Challenges of Annotating a Code-Switching Treebank

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

This paper presents challenges and observations on creating a code-switching treebank based on ongoing annotation efforts of a Turkish–German spoken corpus following the Universal Dependencies annotation scheme. We present and discuss a number of issues that arise because of the need for consistent multilingual annotation within a single treebank, as well as the informal language which is where code-switching is observed most. Besides proposing solutions to these issues, our aim in this paper is to stimulate discussion and facilitate consistency over upcoming code-switching annotation projects.

1 Introduction

Code-switching (CS) is the process of mixing more than one language in written or spoken communication (Myers-Scotton, 1993; Poplack, 2001; Toribio and Bullock, 2012). It is a phenomenon commonly observed in multilingual societies (Auer and Wei, 2007), mainly in informal settings such as social media and spoken communication. For instance, (1) shows a sentence from a dialogue, that mixes Turkish and German (in bold). The speaker starts with Turkish, switches to German, back to Turkish, and ends the sentence with a mixed word where the German noun *Gastfamilie* 'host family' is inflected with the Turkish locative suffix *-de*.

(1) Eh orada iki **Wochen** kaldım **ehm Gastfamilie**de **ehm** .
Eh there two week.Pl stay.Past.1sg ehm guest family.Loc ehm .
'I stayed there two weeks, in a host family.'

The sentence is relatively simple and the overall meaning is derivable from the individual words. Yet, its syntax is not standard. The main predicate *kaldım* 'I stayed' is in Turkish and the whole sentence seemingly follows the Turkish syntax, except the noun phrase *iki Wochen* 'two weeks'. Nouns modified by numbers are in singular in Turkish, but *Wochen* is in plural. The construction is more complex than using the German equivalent of 'week' in a Turkish phrase. It seems, the speaker inherently switches to the German syntax as well, where the noun should be plural, when switching to German on the surface.

Such CS-specific constructions vary from non-canonical morphological marking to creating new syntactic representations, to applying a linguistic phenomenon of one language to the other. They make structural analysis of code-switching linguistically interesting and computationally challenging. Several approaches tackle these challenges by utilising labelled and unlabelled monolingual and parallel data, e.g. by creating artificial CS data and using them in training models for processing CS (Pratapa et al., 2018; Zhang et al., 2018). However to be able to capture unique cases like the singular-to-plural mapping for 'week' in (1), those models need to see such instances. Thus, to observe the characteristics of CS and address them with data-driven tools, we are in the process of annotating Turkish–German transcriptions with part-of-speech, morphology, and dependency layers.

We have chosen Universal Dependencies (UD) (Nivre et al., 2016) as our annotation scheme. The UD project aims to define morphosyntactic annotation guidelines that are consistent across languages. Its unified tag sets and annotation standards facilitate the annotation of multiple languages within a single treebank. Furthermore, annotations parallel to monolingual resources are useful for making use of these resources, e.g., for transfer learning (Bhat et al., 2018).

Despite clear advantages of the UD framework for annotating CS treebanks, the annotation of multiple languages in a single treebank needs additional considerations that have not been studied before. Although there has been a few UD treebanks with code-switching (Bhat et al., 2018; Partanen et al., 2018), the papers describing these treebanks do not document or discuss the code-switching aspects of the annotation process.

In this paper we address this gap and outline some of the challenges and interesting phenomena that surface during the annotation of a Turkish–German code-switching treebank. Our contributions are in two levels. The observations on code-switching, independent of the annotation scheme, help in understanding in what forms it occurs. The annotation solutions we propose explore how to handle CS within the UD framework. Working with spoken data brings another aspect and opens also speech annotation under UD to discussion.

2 Related Work

Many well-known linguistic theories on CS syntax, e.g. Free Morpheme and Equivalence Constraints (Poplack, 1980), Closed-class Constraint (Joshi, 1982), Matrix Language Frame (Myers-Scotton, 1993), Functional Head Constraint (Belazi et al., 1994) define their formalism and constraints on constituency structures . Eppler (2005) argues that these constraints are too restrictive from a data-driven perspective and favours Word Grammar (Hudson, 1990), a dependency-based formalism, where the scope of the constraints is head-dependent pairs. Her annotations on German–English transcriptions and the Chinese–English treebank (Wang and Liu, 2013), which also follows Word Grammar, are the only CS dependency treebanks that do not follow UD to the best of our knowledge.

The starting point for our work is the monolingual UD treebanks of both languages in our study. The recent 2.4 release of UD includes three Turkish and four German treebanks. Turkish treebanks include IMST-UD (Sulubacak et al., 2016b), which is semi-automatically converted from the IMST treebank (Sulubacak et al., 2016a) which, in turn, is a re-annotation of the METU-Sabancı treebank (Oflazer et al., 2003). Turkish GB is a manually annotated treebank consisting of grammar book examples (Çöltekin, 2015). There are PUD treebanks consisting of parallel (translated) sentences for both languages. The PUD treebanks were automatically converted from another dependency scheme for the CoNLL 2017 multi-lingual parsing shared task (Zeman et al., 2017). The first German UD treebank is the GSD treebank (McDonald et al., 2013), which is also automatically converted from a different dependency formalism. There are also two new additions to German treebanks; HDT, a conversion of Hamburg Dependency Treebank (Foth et al., 2014; Hennig and Köhn, 2017), and LIT, a treebank of German literary history. Most of our annotation decisions and the discussions below are based on the version 2.3 of the UD treebanks, particularly Turkish IMST and German GSD. There are, however, inconsistencies across languages, and across treebanks of the same language. For most annotation decisions, we follow the annotations in the monolingual treebanks as much of possible. In case of inconsistencies across treebanks, our policy is to choose the alternative closest to the general UD guidelines, so as to ensure cross-lingual consistency within our multilingual treebank.

None of the treebanks noted above include spoken language, let alone code-switching. Quite a few UD treebanks, on the other hand, contain spoken language partially (Danish DDT, Greek GDT, Latvian LVTB, Persian Seraji, Polish LFG, Swedish LinES) or fully (Cantonese HK, Chinese HK, French Spoken, Naija NSC, Norwegian NynorskLIA, Slovenian SST). These treebanks have extended the UD dependency relations with subtypes in addition to using the existing ones to cover linguistic phenomena mainly observed in speech. For example, Slovenian SST (Dobrovoljc and Nivre, 2016) annotates correcting disfluencies either with `reparandum` or `parataxis:restart`. Another `parataxis` subtype, `parataxis:discourse` is defined to cover sentential parentheticals with fixed semantics that serve as discourse elements (e.g., *you know*). French Spoken (Gerdes and Kahane, 2017) and Naija NSC (Courtin et al., 2018) employ the same tag too. They define a separate tag `parataxis:dislocated` for clauses that precede the sentence they are dislocated from. The other relation that is commonly extended is `discourse`. Slovenian SST separates filler sounds from other discourse elements and assigns them `discourse:filler`. Norwegian NynorskLIA (Øvrelid and Hohle, 2016) follows the same approach.

Cantonese HK and Chinese HK (Leung et al., 2016) define `discourse:sp` for sentence particles common in spoken language. So far we are more conservative in extending relations with subtypes and have introduced one that is described in Section 3.2.

The Hindi-English UD treebank (Bhat et al., 2018) annotates the mixed language of social media and has no extension to UD dependencies. The major annotation augmentation is the language IDs assigned to each token. Komi-Zyrian IKDP (Partanen et al., 2018) consists of spoken language, and some utterances include Russian phrases. In those utterances mixed and Russian tokens are marked with respective language IDs, and the Russian syntax is applied. However, the authors do not claim any consistency with the annotations of the monolingual Russian UD treebank. Similar to these treebanks, we also assign a language ID to each token following the tag set in Çetinoğlu (2016). Many other treebanks include words or phrases from a foreign language. Most of them mark foreign tokens with `Foreign=Yes`, and annotate the internal structure of foreign phrases with `flat` relations. However, a few treebanks, e.g., Irish IDT (Lynn and Foster, 2016), annotate foreign tokens according to their respective language.

3 Annotations

Any annotation project is bound to make non-trivial choices (Gerdes and Kahane, 2016). Most non-trivial choices for a code-switching treebank comes either because of the multilingual nature of the resource, or, as noted earlier, the fact that code-switching is prevalent in informal language, and annotation of informal or spoken language has been more challenging than more standard/written language. Most of the problems related to multilingual nature of the data stem from different annotations choices established for individual languages. Although one of the main motivations behind the UD project is multilingual consistency across treebanks, multilingualism within a treebank has not been one of the motivations for UD. Below we focus on issues that arise due to multilingual nature of the treebank, but also noting some of the issues that are due to the informal and spoken language.

3.1 Annotation Differences in Individual Languages

To be able to benefit maximally form monolingual treebanks, one of the principles we follow is to annotate the tokens that belong to each language following the annotation standards in the monolingual treebank(s) of the corresponding language. In many cases this produces a workable solution in a multilingual treebank. In other cases, however, the interaction of tokens within a sentence results in conflicts. In this section we provide a few examples of both cases.

Titles A relatively simple difference between existing monolingual German and Turkish treebanks is the annotation of titles, e.g., as in *President Obama*. The UD guidelines prescribe the use of `flat` relation here. However, the different treebanks follow slightly different practices.[1] German treebanks seem to annotate names using `appos` relation. In Turkish treebanks, similar to a few other treebanks in the UD distribution, the `nmod` relation is used. Although this is a relatively trivial issue, it demonstrates the trade-offs of the annotation choices. On the one hand, choosing one of three relations and applying to both languages would cause inconsistency with the (larger) monolingual treebanks and tools based on these treebanks. On the other hand, following the conventions of both languages causes inconsistency within the multilingual treebank, potentially confusing users querying the treebank, or automatic tools that are trained on it.

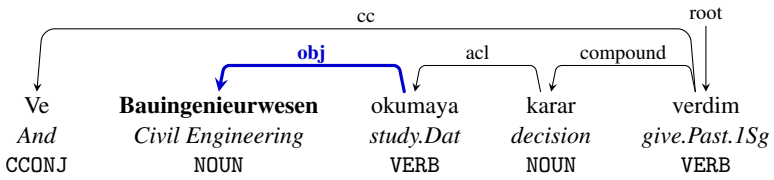
Copula Another similar issue is the annotation of different sort of copula in German. One of the principles of Universal Dependencies is the primacy of the content words. For copular constructions, this means marking the copula as the dependent rather than the head. Since a copula is rarely used in Turkish, the Turkish treebanks naturally follow this for all types of copular constructions. On the other hand, the German GSD and PUD treebanks seem to make distinction where some uses of copula *sein* is annotated as main verb. For example, these treebanks suggest that copula *ist* in *Die Frau ist Ärztin* 'the woman is a doctor' should be annotated using `cop` (with head *Ärztin*), while in *Der Vortrag ist in dem*

[1] See `https://universaldependencies.org/workgroups/mwe.html` for a discussion.

großen Saal ‘The lecture is in the great hall’, it should be marked as the main verb.[2]

Case A particular issue in Turkish–German CS occurs due to different approaches in annotating morphology. Traditionally, morphological annotations in German treebanks are fully disambiguated based on syntax (and possibly larger context) of the sentence. Although not clear-cut, Turkish treebanks annotate only the morphological features that can be inferred from the word form alone. For example, without context, German nouns belonging to some gender classes are ambiguous with respect to their cases. The word *(das) Kind* ‘(the) child’ would be annotated with `Case=Nom` if it is the subject, and with `Case=Acc` if it is the object of the sentence. A similar ambiguity also exists in Turkish. The word *çocuk* ‘child’ may be either the subject or indefinite object of a sentence. However, in both cases it is tagged with `Case=Nom`. The tag `Case=Acc` is only used for definite objects where there is an overt morphological marking for case.

(2)



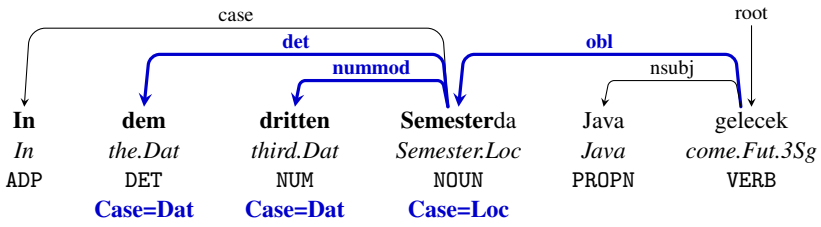
‘And I decided to study Civil Engineering’

(2) presents a sentence involving a German word that functions as an object of a Turkish predicate. According to German annotation standards, the word should be tagged as `Case=Acc`. However, there is no overt case marker,[3] thus the tag should be `Case=Nom` according to Turkish annotation standards. The principle of following the annotation scheme of the token’s language does not work well here, causing the loss of the distinction between definite and indefinite objects in Turkish. In such cases, we chose the language of the head as reference.

3.2 CS-specific Issues

Double case marking Annotating case marking can get more complicated when it is overt in both languages. In (3), the article *dem* ‘the’ and the number *dritten* ‘third’ carry the dative case marking to indicate the static meaning. The noun *Semester* normally does not carry an explicit marker and the German phrase *in dem dritten Semester* ‘in the third semester’ would be completely grammatical. Thus the token *Semester* would normally have the tag `Case=Dat` in its morphological annotation in agreement with its modifiers. However, the speaker has chosen to mark the static meaning *also* in Turkish and following the Turkish grammar rules, there is a locative case marker *-da* attached to the noun, which entails a `Case=Loc` tag in its morphological representation.

(3)



‘Java will start in the third semester.’

The conflict between case markers does not have a perfect solution within the current UD representation. If we choose `Case=Dat` to follow the German rules, the surface form *-da* would not match the morphological tag, furthermore it would change the semantics of the word, as the dative case represents

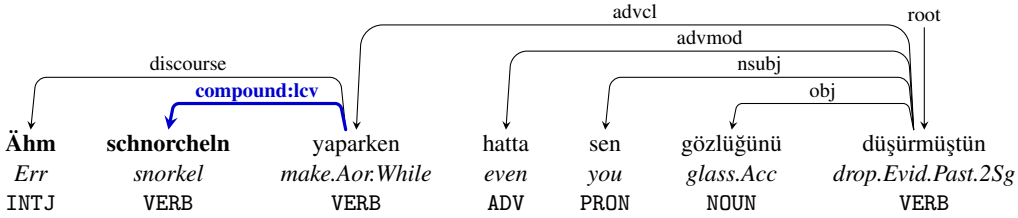
[2]Our design decisions are mainly based on treebanks released in UD version 2.3. As of version 2.4, HDT and LIT treebanks are released for German. While LIT follows GSD and PUD in copula annotation, HDT mark them with `cop`, in accordance with the general and Turkish guidelines. Thus the German copular representation is subject to change.

[3]Note that ***Bauingenieurwesen****i* – the version with the Turkish accusative case marker – would also be grammatical.

motion towards something in Turkish. Thus, we choose the `Case=Loc` tag at the expense of losing the agreement between the determiner and number, and the noun.[4]

Bilingual light verb constructions The use of CS creates new constructions too. One quite common new construction is the use of German verbs followed by a Turkish light verb *etmek* 'do' or *yapmak* 'make', which is also observed in Turkish–German tweets (Çetinoğlu, 2016) as well as Turkish–Dutch (Backus, 2009). The German verb is in infinitive form and the Turkish light verb takes inflectional and derivational suffixes. The core semantics of the construction comes from the German verb. These constructions are similar to noun-light verb constructions common in Turkish (e.g. *yardım etmek* lit.'help do' – 'to help'). In the Turkish UD, noun-verb constructions are labelled with the `compound:lvc` relation where `lvc` denotes light verb constructions. We adopt the same label for German-Turkish constructions. (4) demonstrates a sentence where the German verb *schnorcheln* 'snorkel' is coupled with the Turkish light verb *yap* 'make', that undergoes derivation with the suffix *-ken* 'While'. The combined meaning of the compound is 'while snorkelling'.

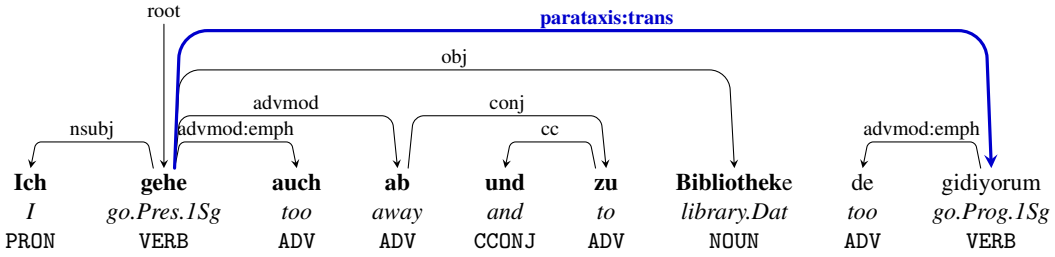
(4)



'While you were snorkelling you had even dropped your glass.'

Translation pairs Another CS-specific language use we have observed is uttering a word, phrase or clause in one language and repeating it as a translation in the other language. (5) shows such an example where German *gehe auch* 'I go too' is repeated again as Turkish *de gidiyorum* Since there are no relations in UD that would capture this phenomenon, we extend the relation `parataxis` by introducing a `trans` subtype. The relation connects the head of the second constituent to the head of the first constituent as a dependent.

(5)



'I also go to the library now and then.'

Bilingual *m*-reduplication In Turkish, it is possible to generalise the meaning of a word by so-called *m*-reduplication (Göksel and Kerslake, 2005). To realise *m*-reduplication, the first word is reduplicated, and an *m* prefixes the duplicate if the word starts with a vowel, or the first character of the duplicate is replaced with an *m* if it is a consonant as in (6).

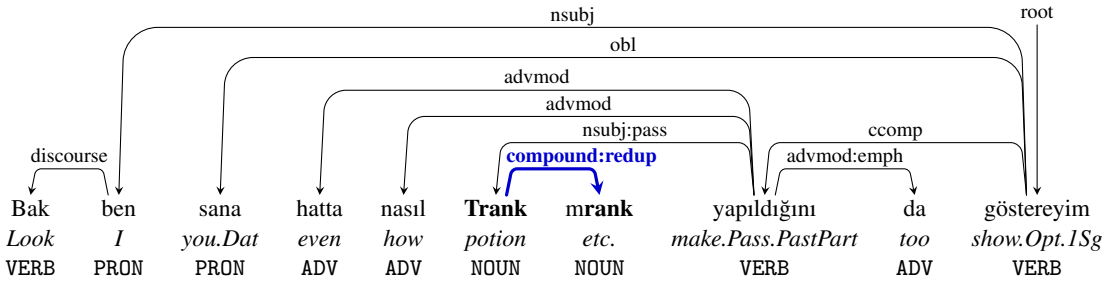
(6) Çay **m**ay içer misin?
Tea etc. drink.Aor Ques.2Sg
'Would you like to drink tea and the like?'

While this is a Turkish-specific phenomenon, bilinguals also apply it to other languages. In (7) we see that the German word *Trank* 'potion' undergoes *m*-reduplication. This is not only a new lexical

[4] Another possibility is indicating both cases with notation `Case=Dat,Loc`. This is used when the word may have one of the values, but it cannot be decided from the available context. In this particular case, however, there is no ambiguity. Both case values are correct depending on the language.

alternation in German, its syntactic representation is new to German UD as well. *m*-reduplications are represented as `compound:redup` in the Turkish UD treebanks; we apply it also to German in this case.

(7)



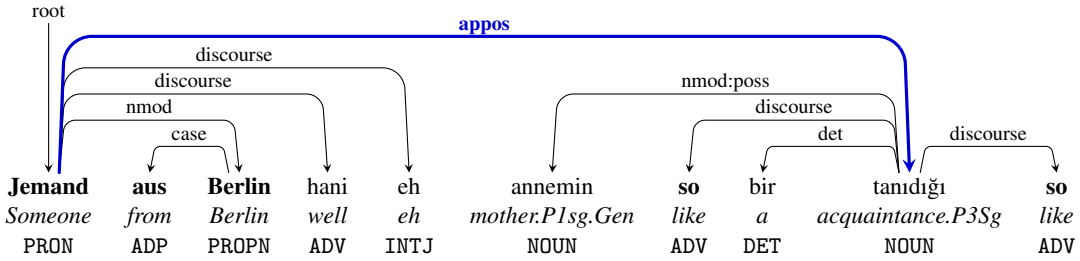
'Look, let me even show you also how potion et cetera is made.'

3.3 Issues Related to Spoken Language

We also observe some linguistic phenomena more frequently than corresponding monolingual treebanks due to the medium we collect the data. Spoken language contains many disfluencies, repetitions, run-on sentences, and uncommon word order. Since these phenomena are orthogonal to mixing languages, their dependencies can cross language boundaries within a sentence. We exemplify two of the commonly observed cases.

Appositions In appositions two consecutive noun phrases define the same referent in different ways. In our corpus these two noun phrases could as well be in different languages. In (8) the speaker mentions 'someone from Berlin' in Turkish then refers to the same person with additional information 'an acquaintance of my mother' in German. Following the UD guidelines, the head of the second phrase is dependent on the head of the first phrase with the relation `appos`.

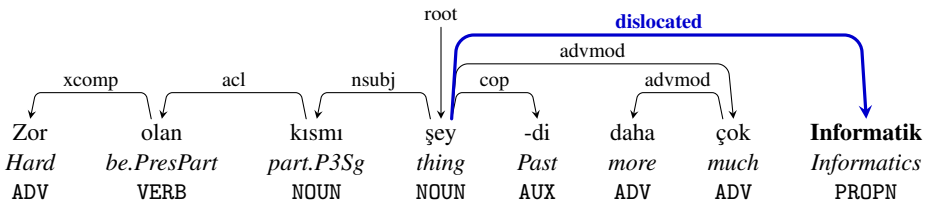
(8)



'Someone from Berlin, well, an acquaintance of my mother.'

Dislocation In spoken Turkish it is quite common to replace a word or phrase that does not come to mind immediately or inappropriate to say with the word *şey* 'thing'. While it is a noun itself, it can also replace verbs or clauses when combined with the light verb *etmek* 'do'. The CS corpus we are collecting has many instances of such use, (9) demonstrates one case.

(9)



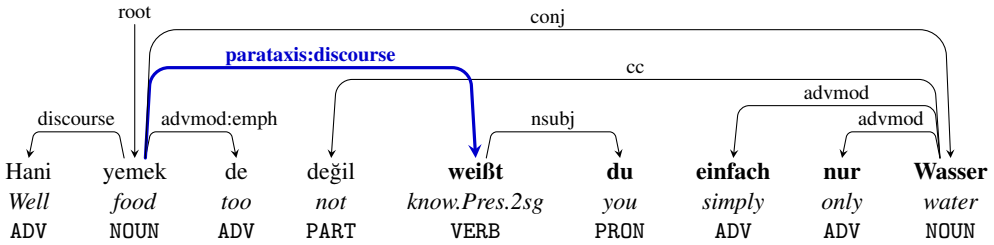
'The hard part of it was mostly this thing, Informatics.'

The speaker first uses *şey* as the nominal predicate of the copular sentence. This way the sentence is grammatically complete with the placeholder *şey* until the last word. Once the word *Informatik* 'Informatics' is uttered, it does not have a role in the sentence other than clarifying *şey*. UD employs the

dislocated tag for these relations. By definition the dislocated item is attached to the head of the placeholder. Here, the head is the placeholder itself, thus *Informatik* is dependent on *şey*.

Clausal discourse elements Spoken language contains many clauses with fixed semantics that function as discourse markers such as *you know*, *say*, *I think*. We observe similar cases in our corpus too; most frequent examples include German *weißt du* 'you know', *ich glaube* 'I think', and Turkish *bak* 'look'. The UD policy for such cases is connecting them to the main clause with a parataxis tag. Some of the UD spoken treebanks (Dobrovoljc and Nivre, 2016; Gerdes and Kahane, 2017; Courtin et al., 2018) keep the discourse information via the subtype parataxis:discourse. We follow their approach and employ the same tag as exemplified in (10) with *weißt du* 'you know'.

(10)



'Well, it is not food, you know, just water.'

4 Conclusions

In this paper we present our experience with an ongoing treebank creation project of a Turkish-German code-switching corpus. In annotations, we follow the general UD guidelines and, Turkish and German UD treebanks as much as differences in individual languages allow. When we encounter new monolingual or bilingual syntactic constructions we apply existing relations to these new conditions; and if not sufficient, we introduce a subtype. Due to annotating spoken data, our sentences contain dependencies that are rare or nonexistent in monolingual Turkish and German treebanks. For those cases also, we follow general UD guidelines and other spoken UD treebanks.

Our observations so far suggest that interesting phenomena we come across and challenges they bring can only increase as we continue to collect and annotate more data. For some of the challenges we propose well-fitting solutions. For others, we take advantage of reporting work in progress and open our decisions up for discussion. Thus we see this paper as an opportunity to share idiosyncrasies of code-switching with any researcher who is interested in CS in particular, or in non-canonical language in general; and to exchange annotation ideas with the UD community.

Acknowledgements

We thank Cansu Turgut and Sevde Ceylan for data collection and annotation, and for discussions on the semantics of examples. We also thank the reviewers for their helpful comments. The first author is funded by DFG via Project CE 326/1-1 "Computational Structural Analysis of German-Turkish Code-Switching".

References

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

A. Backus, 2009. *Codeswitching as one piece of the puzzle of language change: The case of Turkish yapmak*, pages 307–336. Number 41 in Studies in Bilingualism. John Benjamins. Pagination: 20.

Hedi M Belazi, Edward J Rubin, and Almeida Jacqueline Toribio. 1994. Code switching and x-bar theory: The functional head constraint. *Linguistic inquiry*, pages 221–237.

Irshad Bhat, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2018. Universal Dependency parsing for Hindi-English code-switching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 987–998, New Orleans, Louisiana, June. 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.

Marine Courtin, Bernard Caron, Kim Gerdes, and Sylvain Kahane. 2018. Establishing a language by annotating a corpus: the case of Naija, a post-creole spoken in Nigeria. In *Proceedings of the Workshop on Annotation in Digital Humanities*, pages 7–11, August.

Kaja Dobrovoljc and Joakim Nivre. 2016. The Universal Dependencies treebank of spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, may.

Eva Maria Eppler. 2005. *The syntax of German–English code-switching.* Ph.D. thesis, University of London.

Kilian A. Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. Because size does matter: The Hamburg dependency treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2326–2333, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Kim Gerdes and Sylvain Kahane. 2016. Dependency annotation choices: Assessing theoretical and practical issues of Universal Dependencies. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 131–140, Berlin, Germany, August. Association for Computational Linguistics.

Kim Gerdes and Sylvain Kahane. 2017. Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral : le cas de la macrosyntaxe. In *Atelier sur les corpus annotés du français (ACor4French)*.

Aslı Göksel and Celia Kerslake. 2005. *Turkish: A comprehensive grammar*. Routledge.

Felix Hennig and Arne Köhn. 2017. Dependency tree transformation with tree transducers. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 58–66, Gothenburg, Sweden, May. Association for Computational Linguistics.

Richard A. Hudson. 1990. *English Word Grammar*. Oxford:Blackwell.

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.

Herman Leung, Rafaël Poiret, Tak-sum Wong, Xinying Chen, Kim Gerdes, and John Lee. 2016. Developing Universal Dependencies for Mandarin Chinese. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 20–29, Osaka, Japan, December.

Teresa Lynn and Jennifer Foster. 2016. Universal Dependencies for Irish. In *Celtic Language Technology Workshop*, pages 79–92.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

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

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)*, pages 1659–1666.

Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. 2003. Building a Turkish treebank. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, chapter 15, pages 261–277. Springer.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian Universal Dependencies treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 126–132, Brussels, Belgium, November. Association for Computational Linguistics.

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.

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

Adithya Pratapa, Monojit Choudhury, and Sunayana Sitaram. 2018. Word embeddings for code-mixed language processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3067–3072, Brussels, Belgium, October-November. Association for Computational Linguistics.

Umut Sulubacak, Gülşen Eryiğit, Tuğba Pamay, et al. 2016a. IMST: A revisited Turkish dependency treebank. In *Proceedings of TurCLing 2016, the 1st International Conference on Turkic Computational Linguistics*. EGE UNIVERSITY PRESS.

Umut Sulubacak, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016b. Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454, Osaka, Japan.

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

Lin Wang and Haitao Liu. 2013. Syntactic variations in Chinese–English code-switching. *Lingua*, 123:58–73.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.

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, October-November. Association for Computational Linguistics.

Lilja Øvrelid and Petter Hohle. 2016. Universal Dependencies for Norwegian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, may.