Morpho-Syntactic Properties of Kiezdeutsch: Lexicalized POS Regression Analyses

Diego Frassinelli (University of Konstanz); Gabriella Lapesa, Reem Alatrash,
Dominik Schlechtweg, Sabine Schulte im Walde (University of Stuttgart)
diego.frassinelli@uni-konstanz.de

Kiezdeutsch is a variety of German predominantly spoken by teenagers from multi-ethnic urban neighborhoods in casual conversations with their peers (Wiese, 2017). In the last 30 years, it has developed systematic linguistic structures that identify it as an independent variety of German (Wiese, Freywald, & Mayr, 2009). Previous studies have extensively investigated the linguistic properties of Kiezdeutsch, mostly from a qualitative perspective (Wiese & Pohle, 2016). The main differences with standard German at the syntactic level are (see examples in the next page): bare noun phrases lacking (1) determiners or (2) prepositions; (3) lack of copula verbs; (4) verb-first declaratives; and (5) subject-verb-object (SVO) word order in sentences beginning with an adverb. In this study, we perform the first systematic bottom-up data-driven comparison of the lexical and morpho-syntactic properties of Kiezdeutsch as compared to standard German.

Materials We use two corpora of spoken German, transcribed and POS-tagged: KiDKo, (Kiezdeutsch, dialogues; 229,967 tokens; Rehbein, Schalowski, & Wiese, 2014) and GRAIN (standard German, radio interviews; 219,650 tokens; Schweitzer et al., 2018). Using stratified sampling (Levy & Lemeshow, 2013), we extract the same number of unigrams and trigrams from the two corpora while preserving the frequency distributions of the original data. Moreover, we distinguish between non-lexicalized POS n-grams (e.g., PRON-VERB-NOUN for the trigram *Ich geh Kino*, "I go to the cinema") and lexicalized n-grams replacing one POS at the time with the respective lexeme (e.g., *Ich*-VERB-NOUN, PRON-*geh*-NOUN, PRON-VERB-*Kino*).

Method We identify the most distinctive POS features by applying logistic regression models: we predict the categorical variable `corpus_type` (KiDKo or GRAIN) using as predictor the presence/absence of one unigram/trigram type at a time. The sign of the z-score indicates the direction of the effect, that is, whether the POS n-gram feature is more predictive of KiDKo (positive sign) or of GRAIN (negative sign). The absolute value of the z-score is directly related to the level of uncertainty involved in the prediction of one of the two corpora (p-value <0.001).

Study 1: Unigram POS Analysis We compare the distribution of POS in the two corpora by running 10 logistic regression models, one for each coarse-grain POS (e.g., NOUN, VERB, etc.). As shown in Table 1, five POS types are significantly more predictive of GRAIN (left) and five of KiDKo (right). These results are in line with previous qualitative studies (Wiese, 2016): determiners are used significantly less in Kiezdeutsch than in standard German (see Example (1)); similarly, adpositions are much less used by teenagers than adults (see Example (2)). In Table 2, the top-ranked lexicalized unigrams show that GRAIN contains more formal nouns and also that verbs are part of more complex structures (e.g., modal structures). In KiDKo, the first and second singular forms (self-centered conversations) are the most predictive ones together with nouns referring to typical topics for young people (school, home, games).

Study 2: Trigram POS Analysis We analyze the distribution of a total of 1,245 POS trigrams (e.g., DET+ADJ+NOUN) in the two corpora, and run 1,245 logistic regression models. Significant level is reached when the z-value is larger than $\pm$ 3.2 (alpha-corrected p-value <0.0008). Table 3 lists the most predictive triplets for GRAIN (left) and KiDKo (right). Overall, 178 triplets are highly significant for GRAIN and 181 for KiDKo. In line with Study 1, we see how triplets of POS involving nouns and determiners are more predictive of GRAIN, while verbs and pronouns dominate the KiDKo top-ranks. In Kiezdeutsch, the clear preference for pronouns, as opposed to nouns, can be explained by the topics of spontaneous speech being much more related to conversations involving actors present in the scene. Nouns, on the other side, are essential when referring to events far from the proximity of the speech act as in political interviews.

Conclusion Our results are consistent with predictions drawn from the theoretical literature. In addition, we provide a quantitative empirical framework for larger-scale bottom-up analyses that result in a multi-faceted set of lexical and morpho-syntactic observations.

Examples

(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 & 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.")

GRAIN	**z**	**KiDKo**	**z**
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 single POS predicting GRAIN (left) vs. KiDKo (right).

GRAIN-V	trans.	**z**	**KiDKo-V**	trans.	**z**
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
GRAIN-N	trans.	**z**	**KiDKo-N**	trans.	**z**
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

Table 2: The five most predictive verbs (top) and nouns (bottom) in GRAIN (left) vs. KiDKo (right) with the corresponding z-scores.

GRAIN	**z**	**KiDKo**	**z**
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

Table 3: Distribution of z-scores of most predictive POS triplets for GRAIN (left) vs. KiDKo (right).

References

Levy, P. S., & Lemeshow, S. (2013). *Sampling of populations: Methods and applications*.

Rehbein, I., Schalowski, S., & Wiese, H. (2014). The KiezDeutsch Korpus (KiDKo) Release 1.0. In *Proceedings of LREC* (pp. 3927–3934). Reykjavik, Iceland.

Schweitzer, K., Eckart, K., Gärtner, M., Falenska, A., Riester, A., Rösiger, I., ... Kuhn, J. (2018). German radio interviews: The GRAIN release of the SFB732 Silver Standard Collection. In *Proceedings of LREC.*

Wiese, H. (2016). What can new urban dialects tell us about internal language dynamics? The power of language diversity. *Linguistische Berichte*, *19*, 207–245.

Wiese, H. (2017). Urban contact dialects. In S. S. Mufwene & A. M. Escobar (Eds.), *Cambridge Handbook of Language Contact.* Cambridge: Cambridge University Press.

Wiese, H., Freywald, U., & Mayr, K. (2009). Kiezdeutsch as a test case for the interaction between grammar and information structure. *Working Papers of the SFB 632*, *12*.

Wiese, H., & Pohle, M. (2016). "Ich geh Kino" oder " ... ins Kino"? *Zeitschrift für Sprachwissenschaft*, *35*(2), 171–216.