# Morpho-Syntactic Properties of Kiezdeutsch: Lexicalised POS Regression Analyses

Poster 118

Diego Frassinelli (University of Konstanz);
Gabriella Lapesa, Reem Alatrash, Dominik Schlechtweg,
Sabine Schulte im Walde (University of Stuttgart)

# Our Work

We propose a quantitative **empirical framework** for larger-scale bottom-up analyses of the **lexical** and **morpho-syntactic** properties of language varieties

**Case study:** KiezDeutsch vs. Standard German

**Results:**

- Large scale analysis of linguistic properties
- Consistent with the predictions drawn from the theoretical literature
- Successful method to identify the most salient properties that distinguish one language variety from another

# Kiezdeutsch: an independent variety of German

Spoken by teenagers from multi-ethnic urban neighborhoods in casual conversations among peers → **interesting morpho-syntactic properties**:

1. Bare nouns lacking determiners:
   - Hast du ~~**ein**~~ Problem? (Do you have problem?)
2. Bare nouns lacking prepositions:
   - Ich gehe ~~**ins**~~ Kino. (I go cinema)
3. Lack of copula verbs:
   - Er ~~**ist**~~ aus Kreuzberg. (He from Kreuzberg.)
4. Verb first declaratives:
   - ~~**Ich**~~ Wollte **ich** keine Hektik machen da drinne. (I didn't want to make any hectic in there.)
5. Subject-Verb-Object word order in sentences beginning with an adverb:
   - Jetzt ~~**bin**~~ ich **bin** 18. (Now I am 18.)

# Materials

## KiDKo
## The KiezDeutsch Korpus*

Spoken

Casual everyday conversations in Berlin

Teenagers (14 - 17 yo), self-recording

Daily life topics

63K sentences, 359K tokens

POS tagged: Stuttgart-Tübingen tagset

Very short sentences: 8.8 tokens/sentence

** Multi-ethnic subcorpus*

## GRAIN
## The German RAdio INterviews corpus

Spoken

Interviews on the public radio

Professionals (adults), recorded

Social and political topics

14K sentences, 221K tokens

POS tagged: Stuttgart-Tübingen tagset

Longer sentences: 26.7 tokens/sentence

# Method: Z-score Analysis

- Logistic Regression models

$$\textbf{corpus type}_{\textbf{KidKo/GRAIN}} \sim \textbf{POS}_{\textbf{0/1}}$$

- One POS at the time: $NOUN_{0/1}$, $VERB_{0/1}$, …
- Analysis of the z-scores:
  - **Sign** for the direction of the effect (KidKo positive, GRAIN negative)
  - **Value** for the level of uncertainty of the prediction (bigger is more predictive)
  - **Significance** test (alpha corrected p-value < 0.001)

# Study 1a: Unigram POS Analysis

10 logistic regression models, one for each coarse-grain POS

Results for KiezDeutsch:

1. Significantly less determiners
    - Hast du ~~ein~~ Problem? (Do you have problem?)
2. Significantly less adpositions (prepositions)
    - Ich gehe ~~ins~~ Kino. (I go cinema)
3. More pronouns and verbs than nouns
    - Short sentences and a lot of direct speech

| 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 |

# Study 1b: Lexicalized Unigram POS Analysis

## Nouns

| GRAIN | trans. | z-score | KidKo | trans. | 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 |

1. Slang forms (Alter)
2. Daily topics (school, fun)

## Verbs

| GRAIN | trans. | z-score | KidKo | trans. | 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 |

1. Needing, having, obligations
2. Simpler structures

# Study 2: Trigram POS Analysis

1,245 logistic regression models, one for each POS triplet

| GRAIN | | KidKo | |
|---|---|---|---|
| **POS** | **z-sc.** | **POS** | **z-sc.** |
| 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 |

1. Preference for VERBS and PRON
2. Three very frequent POS triplets
3. Longer tail of infrequent POS
4. Steeper curve: not many mid-frequency POS

## Frequency Distributions