Regression Analysis of Lexical and Morpho-Syntactic Properties of Kiezdeutsch

Diego Frassinelli[1], **Gabriella Lapesa**[2],
Reem Alatrash[2], **Dominik Schlechtweg**[2], **Sabine Schulte im Walde**[2]
[1]Department of Linguistics, University of Konstanz
[2]Institute for Natural Language Processing, University of Stuttgart
[1]diego.frassinelli@uni-konstanz.de
[2]name.surname@ims.uni-stuttgart.de

Abstract

Kiezdeutsch is a variety of German predominantly spoken by teenagers from multi-ethnic urban neighborhoods in casual conversations with their peers. In recent years, the popularity of Kiezdeutsch has increased among young people, independently of their socio-economic origin, and has spread in social media, too. While previous studies have extensively investigated this language variety from a linguistic and qualitative perspective, not much has been done from a quantitative point of view.

We perform the first large-scale data-driven analysis of the lexical and morpho-syntactic properties of Kiezdeutsch in comparison with standard German. At the level of results, we confirm predictions of previous qualitative analyses and integrate them with further observations on specific linguistic phenomena such as slang and self-centered speaker attitude. At the methodological level, we provide logistic regression as a framework to perform bottom-up feature selection in order to quantify differences across language varieties.

1 Introduction

Over the past 50 years, Europe has seen a substantial increase in the number of immigrants and in the diversity of their origin. A direct consequence of this situation is the rise of the so-called "Urban Youth Languages" (Wiese, 2017): specific linguistic practices used by young people in multi-ethnic urban areas. One example of urban youth languages is Kiezdeutsch ('hood German'), which is a linguistic variety of German spoken primarily by teenagers from multi-ethnic urban neighborhoods in casual conversations with their peers. Kiezdeutsch first appeared over 30 years ago and has since then developed systematic linguistic structures that identify it as an independent variety of German (Wiese et al., 2009).

Recent studies have shown that the stylistic elements of Kiezdeutsch have spread in the repertoires of many young German speakers without immigrant background (Freywald et al., 2011; Stevenson et al., 2017). At the syntactic level, the main differences with standard German are (see examples below): bare noun phrases (1) lacking determiners or (2) lacking prepositions; (3) lack of copula verbs; (4) verb-first declaratives; and (5) subject-verb-object (SVO) word order in sentences beginning with an adverb.

1. Hast du Problem? (vs. Hast du **ein** Problem?)
   Have you problem? (Do you have a problem?)

2. Ich geh Kino. (vs. Ich geh**e ins** Kino.)
   I go cinema. (I go to the cinema.)
   (Wiese and Pohle, 2016)

3. Er aus Kreuzberg. (vs. Er **ist** aus Kreuzberg.)
   He from Kreuzberg. (He is from Kreuzberg.)

4. Wollte ich keine Hektik machen da drinne. (vs. **Ich wollte** keine Hektik machen da drinne.)
   Wanted I no hectic make there inside.
   (I didn't want to make any hectic in there.)

5. Jetzt ich bin 18. (vs. Jetzt **bin ich** 18.)
   Now I am 18. (Now, I am 18.)

In previous work, researchers have studied various linguistic aspects of Kiezdeutsch focusing on either qualitative analyses (Tertilt, 1996; Auer, 2003; Keim and Knöbl, 2011; Wiese et al., 2009; Wiese, 2012, 2013; te Velde, 2017; Preseau, 2018) or small-scale quantitative analyses (Fuchs et al., 2010; Jannedy, 2010; Wiese and Pohle, 2016).

Proceedings of the 8th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 21–27

In this work we suggest logistic regression as a general framework to perform bottom-up data-driven feature selection in order to quantify differences across language varieties. We use Kiezdeutsch as a test case by comparing against standard German, and we deliberately select simple lexical and morpho-syntactic features that can easily be obtained from standard part-of-speech (POS) taggers and lemmatisers. In this vein, we present three studies: morpho-syntactic variation in terms of part-of-speech unigram distributions (Study 1) and in terms of part-of-speech trigram distributions (Study 2); and lexical variation in the usage of nouns and verbs (Study 3).

2 Previous Work on Kiezdeutsch

In the mid-1990s the release of two books sparked interest in migrant varieties of German in Germany, a collection of semi-fictitious interviews with young men from Turkish backgrounds living in Berlin (Zaimoglu, 1995), and a documentation of a Turkish youth gang regarding their daily activities (Tertilt, 1996). Although these ethnographic-centered analyses were not mainly concerned with language change or variation, they brought to light the notion of groups of young people living in the urban centres of Germany who had developed their own language practices as part of their identity.

With the turn of the century, more language-centered studies of this urban vernacular began to appear. In an effort to gain a holistic understanding of Kiezdeutsch, Androutsopoulos (1998a,b, 2001) used media text analyses, ethnographic observations and interviews to analyze the speech style of teenagers speaking Kiezdeutsch. He then compiled a list of language features on the phonological/phonetic, lexical, and grammatical levels. Moreover, Androutsopoulos studied the various socio-cultural aspects of Kiezdeutsch and their effects on the German language. Similarly, Auer (2003) and Keim and Knöbl (2011) identified features of Kiezdeutsch through analyses of speech sequences which were then linked to their social interactions and functions as well as their discourses, in order to asses the social and linguistic effects of these features on the German language. Their research suggested that Kiezdeutsch speakers exhibited a high level of linguistic proficiency and communicative competence, thus contradicting previous views that grammatical simplifications were due to deficiency in language acquisition. These conclusions were in agreement with studies by Freywald et al. (2011) who considered Kiezdeutsch a multi-ethnolect, and by Wiese (2013) who categorized Kiezdeutsch as an urban dialect.

The introduction of a corpus of spoken Kiezdeutsch (Rehbein et al., 2014) led to research across linguistic levels. For example, te Velde (2017) investigated phonological form using syntax of verb-second constructions found in the German dialects Kiezdeutsch, Yiddish, Bavarian, Cimbrian, and colloquial German. More recently, the effects of English on Kiezdeutsch constructions were examined by Preseau (2018) who argued that it was necessary to reconsider Kiezdeutsch as a native dialect of German, given the role of English as a Lingua Franca (ELF) in urban Germany and the effect it has on Kiezdeutsch-speaking communities.

The above-mentioned studies illustrate the broad spectrum of qualitative evidence and analyses on Kiezdeutsch. On the other hand, up to date only a few studies have provided quantitative evidence on Kiezdeutsch. One such study was conducted by Fuchs et al. (2010) who used a Gaussian mixture model to explore the durational properties of the particle *so* in various prosodic positions within utterances of Kiezdeutsch speech. The authors supported their findings by predicting distinctions (e.g., utterance-final and phrase-final) from text using punctuation as marker. The work by Fuchs and her colleagues examined a single test case of a very specific phonological phenomenon using audio signals as the main source of information. Another study by Jannedy (2010) complemented the work by Fuchs and her colleagues by investigating the usage patterns of the particle *so* using a contingency table and χ^2 tests. The usage-patterns of the same particle were also analyzed quantitatively by Wiese (2012) using χ^2 tests. While being close to our studies, the contributions by Jannedy and Wiese focused on a single test case while we focus on large-scale analyses. Moreover, the studies by Fuchs et al. and Jannedy used a closed-access corpus of speech obtained from interviews with teenagers who speak Kiezdeutsch.

The work by Wiese and Rehbein (2016) combined qualitative and quantitative analyses of several well-established phenomena in Kiezdeutsch with the aim of demonstrating the linguistic coherence of this urban vernacular. In their top-down approach, Wiese and Rehbein started with predefined linguistic patterns and then performed a χ^2 test on

the raw corpus frequencies of these patterns. The results pointed to systematic differences between data from the sub-corpus of multi-ethnic speakers and the sub-corpus of mono-ethnic speakers of German. In contrast, our study takes a bottom-up approach and does not define features a priori, but instead allows them to emerge in a data-driven fashion. Furthermore, we operate on a large scale and consider all patterns that emerge.

To our knowledge, there have been no studies on Kiezdeutsch which employ logistic regression models to gather large-scale evidence regarding the various claims and theories in the literature.

3 Materials

In our study, we use two transcribed German spoken corpora: the KiDKo corpus containing dialogues in Kiezdeutsch, and the GRAIN corpus containing radio interviews in standard German.

KiDKo The KiezDeutsch Korpus (KiDKo, Rehbein et al. (2014)) is a collection of casual everyday conversations between teenagers (14-17yo) from multi-ethnic and mono-ethnic communities in Berlin. The collection took place from 2008 until 2015 using self-recordings in the absence of adults and non-members of their social group. In order to capture the most salient emerging properties in such a dynamic language variety, in our studies we focus on the multi-ethnic sub-corpus (Rehbein and Schalowski, 2013). In total this part of the corpus contains the transcription of 43 hours of conversations with a total of 63,604 sentences (359,000 normalised tokens). Part-of-speech tagging has been performed with a tagger developed specifically for KiezDeutsch by Rehbein et al. (2014), based on a version of the Stuttgart-Tübingen tagset (STTS) augmented by 11 additional tags tailored to spoken German.

GRAIN The German RAdio INterviews corpus (GRAIN, Schweitzer et al. (2018)) is a collection of interviews broadcast on the German public radio. The hosts from the radio interviews are professionals talking about social and political topics (e.g., a chairman of a council talking about city pollution). In total, 14,097 sentences (221,000 tokens) have been extracted from 23 hours of recordings. The materials have been automatically pos-tagged with the Tree Tagger (Schmid, 1994) and according to the STTS tagset.

KidKo vs. GRAIN GRAIN was selected as the corpus representative of standard (spoken) German because among the available spoken German corpora it is the most comparable to the KiDKo for its size and its collection time frame (see above). Both corpora contain transcriptions of recorded speech, however the dialogues in KiDKo are spontaneous conversations about everyday topics, whereas the dialogues in GRAIN are more controlled in their content and setting. Moreover, speakers of Kiezdeutsch are teenage students, while the speakers in the GRAIN corpus are adults holding professional roles.

With respect to size, both corpora are relatively small, with KiDKo being one third bigger than GRAIN. The sentence length of the two corpora is extremely different: the average sentence length in KidKo (8.8 tokens/sentence) is much shorter than the one in GRAIN (26.7 tokens/sentence). As basis for comparison, we extracted the same number of n-grams (unigrams and trigrams) from GRAIN and KiDKo using a stratified sampling algorithm (Levy and Lemeshow, 2013). In this way, we created a basis for lexical and morpho-syntactic analyses on the individual token level and on a token-sequence level, while maintaining unchanged the underlying distribution of the respective n-grams from the original corpora.

4 Logistic Regression Analyses

To identify the most distinctive features in the two corpora we use logistic regression models. In all the models reported below, we predict the categorical variable `corpus_type` (KidKo vs. GRAIN) using as predictor the presence/absence of one feature at a time. Running a unique model including all the lexicalised features would lead to convergence issues; for this reason, we do not use a bag-of-feature classifier approach consistently in all three studies.

After fitting the model, we take the z-score corresponding to the predicted variable. In logistic regression, a z-score is the ratio of the coefficient estimate divided by its standard error. The larger the z-score, the less uncertain the prediction is and, consequently, the stronger the difference between the feature in the two corpora. Compared to frequency analysis or more traditional estimates like χ^2, the analysis of the z-scores conveys richer information: the sign of the z-score indicates the direction of the effect if the feature is more predic-

tive of the KidKo (positive sign) or of the GRAIN corpus (negative sign); moreover, its absolute value is directly related to the level of uncertainty involved in the prediction: larger numbers indicate more reliable predictions.

For this reason, we systematically look at the largest positive and negative z-scores in the analyses as the most informative ones. In such way, we filter out features that show a comparable distribution in both corpora and consequently have no discriminative power. Finally, by looking at the z-score values it is possible to detect if the probability of selecting one of the two corpora is significantly different from zero (i.e., p-value < 0.001). In order to reduce type I errors (false positives due to chance), we correct the alpha values by dividing our significance threshold (0.001) by the total number of models we run.

5 Studies and Results

Study 1: Unigram POS Analysis The aim of this first study is to compare the unigram POS distributions in the two corpora. We run 10 logistic regression models predicting corpus type given the presence/absence of each POS (such as NOUN).

Given the extreme granularity of the POS types in the original collections, we decided to use a coarse-grained classification of 10 POS types only, encoding the word class but not the inflectional categories: nouns (NOUN), pronouns (PRON), verbs (VERB), adverbs (ADV), adpositions (ADP), determiners (DET), conjunctions (CONJ), adjectives (ADJ), particles (PRT), and numerals (NUM).[1] Besides the fact that it alleviates sparsity, such coarse-grained approach allows us to uncover differences between two corpora that are systematic across classes and go beyond the idiosyncratic use of extremely corpus-specific tags.

Table 1 reports the z-scores associated to each POS. Across the ten POS tags under analysis, five are significantly more predictive of GRAIN (negative values) and five of KidKo (positive values). These results are in line with previous qualitative studies (Wiese and Pohle, 2016): determiners are used significantly less in Kiezdeutsch compared to standard German (see Example (2)); similarly, adpositions (mainly prepositions) are much less used by teenagers than adults (see Example (1)).

[1]For the full set of STTS part-of-speech tags, see https://www.ims.uni-stuttgart.de/en/research/resources/lexica/germantagsets/.

GRAIN		**KidKo**	
POS	**z-sc.**	**POS**	**z-sc.**
DET	-54.27	PRT	59.32
NOUN	-44.10	PRON	37.50
ADP	-38.45	ADV	22.10
CONJ	-18.17	VERB	21.92
ADJ	-11.67	NUM	7.30

Table 1: Distribution of z-scores for each coarse-grained POS for standard German (negative) vs. Kiezdeutsch (positive).

Study 2: Trigram POS Analysis In this study we analyse the distribution of trigrams of consequent POS (e.g., DET+ADJ+NOUN) that we extracted from each sentence in the two corpora. In this way we approach syntactic structural differences in the language varieties, while still relying on simple POS information. In total we have 1,245 trigram types and, consequently, we run 1,245 logistic regression models where we predict corpus type using each of those trigrams as binary predictors (0/absence vs. 1/presence of each trigram). Significant level is reached when the z-score is larger than $\pm$ 3.2 (p-value < 0.0008).

Table 2 lists the most predictive trigrams for GRAIN (left) and for KidKo (right). Overall, 178 trigrams are highly significant for GRAIN and 181 for KidKo. Three trigrams of POS have an extremely strong predictive power for Kiezdeutsch: PRON+VERB+PRON, PRON+VERB+ADV, and VERB+PRON+ADV. In line with the evidence from Study 1, we see how trigrams of POS involving verbs and pronouns predominate in KidKo, while nouns and determiners are more predictive of GRAIN. The clear preference for pronouns in Kiezdeutsch, as opposed to nouns, can be explained by the topics of spontaneous speech being much more related to conversations involving self-reference and reference to further actors present in the scene. Corpus examples for the three most predictive KidKo trigrams in this respect are *ich habe deine* 'I have yours' (PRON+VERB+PRON); *wir reden hier* 'we talk here' (PRON+VERB+ADV); and *machen wir jetzt* 'do we now' (VERB+PRON+ADV). Nouns, on the other side, are essential when referring to events far from the proximity of the speech act, as in political interviews, e.g., *Gestaltung des Lebens* 'shaping of life' (NOUN+DET+NOUN); *ein Einsatz in* 'a mission in' (DET+ADP+NOUN); and *Menschen in Sorge* 'humans in fear' (NOUN+ADP+NOUN).

GRAIN		KidKo	
POS	**z-score**	**POS**	**z-score**
NOUN+DET+NOUN	-25.08	PRON+VERB+PRON	37.79
DET+NOUN+ADP	-23.62	PRON+VERB+ADV	33.92
NOUN+ADP+NOUN	-22.99	VERB+PRON+ADV	29.47
NOUN+ADP+DET	-22.96	VERB+PRON+PRON	21.79
DET+ADJ+NOUN	-22.71	PRON+PRON+VERB	19.78
ADP+DET+NOUN	-21.67	VERB+ADV+ADV	19.54
DET+NOUN+DET	-19.07	VERB+PRON+PRT	19.39
ADJ+NOUN+VERB	-18.33	PRON+VERB+PRT	19.25
ADP+DET+ADJ	-17.88	VERB+ADV+PRT	18.03
ADJ+NOUN+ADP	-17.64	PRON+VERB+ADJ	17.21

Table 2: Distribution of z-scores of the most predictive POS trigrams for GRAIN (left) vs. KidKo (right).

Figure 1 shows the distributions of POS trigrams sorted by their relative frequencies. As we can see, even though both lines follow a Zipfian distribution, in KidKo there are three trigrams that are much more frequent than all the rest; we also see a longer tail indicating a higher number of trigrams occurring only once. Moreover, if we look at mid-frequent trigrams (Figure 1), we see how the slope in the distribution from KidKo is much steeper than the one from GRAIN. KidKo thus shows a higher number of extremely frequent and extremely rare trigrams indicating the more idiosyncratic nature of Kiezdeutsch. On the other hand, in GRAIN we can find more mid-frequency trigrams indicating the more standardised nature of the variety of German used in this corpus.

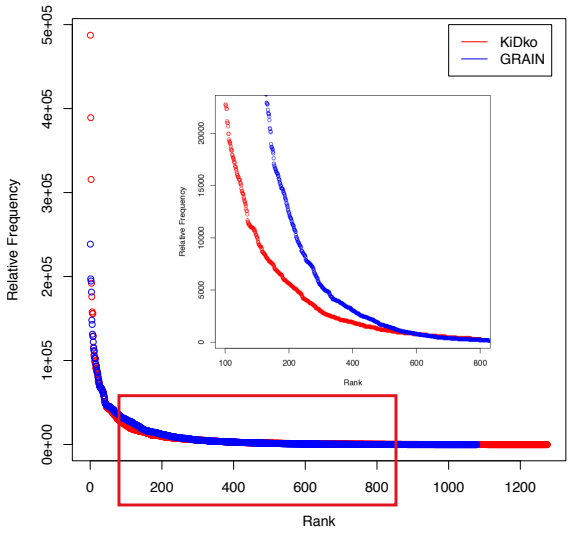

Figure 1: Overall distributions of POS trigrams sorted by relative frequency, accompanied by zoom into mid-values frequency scores.

Study 3: Noun and Verb Distributions In this final study we analyse the distributions of nouns and verbs in the two corpora (Tables 3 and 4). When looking at the most predictive verbs for each corpus, we find that GRAIN contains verbs which are part of more complex structures (such as modal structures requiring the presence of an infinitive form), and more formal nouns.

On the other hand, in KidKo verbs of needing, having, existence and obligation are the most predictive ones, together with nouns referring to typical topics for young people (school, home, fun, games). Important to highlight is the extreme predictive nature of *Alter* as the most frequent form of addressing among members of the younger generation: what the regression model has picked up here is a clear case of slang.

Once more, such distributions highlight the more self-centered type of conversation among teenagers with topics related to everyday life events that affect them, thus directly showing the simplified nature of KidKo. On the other hand, GRAIN shows the usage of more formal and detached forms (human, question, topic).

6 Conclusion

The aim of this study was to introduce logistic regression as a general framework to perform bottom-up data-driven feature selection in order to quantify differences across language varieties. We applied the framework to Kiezdeutsch in comparison to standard German as a test case, which allowed us to identify significant differences at the level of part-of-speech, part-of-speech sequences, as well as lexical choices.

GRAIN		z-score	KidKo		z-score
Menschen	'humans'	-8.23	*Alter*	'age'	11.19
Frage	'question'	-6.23	*Schule*	'school'	8.11
Thema	'topic'	-6.20	*Euro*	'euro'	7.85
Land	'country'	-6.02	*Stunden*	'hours'	7.33
Herr	'Mr.'	-5.99	*Spiel*	'game'	7.23
Prozent	'percent'	-5.65	*Hause*	'home'	7.17
Europa	'Europe'	-5.31	*Ahnung*	'idea'	6.88
Jahren	'years'	-4.70	*Spaß*	'fun'	6.72
Gesellschaft	'society'	-4.29	*Minuten*	'minutes'	6.34
Ende	'end'	-4.07	*Mal*	'times'	6.20

Table 3: The 10 most predictive nouns in GRAIN (left) vs. KidKo (right) with the corresponding z-scores.

GRAIN		z-score	KidKo		z-score
habe	'have'	20.13	*werden*	'will be'	-16.81
war	'was'	11.72	*haben*	'have'	-13.68
weiß	'know'	10.84	*wird*	'will'	-12.61
gesehen	'seen'	6.58	*sind*	'are'	-12.35
warte	'wait'	6.27	*müssen*	'must'	-12.27
mache	'make'	6.18	*gibt*	'give'	-11.34
gesagt	'said'	6.18	*können*	'can'	-9.30
bin	'am'	6.09	*wollen*	'want'	-8.69
mach	'make'	6.05	*brauchen*	'need'	-7.08
gemacht	'made'	5.90	*sagen*	'say'	-7.03

Table 4: The 10 most predictive verbs in GRAIN (left) vs. KidKo (right) with the corresponding z-scores.

Our results show consistent trends: on the one hand, we confirm the predictions drawn from the theoretical literature; on the other hand, our automatic bottom-up process results in a multi-faceted set of observations including slang, specific topics and reporting attitudes. Our studies thus confirm our framework as a useful tool to detect and quantify language variation properties, while relying on simple and easy-to-obtain lexical and morpho-syntactic features.

Current work targets both the scope of the experiments and the methodological investigation. We are further experimenting with the introduction of semi-lexicalised patterns (e.g., PRON+VERB+*Kino*, 'cinema' vs. PRON+VERB+*Schule*, 'school') in the regression to investigate whether specific syntactic patterns are more salient in certain domains, e.g., leisure vs. non-leisure activities. Methodologically, we plan to support our insights with a thorough comparison with other feature selection methodologies, such as random forests Tagliamonte and Baayen (2012).

The relevance of the framework is not limited to the sociolinguistic issues we address: it also proposes a robust strategy to select distinctive features and to demonstrate their use in concrete feature selection settings. Kiezdeutsch is a spoken variety, but we expect its most salient features to emerge also in less controlled varieties of written language, posing a significant challenge to NLP tools developed for standard German. From this perspective, our work has a straightforward application in social-media use-cases, for example in the detection and handling of German/Kiezdeutsch code-switching on Twitter and forums.

Acknowledgments

We thank the three anonymous reviewers for the very detailed and to-the-point comments, and acknowledge funding from the following institutions: G. Lapesa (German Ministry of Education and Research, project E-DELIB); R. Alatrash (CRETA center funded by the German Ministry for Education and Research); D. Schlechtweg (CRETA center and Konrad Adenauer Foundation).

References

Jannis Androutsopoulos. 1998a. *Deutsche Jugendsprache: Untersuchungen zu ihren Strukturen und Funktionen*. Peter Lang, Frankfurt am Main.

Jannis Androutsopoulos. 1998b. Forschungsperspektiven auf Jugendsprache: Ein integrativer Überblick. In Jannis Androutsopoulos and Arno Scholz, editors, *Jugendsprache – Langue des Jeunes – Young People's Language*. Peter Lang, Frankfurt am Main.

Jannis Androutsopoulos. 2001. Ultra korregd Alder! Zur medialen Stilisierung und Aneignung von 'Türkendeutsch'. *Deutsche Sprache*, 29:321–339.

Peter Auer. 2003. "Türkenslang": Ein jugendsprachlicher Ethnolekt des Deutschen und seine Transformationen. In Annelies Häcki-Buhofer, editor, *Spracherwerb und Lebensalter*, pages 255–264. Tübingen: Francke.

Ulrike Freywald, Katharina Mayr, Tiner Özçelik, and Heike Wiese. 2011. Kiezdeutsch as a multiethnolect. *Ethnic Styles of Speaking in European Metropolitan Areas*, pages 45–73.

Susanne Fuchs, Jelena Krivokapic, and Stefanie Jannedy. 2010. Prosodic boundaries in German: Final lengthening in spontaneous speech. *Journal of the Acoustical Society of America*, 127(3):1851–1851.

Stefanie Jannedy. 2010. The usage and distribution of "so" in spontaneous Berlin Kiezdeutsch. *ZASPiL Papers from the Linguistics Laboratory*, 43(52).

Inken Keim and Ralf Knöbl. 2011. Linguistic variation and linguistic virtuosity of young "ghetto"-migrants in Mannheim. In Friederike Kern and Margret Selting, editors, *Ethnic Styles of Speaking in European Metropolitan Areas*, pages 239–264. Amsterdam: Benjamins.

Paul S. Levy and Stanley Lemeshow. 2013. *Sampling of populations: Methods and applications*. John Wiley & Sons.

Lindsay Preseau. 2018. *Kiezdeutsch, Kiezenglish: English in German Multilingual/-ethnic Speech Communities*. Ph.D. thesis, UC Berkeley.

Ines Rehbein and Sören Schalowski. 2013. STTS goes Kiez–Experiments on annotating and tagging urban youth language. *Journal for Language Technology and Computational Linguistics*, 28(1).

Ines Rehbein, Sören Schalowski, and Heike Wiese. 2014. The KiezDeutsch Korpus (KiDKo) Release 1.0.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Katrin Schweitzer, Kerstin Eckart, Markus Gärtner, Agnieszka Falenska, Arndt Riester, Ina Rösiger, Antje Schweitzer, Sabrina Stehwien, and Jonas Kuhn. 2018. German radio interviews: The GRAIN release of the SFB732 Silver Standard Collection. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Patrick Stevenson, Kristine Horner, Nils Langer, and Gertrud Reershemius. 2017. *The German-speaking world: A practical introduction to sociolinguistic issues*. Routledge.

Sali A. Tagliamonte and R. Harald Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, 24(2):135–178.

Hermann Tertilt. 1996. *Turkish Power Boys. Ethnographie einer Jugendbande*. Suhrkamp, Frankfurt am Main.

John R. te Velde. 2017. German V2 and the PF-interface: Evidence from dialects. *Journal of Germanic Linguistics*, 29(2):147–194.

Heike Wiese. 2012. *Kiezdeutsch: Ein neuer Dialekt entsteht*. C.H. Beck.

Heike Wiese. 2013. What can new urban dialects tell us about internal language dynamics? The power of language diversity. *Linguistische Berichte*, 19:208–245.

Heike Wiese. 2017. Urban contact dialects. In Salikoko S. Mufwene and Anna Maria Escobar, editors, *Cambridge Handbook of Language Contact*. Cambridge: Cambridge University Press.

Heike Wiese, Ulrike Freywald, and Katharina Mayr. 2009. Kiezdeutsch as a test case for the interaction between grammar and information structure. *Interdisciplinary Studies on Information Structure. Working Papers of the SFB 632*, 12.

Heike Wiese and Maria Pohle. 2016. "Ich geh Kino" oder "... ins Kino"? *Zeitschrift für Sprachwissenschaft*, 35(2):171–216.

Heike Wiese and Ines Rehbein. 2016. Coherence in new urban dialects: A case study. *Lingua*, 172:45–61.

Feridun Zaimoglu. 1995. *Kanak Sprak*. Rotbuch Verlag, Berlin.