ID annotation of multilingual texts, see Zhang et al. (2018).

Predicting CS Points CS is influenced by various sociolinguistic characteristics, such as topic and setting or the speakers involved in the conversation and their level of familiarity. It can serve several sociopragmatic functions such as direct quotation, emphasis, clarification, parenthetical comments, etc. Several linguistic features can be exploited for predicting CS points. Soto et al. (2018) showed that POS-tags, cognates, and entrainment of a word can trigger switches on the succeeding word, but not on the preceding word. This suggests that predicting CS points from the previous words alone is possible. Solorio and Liu (2008) predicted CS points using lexical and syntactic features, such

as tokens, part-of-speech (POS) tags, and tree tags. Recent works show that the strong relationship between CS and cognate words, as proposed by Clyne (1967, 1980, 2003) in the *Triggering Hypothesis*, can be used to improve language models (Solorio and Liu, 2008; Soto and Hirschberg, 2019).

It is important to note that predicting CS is a difficult task because CS is a subjectively motivated process, subject to the speaker's preferences and background. Clearly, bilingual speakers do not *have* to code-switch, as by definition they can converse in any of their two languages. Understanding when and where they do code-switch is an ultimate goal of our research program, but undoubtedly some degree of arbitrariness is inherent to the phenomenon. Solorio and Liu (2008) therefore proposed the use of human judgments additionally to standard statistical evaluation measures.

3 Experimental Setup

Data *Reddit* is a large-scale social news and discussion platform, with several hundreds of thousands of sub-categories (*sub-reddits*) on different topics, and over 100 million new posts a year. There are many region-based sub-reddits, which attract large bilingual communities. The posts and comments are length-unlimited, and unlike in lab-settings the interlocutors produce language spontaneously, which allows us to analyze natural conversation flow.

German is one of the most widely-used languages in the world. With approximately 100 million native speakers, it is the most prevalent mother-tongue and, after English, the most widely understood language in Europe.[2] Since English is the world's main lingua franca, that non-native speakers across Europe use on a regular basis, German speakers are constantly exposed to English (through movies, music, the Internet, etc.) and CS exists in their daily life. It is thus worthwhile to investigate CS in German-English. However, to the best of our knowledge, no corpus or any work on written German-English CS is available, although a German-Turkish corpus of Twitter posts does exist (Çetinoğlu, 2016).

Most existing CS corpora and studies on CS use language pairs in communities where both languages are either co-official or co-native to the community (e.g., Hindi-English, where English is an official language and a lingua franca throughout India (Ganji et al., 2019); Maltese-English, where Malta as a former British colony maintained English as a lingua franca (Camilleri Grima, 2013); Turkish-German in the German-Turkish community (Çetinoğlu, 2016); etc.) Here, we address CS in a country that is officially monolingual (German) and neither has a major community of English-natives nor uses English as a lingua franca.

We investigate German-English CS using country-specific sub-reddits for German-speaking countries/regions, like `r/Germany` or `r/DE`. Since these sub-reddits contain discussions about region-based topics, we expect authors in these communities to be speakers of both German and English.

Statistical Classification We use (supervised) statistical classification in order to identify CS points. *Statistical classification* is the problem of identifying to which of at least two categories a given observation belongs. A classifier is trained on labeled examples, i.e., instances of which the classification is known a priori. Each instance is represented by a set of features, to which the classifier assigns weights during training. Given that the chosen features are actually relevant for the classification and given that the training set is large enough, the classifier can then predict the category of a new unseen instance. We use *Conditional Random Fields (CRF)* for the classification (Lafferty et al., 2001); CRF is a sequence to sequence classifier that uses its predictions on the previous instance in order to predict the label of the current instance.

Linguistic interpretation of the results can help us extend our knowledge of CS. By predicting CS points, we can learn about the specific features of language that trigger CS or discourage it. Such linguistic insights into the CS process can be used to build NLP systems that better cope with CS and multilingual discourse.

4 Shared Lexicon

The key to identifying CS points is precise annotation of the language ID of each token in the text. In multilingual texts, this problem is non-trivial (Alvarez-Mellado and Lignos, 2022). We now discuss some of its challenges. We provide examples from German-English, but most of the observations are valid for any language pair.

Many words are shared across the German and English lexicons. We differentiate between *inher-*

[2] https://en.wikipedia.org/wiki/List_of_European_languages_by_number_of_speakers

ited words, or *cognates*, which developed from words in an earlier stage of the language, and *borrowed words*, which are taken from or developed from words of another language. Borrowing is a continuous process with different stages: words are first introduced into the language as a completely foreign sounding word; they are gradually adapted to the phonological and morphological rules of the borrowing language until eventually they are considered to be common words of the language (Haspelmath, 2009; Grant, 2015; Campbell, 2020).

Loan words are fully integrated borrowings, i.e., fully adapted to the borrowing language in flexion, phonology and orthography. Borrowings without (or with minimal) adaptions are called *foreign words*. A *pseudo-borrowing* is a word created from elements of a borrowing language, but which does not exist in the donor language (e.g., *Handy–cellphone*) (Bussmann, 2008; Campbell, 2020).

The reasons that words are borrowed or shared across languages include geographical language contact, phylogenetic closeness, and common cultural background (Haspelmath, 2009; Grant, 2015; Campbell, 2020). It is not always easy to tell whether a word is borrowed, native, or a switch.

German and English are both Germanic languages, which share a common ancestral lexicon and many similar-looking words. Both languages were religiously and culturally influenced by Greek and Latin. Nevertheless, words can be marked by native morphology or orthography and some of these adaptions may intuitively look more German than others (e.g., *-**ier**en*, which is usually used on long integrated Latin words instead of *-en*). Further, not all of these words are actually shared. Many Latinate words entered English through Old French and by today either displaced their Germanic equivalents or shifted their meaning.

Many cultural terms are borrowed into other languages as full new concepts without translations. This is the case for modern inventions, but many everyday words entered the German lexicon centuries ago, and native speakers are often unaware of their foreign roots (e,g, *meschugge, Schal, cotton, assassin*).

Named entities are usually borrowed without translation, but they may take different forms: they can be shared completely or adapted orthographically, phonologically, and morphologically, sometimes with very distant looking forms, or even be taken from different etymological roots. Additionally, they can take derivational or inflectional morphemes of the borrowing language or even be used in compounds with native words.

Additional challenges are due to the fact that some very high-frequency words share spelling with a word of the other language (e.g., *was*, *die*) although they are totally unrelated. Furthermore, words can be composed of components of two different languages (e.g., *Pushnachrichten–push notifications*).

Using English entities like *Fifth Avenue*, or untranslatable terms like *hamburger*, in a German sentence cannot be considered a regular switch, since there is no actual German equivalent for such terms. Nevertheless, the use of these terms might activate the English lexicon and trigger a future switch. The extent of such triggering may be reduced for entities or terms that are adapted to German in orthography and morphology. These considerations are the motivating principles for our annotation scheme, which we now present.

5 Annotation Scheme

We introduce a novel, highly-detailed annotation scheme that reflects the observations of Section 4 above. We present the scheme in Section 5.1, and then propose a flattened version of it in Section 5.2. Crucially, while we define the schemes and exemplify annotated instances in terms of English and German, the schemes are applicable to any language pair.

5.1 Detailed Annotation Scheme

The annotation scheme is summarized in Table 1. We defined the following basic categories:

English (1): pure/regular English words.

German (2): pure/regular German words.

Overlap (3): words that belong to both mental lexicons, including shared and adapted named entities (`3a`), borrowed words (`3`), language-mixed words (`3c`), and words that overlap in the given context (`3b`).

Neutral (4): tokens that are language universal, including numbers (`4b`), emoticons (`4c`), interjections (`4d`), and words of other languages than English and German (`4a`).

In addition, we sometimes add the origin of the word to the tag, as a suffix `-E` for English, `-D` for German, and `-O` for other. We now explain how we assign labels to the problematic cases described in Section 4.

1	English				
2	German				
3	Overlaps				
	3a Named Entities		3c Merge-Words		3b Ambiguous Words
	3a-E English Origin		3c-C Compounds		
	3a-D German Origin		3c-M Morphology		3-E Untranslatable English
	3a-AE Adapted to English		3c-EC Entity Compounds		3-D Untranslatable German
	3a-AD Adapted to German		3c-EM Entity Morphology		3-O Untranslatable Other
4	Neutral				
	4a Foreign	4b Numbers	4d Interjections		<url> URL
		4b-E English only	4d-E English only		<punct> Punctuation
		4b-D German only	4d-D German only		<EOS> End of Sentence
		4c Smiley	4e-E English abbr.		<EOP> End of Paragraph

Table 1: Detailed Annotation Scheme.

Named Entities are often borrowed and shared across languages. They can be adapted to the borrowing language on all linguistic levels. We introduce the following subcategories: *NE of German Origin* (3a-D), *NE of English Origin* (3a-E), *NE Adapted to German* (3a-AD), *NE Adapted to English* (3a-AE), *NE of Other origin* (3a). We differentiate among the following adaption cases:

Unadapted entities: entities that do not show any kind of adaption to the borrowing language or are native to the language (*Paris*, *Berlin*) are tagged according to their origin (3a-E for English, 3a-D for German, 3a for Other).

Translated entities: entities that are full translations (*United Kingdom–Vereiniges Königreich*) or stem from different etymologies (*Germany–Deutschland*) are considered regular words (1 / 2).

Orthographic adaptions: entities that have only spelling differences due to orthographical rules (English */c/* vs. German */k/*) or pronunciation are tagged equally to the original name.

Morphologic adaptions: major phonological and morphological adaptions in the entity itself affect the annotation in case they identify one of the languages (*Kalifornien–California*, where *-ien* is a German location morpheme). Such entities are tagged as Adapted Entities (3a-AE for adapted to English, 3a-AD for adapted to German). Entities that show case or plural markings (*Münchens*, where *-s* is a genitive morpheme) are also Adapted Entities.

Lexical adaptions entities containing translated word parts (*New Zealand–Neuseeland*) are considered Adapted Entities. Prefixes of other languages than German and English ('*anti-*') were not relevant for the annotation.

We consider the following to be NEs: geographical location as well as their demonyms, including religious and ethnic or tribal groups, as well as language communities, persons, companies and organizations, names of weekdays and months, units, and measures. The origin of an entity was identified by etymological roots, and phonetic, phonological and lexical features of the word.

Borrowings Often, words are borrowed as new concepts without any native translation. This is especially the case for modern inventions (*Smartphone*) and cultural terms related to food (*Döner*), religion (*Hijab*), festive activities and traditions (*Oktoberfest*) and philosophy/ideology (*LGBTQ*, *Feng Shui*), including academic and honorific titles (*Tsar*, *Shah*).

We differentiate among the following cases:

Established untranslatables: well-established cultural and technological terms without native translations are tagged as 3-E/D/O according to their origin.

Unestablished untranslatables: unestablished technical terms common only to certain groups (*Blockchain*) and terms that only recently entered the lexicon (*Lockdown*) are tagged as regular English words (1).

Translatables: Borrowings that have translation equivalents that could have been chosen instead (*Bildschirm–Display*), are tagged as regular words (1 / 2).

Integrated old loans: Words that originate from a third language, e.g., Old French, Latin, Greek, Arabic, or Persian, and have been fully integrated in the language (e.g., *cemetery*, *origin*, *assassin*, *coffee*, *cotton*), including Greek or Latin prefixes, are considered regular words (1 / 2).

Unintegrated old loans: Many unintegrated Latin words are found in abbreviations (e.g., *PS*) and were tagged as 4a. Those Latin abbreviations that are spelled out with English

words and are not used in German (*e.g.*) are tagged as English (1).

Neologisms and pseudoborrowings: Borrowed Greek and Latin neologisms (*video*) are tagged as 3-O. Pseudo-borrowings (*Handy*) are tagged as 3-E.

Mixes Borrowed words can be compounded with native words (***Wohlstands****bubble*) or morphologically adapted to the borrowing language (***ge****sterotyped*). Such words contain intra-word switches. We differentiate:

Compounding: Compounds of an English and a German word are tagged 3c-C.

Flexion: English words with German flexion morphemes are tagged 3c-M.

The same is possible with borrowed NE:

Entity Compounds: Borrowed English entities (3a-E, 3a-AE) with German flexion (*google****n***) or vice versa are tagged as 3c-EM.

Inflected Entities: English Entities compounded with German words (*NRA-****mäßig***) are tagged as 3c-EC.

Compounds and flexion on NEs of the same or a third language are tagged as Adapted Entities.

Ambiguous Cases Words that cannot be identified as German or English in the given context due to overlapping spelling and meaning and switches occurring around them (*taxes with a* ***separate*** *Einnahmen-Überschussrechnung* ***plus*** *Umsatzsteuererklärung*) are tagged as 3b.

Language Markings on Neutral Items Neutral language-universal tokens like numbers and interjections can bear cues to the active language lexicon (*90s–90er*, *10th–10ter*, *ähm–erm*, *achso*). Those tokens that are specific to one lexicon are tagged as *English/German use only* (4b-E/D, 4d-E/D), those used in both languages as 4b, 4d. English language abbreviations that are used as interjections across languages (*lol*, *rofl*) are tagged 4e-E.

5.2 Collapsed Annotation Scheme

These categories were over-refined, and some of them had relatively few occurrences in our corpus. We therefore defined a collapsed version of the scheme, as shown in Table 2.

English (E): all English words (1), English numbers and interjections (4b-E, 4d-E).

German (D): all German words (2), German numbers and interjections (4b-D, 4d-D).

Mix (M): words containing properties of both languages, including intra-word switches (3c-(E)M/(E)C).

Shared English (SE): all English words that are used in both languages (3a-(A)E, 3-E, 4e-E).

Shared German (SD): all German words that are used in both languages (3a-(A)D, 3-D).

Shared Other (SO): all words of other origin that are used in both languages (3a, 3-O, 4a), including shared interjections (4d) and other overlaps (3, 3b).

Other (O): all tokens that are language independent, including neutral number constructions, emoticons, and punctuation (4b, 4c, <punct>, <url>, 4).

E	English	1, 4b-E, 4d-E
D	German	2, 4b-D, 4d-D
M	Mix	3c, 3c-C, 3c-M, 3c-EC, 3c-EM
SE	Shared English	3a-E, 3a-AE, 3-E, 4e-E
SD	Shared German	3a-D, 3a-AD, 3-D
SO	Shared Other	3, 3a, 3b, 3-O, 4a, 4d
O	Other	4, 4b, 4c, <punct>, <url>

Table 2: Collapsed Annotation Scheme.

6 Corpus Creation

We used a modified version of the method used by Rabinovich et al. (2019) to collect and extract our data. We downloaded approximately 17 million comments from the German-language sub-reddits *r/DE*, *r/Deutschland*, *r/Germany*, *r/Berlin* using the Pushshift Reddit API. We extracted 10,000 comments that potentially contained CS using the Polyglot language detector. We[3] annotated 950 of the extracted comments manually following the detailed scheme of Section 5. These contained over 75,000 tokens in 4,200 sentences, of which 1,250 contained intra-sentential switches. We then generated a version with the collapsed annotation scheme.

We then downloaded another set of 25.5 million comments from German-language sub-reddits, including also sub-reddits dedicated to cities and regions in Austria and Switzerland, as well as a few general topics. Of those, 21,500 comments were extracted as potentially including switches. These comments, together with the remainder of the initially downloaded comments, were used to create a larger automatically annotated corpus. The data for the automatic annotation thus consists of 31,500

[3]All annotation was done by the first author. We therefore cannot report inter-annotator agreement.

comments containing 230,000 sentences with over 5 million tokens.

To identify code-switches in the automatically-tagged corpus we use two different criteria. The *strict* definition requires a sentence to contain at least one word annotated "pure English" (1), and at least one tagged as "pure German" (2). The *relaxed* definition only requires a token tagged as English-origin, excluding named entities (1,4b-E,4d-E,3-E) and a token similarly annotated as German-origin, excluding NEs (2,4b-D,4d-D,3-D), or a token tagged as Merge-Word, excluding NE-Merger (3c-M,3c-C). Table 3 lists data on the manually-tagged and automatically-tagged corpora. It reports the total number of sentences in each corpus, the number of sentences containing CS (both strict and relaxed), and the number of posts containing CS (for posts, the strict and relaxed numbers are almost identical).

We now provide some observations on the manually-annotated portion of the corpus.

Amount of Switches The portion of bilingual posts was very small, only 0.62‰ of the downloaded comments. A considerable amount of the bilingual raw data contained the second language only as citations or as titles (of books, movies, songs, etc.)

Types of Switches Many of the extracted posts contained switches on sentence boundaries. Intra-sentential switches were often *insertional*, i.e., comprised of only a single switched word or construct of a few switched words in an otherwise monolingual sentence. Intra-word switches do exist, especially as German flexion and derivation on English words and entities.

Topics A few topical peculiarities were striking: computer and gaming related terms as well as social media related terms were often switched to English in otherwise German comments; terms related to politics, authorities, law or regulations were often switched to German in otherwise English comments.

7 Identifying Switches

In order to identify switches in an unseen utterance, we need to identify the language ID tag of the words in the sentence. We now describe a classifier that establishes this task.

7.1 Word-Level Classification

We used CRF to train a sequence to sequence classifier, using various features we list below. We opted for more traditional, statistical classification rather than neural classification both because we were interested in interpreting the features and because Shehadi and Wintner (2022), on a very similar task, report that both methods yielded almost identical accuracy.

Orthography: the word in lower case; whether the word is in upper, lower or all-upper case; whether the word is is an emoji or emoticon; whether it contains digits, punctuation, or special German letters (*ü, ö, ä, ß*).

N-Grams: whether the word contains one of the most frequent English or German letter bi- and trigrams; 400 most frequent n-grams in the corpus as separate features.

Morphology: whether the word contains German or English derivational or inflexional affixes, including common verbal prefixes and noun and adjective suffixes.

Function Words: whether the word is included in German or English lists of function words.[4]

Frequency: whether the word is in the 207 most frequent German words, or the 5050 most frequent English words, taken from the one billion word Corpus of Contemporary American English.

Lexical Components: whether the word contains lexical parts that are regularly used in German or English named entities, e.g., *weiler, burg, neu; borough, dale, port.*

Word Lists: several word lists for named entities and cultural terms, e.g., the names of the biggest German cities or companies.

We used 10-fold cross-validation for evaluation. The evaluation results are listed in Table 4, reflecting an overall accuracy of 0.965.

7.2 Sentence-Level Classification

Following Shehadi and Wintner (2022) we combined the results of the word level annotation to form bit-vector annotations for sentences. Each sentence is thus associated with a single bit-vector indicating which of the language category tags are present in it. We then trained a classifier to predict the full bit-vectors at the sentence level. The results, reflecting the accuracy of the sentence-level classifier on each category, are presented in Table 5.

[4]We compiled these lists and will make them available.

Corpus	Sentences	Strict CS	Relaxed CS	Posts with CS
Manually-tagged	4,200	1,250	1,400	950
Automatically-tagged	228,800	72,250	74,000	30,150
Total	233,000	73,500	75,400	31,100

Table 3: Statistics of the corpora: the total number of sentences in each corpus, the number of sentences containing CS (both strict and relaxed), and the number of posts containing CS.

Tag	Prc	Rcl	F1	Support
English	0.97	0.98	0.98	29918
German	0.96	0.98	0.97	29730
Mix	0.50	0.19	0.28	246
Shared English	0.82	0.55	0.66	699
Shared German	0.78	0.54	0.64	807
Shared Other	0.75	0.50	0.60	1108
Other	0.99	0.98	0.99	12505
Micro Avg	0.96	0.96	0.96	75013
Macro Avg	0.82	0.68	0.74	75013
Weighted Avg	0.96	0.96	0.96	75013

Table 4: Results: Word-Level Classification.

The overall accuracy of predicting the full bit-array of a sentence correctly is 0.764.

Tag	Acc	Prc	Rcl	F1
English	0.95	0.96	0.96	0.96
German	0.96	0.97	0.98	0.97
Mix	0.96	0.59	0.26	0.36
Shared English	0.95	0.86	0.68	0.76
Shared German	0.95	0.81	0.65	0.72
Shared Other	0.92	0.83	0.61	0.70
Other	1.00	1.00	1.00	1.00

Table 5: Results: Sentence-Level Classification.

7.3 Analysis

Mix Many of the words classified as Mix were seen in the training corpus. Some of the misclassifications of `3c-M` and `3c-C` on full-German words (*ge**back**en–baked*, *Krisenter**mine**–crisis dates*) indicate that the classifier actually learns to classify words as Mixed that could be decomposed to parts reflecting both languages.

Untranslatables Identifying untranslatables works relatively well even with only few training instances, probably due to the word lists. Most of the words tagged as `3-E/D/O` were actually contained in the word lists.

NEs Most of the words classified as Adapted Entities contain one of the derivation suffixes (*-ish*, *-ian*, etc.) Many words classified as `3a-D` contained lexical features of the Lexical Components lists, this was not observed for `3a-E`. This might be due to the training corpus containing several German person and town names, but not many English ones.

Ambiguous The classification of ambiguous words is rather poor, probably because identifying whether the word can be disambiguated in the context is a very subjective feature and only very few examples were seen in training. It mainly classifies some of the instances of the words seen as `3b` in training as `3b`.

8 Conclusion

We presented a corpus of German-English code-switched utterances from user generated social media content, which contains precise language annotation indicating code switches. Our corpus is partly hand-annotated and partly automatically annotated. We addressed some challenges in annotation of multilingual data by introducing various types of shared and mixed categories. We trained classifiers to predict our word-level annotation and switch-points. First experimental results from the prediction of switch-points indicate that properties of shared and mixed words are relevant factors for CS. This encourages us to use our corpus as a basis for further sociolinguistic research on spontaneous written CS, specifically for studying the use and effects of Shared and Mixed words on switches in German-English and how these compare to other language pairs. Such work is currently underway.

Ethical considerations

This research was approved by the University of Haifa IRB. We collected data from a social media outlet, Reddit, in compliance with its terms of service. For anonymity, we systematically replaced all user IDs by unique IDs; we do not have, and therefore do not distribute, any personal information of the authors. With this additional level of anonymization, we anticipate very minimal risk of abuse or dual use of the data.

Limitations

Like any other dataset, the corpus we report on here is not representative. In particular, it probably includes German as used mainly by users highly fluent in English. It is very likely unbalanced in terms of any demographic aspect of its authors. Clearly, the automatic annotation of language IDs is not perfect, and may introduce noise, especially on the smaller and more subjective categories (e.g., `3b`, `M`). Further, when extracting the comments for the final corpus, very short comments were not included and comments with only a single switch or borrowed word might have been skipped, due to the rather low sensitivity of the language detector. Use of this corpus for linguistic research must therefore be done with caution. Nevertheless, we trust that the sheer size of the dataset would make it instrumental for research on code-switching in general and in German-English in particular.

Acknowledgements

We are grateful to Yuli Zeira and Safaa Shedahi for great ideas and fruitful discussions, and to the anonymous reviewers for their constructive comments. This work was supported in part by grant No. 2019785 from the United States-Israel Binational Science Foundation (BSF), and by grants No. 2007960, 2007656, 2125201 and 2040926 from the United States National Science Foundation (NSF).

References

Elena Alvarez-Mellado and Constantine Lignos. 2022. Borrowing or codeswitching? Annotating for finer-grained distinctions in language mixing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3195–3201, Marseille, France. European Language Resources Association.

Hadumod Bussmann. 2008. *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag, Stuttgart.

Antoinette Camilleri Grima. 2013. Challenging Code-Switching in Malta. *Revue Française de Linguistique Appliquée*, 18:45–61.

Lyle Campbell. 2020. *Historical Linguistics - An Introduction*, fourth edition. Edinburgh University Press.

Özlem Çetinoğlu. 2016. A Turkish-German code-switching corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4215–4220, Portorož, Slovenia. European Language Resources Association (ELRA).

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. Association for Computational Linguistics.

Brian Hok-Shing Chan. 2009. English in Hong Kong Cantopop: Language choice, code-switching and genre. *World Englishes*, 28(1):107–129.

Michael G. Clyne. 1967. *Transference and triggering: Observations on the language assimilation of post-war German-speaking migrants in Australia*. Martinus Nijhoff, The Hague, Netherlands.

Michael G. Clyne. 1980. Triggering and language processing. *Canadian Journal of Psychology*, 34(34):400–406.

Michael G. Clyne. 2003. *Dynamics of language contact*. Cambridge University Press, Cambridge, UK.

Sreeram Ganji, Kunal Dhawan, and Rohit Sinha. 2019. IITG-HingCoS Corpus: A Hinglish Code-Switching Database for Automatic Speech Recognition. *Speech Communication*, 110:76–89.

Penelope Gardner-Chloros. 2009. *Code-Switching*. Cambridge University Press.

Penelope Gardner-Chloros and Daniel Weston. 2015. Code-Switching and Multilingualism in Literature. *Language and Literature*, 24(3):182–193.

Anthony P. Grant. 2015. Lexical Borrowing. In *The Oxford Handbook of the Word*. Oxford University Press.

François Grosjean. 2010. *Bilingual: Life and Reality*. Harvard University Press.

François Grosjean and Ping Li. 2013. *The Psycholinguistics of Bilingualism*. Wiley-Blackwell.

Richard Jackson Harris and Elizabeth Marie McGhee Nelson. 1992. Bilingualism: Not the exception any more. In Richard Jackson Harris, editor, *Cognitive Processing in Bilinguals*, volume 83 of *Advances in Psychology*, pages 3–14. North-Holland.

Martin Haspelmath. 2009. Lexical Borrowing: Concepts and issues. In *Loanwords in the World's Languages: A Comparative Handbook*. De Gruyter Mouton.

Aravind K. Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*.

Gerrit J. Kootstra, Janet G. van Hell, and Ton Dijkstra. 2012. Priming of Code-Switches in Sentences: The Role of Lexical Repetition, Cognates, and Language Proficiency. *Bilingualism: Language and Cognition*, 15(4):797–819.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Shana Poplack. 1980. Sometimes I'll start a sentence in spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching 1. *Linguistics*, 18(7-8):581–618.

Nanda Poulisse. 1990. *The Use of Compensatory Strategies by Dutch Learners of English*, volume 8 of *Studies on Language Acquisition*. Cambridge University Press.

Ella Rabinovich, Masih Sultani, and Suzanne Stevenson. 2019. CodeSwitch-Reddit: Exploration of written multilingual discourse in online discussion forums. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, page 446, Hong Kong, China. Association for Computational Linguistics.

Mark Sebba, Shahrzad Mahootian, and Carla Jonsson. 2011. *Language Mixing and Code-Switching in Writing: Approaches to Mixed-Language Written Discourse*. Routledge.

Safaa Shehadi and Shuly Wintner. 2022. Identifying code-switching in Arabizi. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 194–204, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Victor Soto, Nishmar Cestero, and Julia Hirschberg. 2018. The Role of Cognate Words, POS Tags and Entrainment in Code-Switching. In *Proceedings of Interspeech 2018*, pages 1938–1942, Hyderabad, India. ISCA.

Victor Soto and Julia Hirschberg. 2019. Improving Code-Switched Language Modeling Performance Using Cognate Features. In *Proceedings of Interspeech 2019*, pages 3725–3729, Graz, Austria.

Yuan Zhang, Jason Riesa, Daniel Gillick, Anton Bakalov, Jason Baldridge, and David Weiss. 2018. A fast, compact, accurate model for language identification of codemixed text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 328–337, Brussels, Belgium. Association for Computational Linguistics.