A Large-scale Logistic Regression Analysis of Kiezdeutsch Syntax

Reem Alatrash

Institut fur Maschinelle Sprachverarbeitung
Universität Stuttgart

January 15, 2020

Outline

1. Overview
2. Methodology
3. Data
4. Experiments & Results
5. Conclusion
6. Discussion

What is Kiezdeutsch?

Definition

German-language variety spoken primarily by teenagers from multi-ethnic urban neighborhoods in casual conversations with their peers.

What is Kiezdeutsch?

- A way of self-identification.

Figure: German rappers Eko Fresh and Ali Bumaye sing

"Lan lass ma' ya!" (Dude, let's go!). Source: YouTube.com

What is Kiezdeutsch?

- Part of bigger phenomenon 'Urban Youth Languages'.
  Other examples: Multicultural London English (UK), straattaal (Netherlands), Rinkebysvenska (Sweden), Isamto (Africa)

What is Kiezdeutsch?

- Part of bigger phenomenon 'Urban Youth Languages'.
  Other examples: Multicultural London English (UK), straattaal (Netherlands), Rinkebysvenska (Sweden), Isamto (Africa)
- **A dialect in its own standing** - Heike Wiese

Syntactic Phenomena in Kiezdeutsch 1

- **bare NPs:** Noun phrases (NPs) lacking determiners and/or prepositions.

Syntactic Phenomena in Kiezdeutsch 1

- **bare NPs:** Noun phrases (NPs) lacking determiners and/or prepositions.

(1) Können wir Party machen?
Can we party make?
'Can we have [a] party?'

Syntactic Phenomena in Kiezdeutsch 1

- **bare NPs:** Noun phrases (NPs) lacking determiners and/or prepositions.

(1) Können wir Party machen?
Can we party make?
'Can we have [a] party?'

Standard German: Können wir eine Party machen?

Syntactic Phenomena in Kiezdeutsch 2

- **Directive Particles:** New particles “Lassma”, “mussttu” at start of sentence.

Syntactic Phenomena in Kiezdeutsch 2

- **Directive Particles:** New particles “Lassma”, “mussttu” at start of sentence.

(2) Lass mal morgen saufen gehen SPK19.
let once tomorrow drinking go SPK19.
‘Let’s go drinking tomorrow SPK19.’

Syntactic Phenomena in Kiezdeutsch 3

- **V1:** Verb-first (V1) declaratives.

Syntactic Phenomena in Kiezdeutsch 3

- **V1:** Verb-first (V1) declaratives.

(3) Mache ich so.
Make I so.
'I do that.'

Syntactic Phenomena in Kiezdeutsch 3

- **V1:** Verb-first (V1) declaratives.

(3) Mache ich so.
Make I so.
'I do that.'

Standard German: Ich mache das so.

Other Phenomena in Kiezdeutsch

- Many other phenomena both syntactic (e.g., verb-first declaratives) and non-syntactic (pronouncing ‘ich’ as ‘ish’).

Motivation

- Nowadays many young Germans speak Kiezdeutsch regardless of background.

Motivation

- Nowadays many young Germans speak Kiezdeutsch regardless of background.
- Research to date has focused on either qualitative analysis or small-scale quantitative studies of hand picked phenomena in Kiezdeutsch.

Motivation

- Nowadays many young Germans speak Kiezdeutsch regardless of background.
- Research to date has focused on either qualitative analysis or small-scale quantitative studies of hand picked phenomena in Kiezdeutsch.
- **Gap in research:** no large-scale computational analysis of Kiezdeutsch.

Motivation

- Nowadays many young Germans speak Kiezdeutsch regardless of background.
- Research to date has focused on either qualitative analysis or small-scale quantitative studies of hand picked phenomena in Kiezdeutsch.
- **Gap in research:** no large-scale computational analysis of Kiezdeutsch.

Motivation

- Nowadays many young Germans speak Kiezdeutsch regardless of background.
- Research to date has focused on either qualitative analysis or small-scale quantitative studies of hand picked phenomena in Kiezdeutsch.
- **Gap in research:** no large-scale computational analysis of Kiezdeutsch.

Goal $\rightarrow$ Fill the gap!

Contributions

- Perform a large-scale logistic regression analysis of Kiezdeutsch syntax with respect to standard German to reveal part-of-speech (POS) n-grams characteristic of Kiezdeutsch.

Contributions

- Perform a large-scale logistic regression analysis of Kiezdeutsch syntax with respect to standard German to reveal part-of-speech (POS) n-grams characteristic of Kiezdeutsch.
- Test whether the distributional properties of the POS n-grams affect their predictability of Kiezdeutsch.

Contributions

- Perform a large-scale logistic regression analysis of Kiezdeutsch syntax with respect to standard German to reveal part-of-speech (POS) n-grams characteristic of Kiezdeutsch.
- Test whether the distributional properties of the POS n-grams affect their predictability of Kiezdeutsch.
- Test the impact of POS granularity and interaction between POS tags within an n-gram.

Contributions

- Perform a large-scale logistic regression analysis of Kiezdeutsch syntax with respect to standard German to reveal part-of-speech (POS) n-grams characteristic of Kiezdeutsch.
- Test whether the distributional properties of the POS n-grams affect their predictability of Kiezdeutsch.
- Test the impact of POS granularity and interaction between POS tags within an n-gram.
- Test the impact of adding positional information.

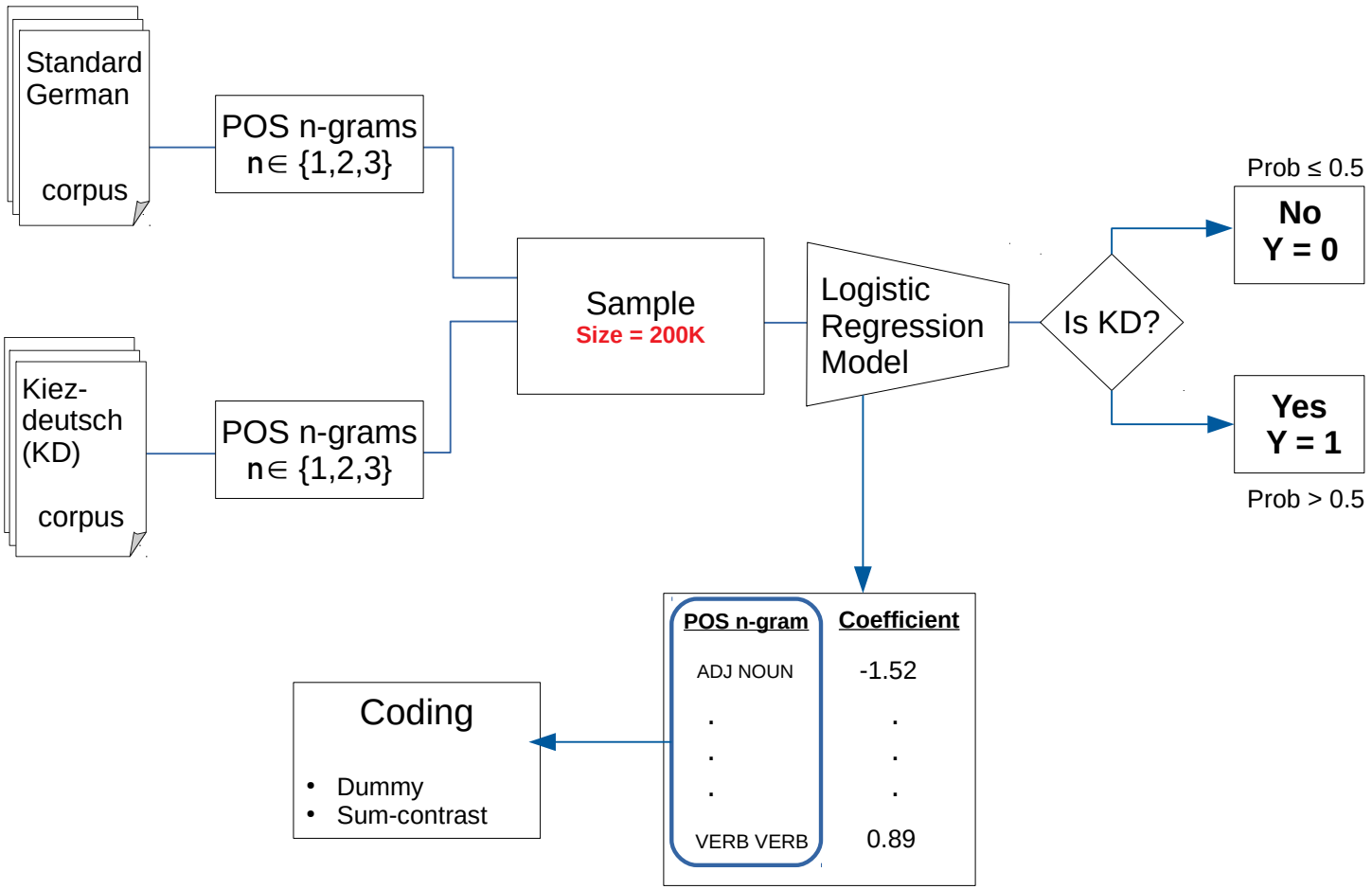
Contributions

- Perform a large-scale logistic regression analysis of Kiezdeutsch syntax with respect to standard German to reveal part-of-speech (POS) n-grams characteristic of Kiezdeutsch.
- Test whether the distributional properties of the POS n-grams affect their predictability of Kiezdeutsch.
- Test the impact of POS granularity and interaction between POS tags within an n-gram.
- Test the impact of adding positional information.
- Outline a robust approach to model selection parameter selection.

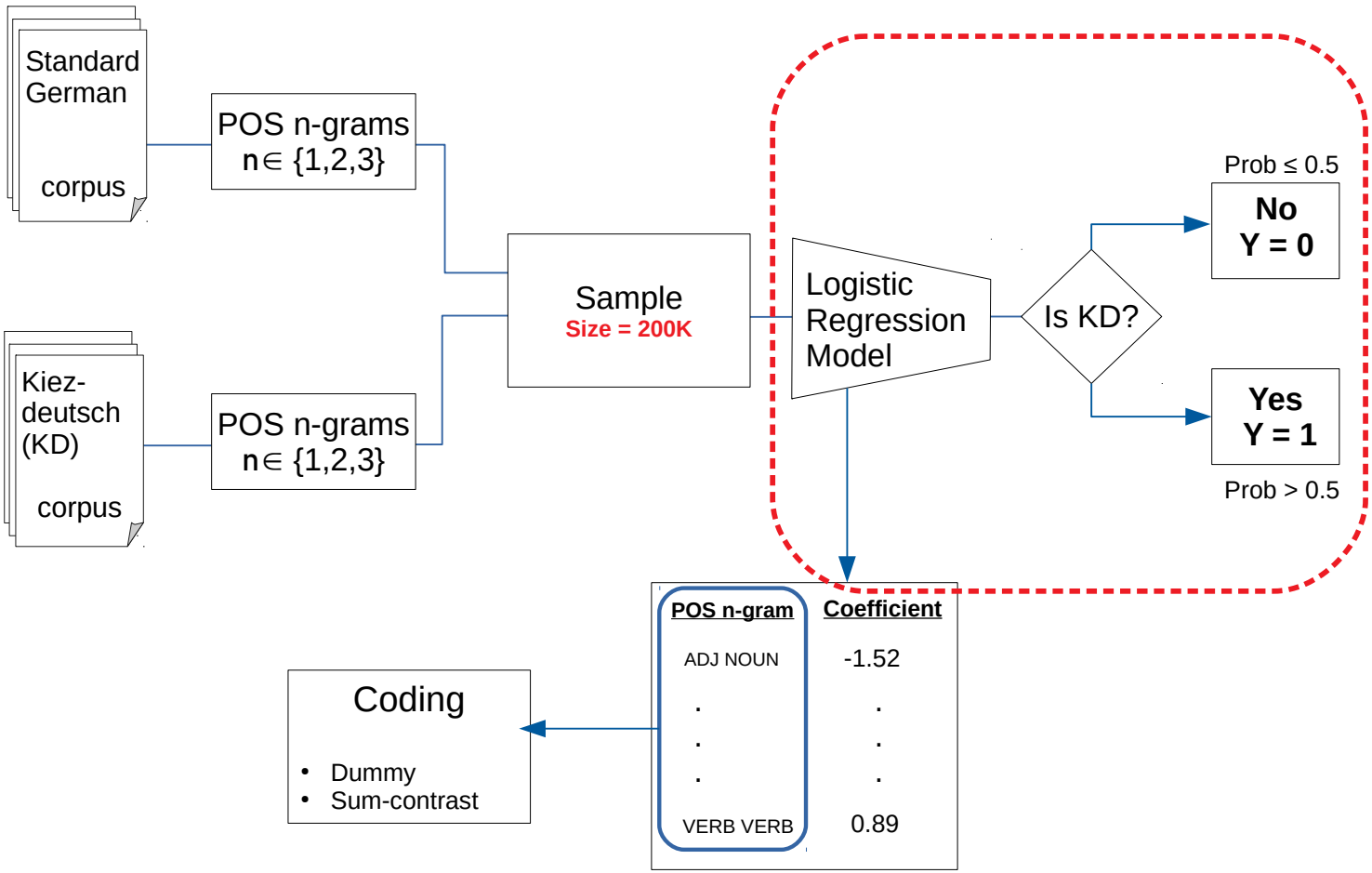
Contributions - This Talk

- Perform a large-scale logistic regression analysis of Kiezdeutsch syntax with respect to standard German to reveal part-of-speech (POS) n-grams characteristic of Kiezdeutsch.
- Test whether the distributional properties of the POS n-grams affect their predictability of Kiezdeutsch.
- Test the impact of POS granularity and interaction between POS tags within an n-gram.
- Test the impact of adding positional information.
- Outline a robust approach to model selection parameter selection.

Methodology - Overview



Methodology - Model



Logistic Regression 1

- Logistic regression is a supervised machine learning approach commonly used for binary classification.

Logistic Regression 1

- Logistic regression is a supervised machine learning approach commonly used for binary classification.
- It uses the logistic/sigmoid function to calculate the probability of the outcome.

Logistic Regression 2

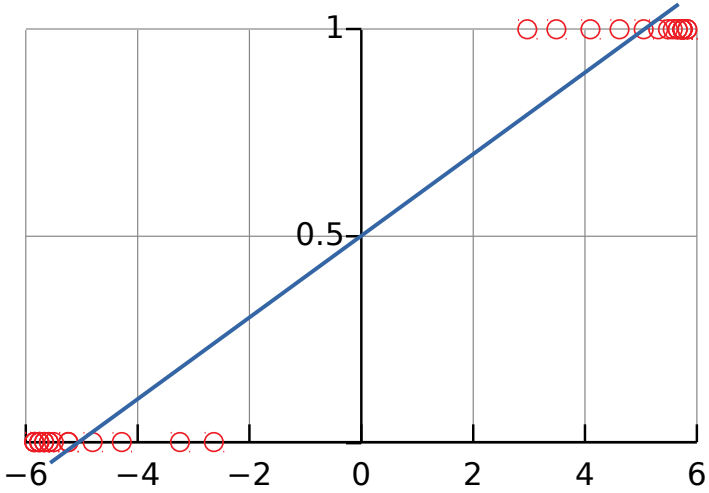


Figure: Linear Regression Model

$$Y = \alpha + \beta X + \epsilon$$

α = intercept, β = slope, ϵ = random error

Logistic Regression 2

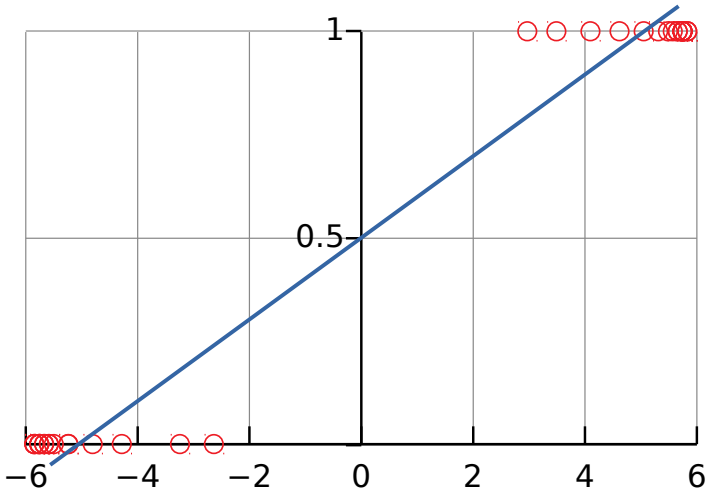
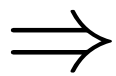


Figure: Linear Regression Model



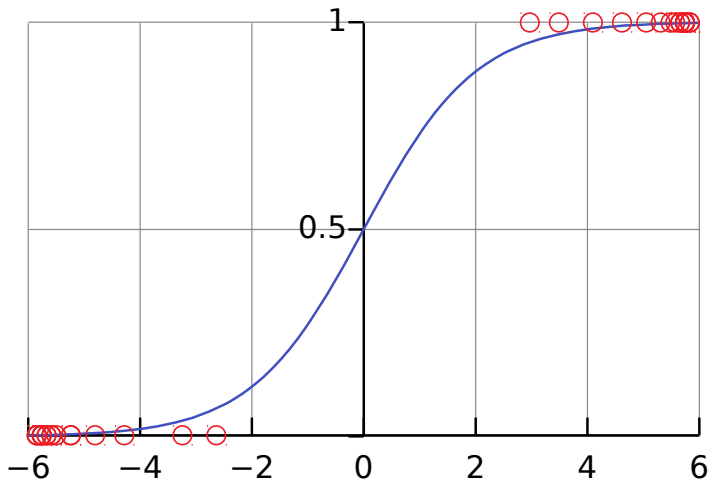


Figure: Logistic Regression Model

$$Y = \alpha + \beta X + \epsilon$$

α = intercept, β = slope, ϵ = random error

$$f(Y) = \alpha + \beta X$$

Logistic Regression 3

- A logistic regression model is a type of Generalized Linear Model (GLM).

Logistic Regression 3

- A logistic regression model is a type of Generalized Linear Model (GLM).
- GLM extends the linear model by allowing non-normal distributions for the outcome.

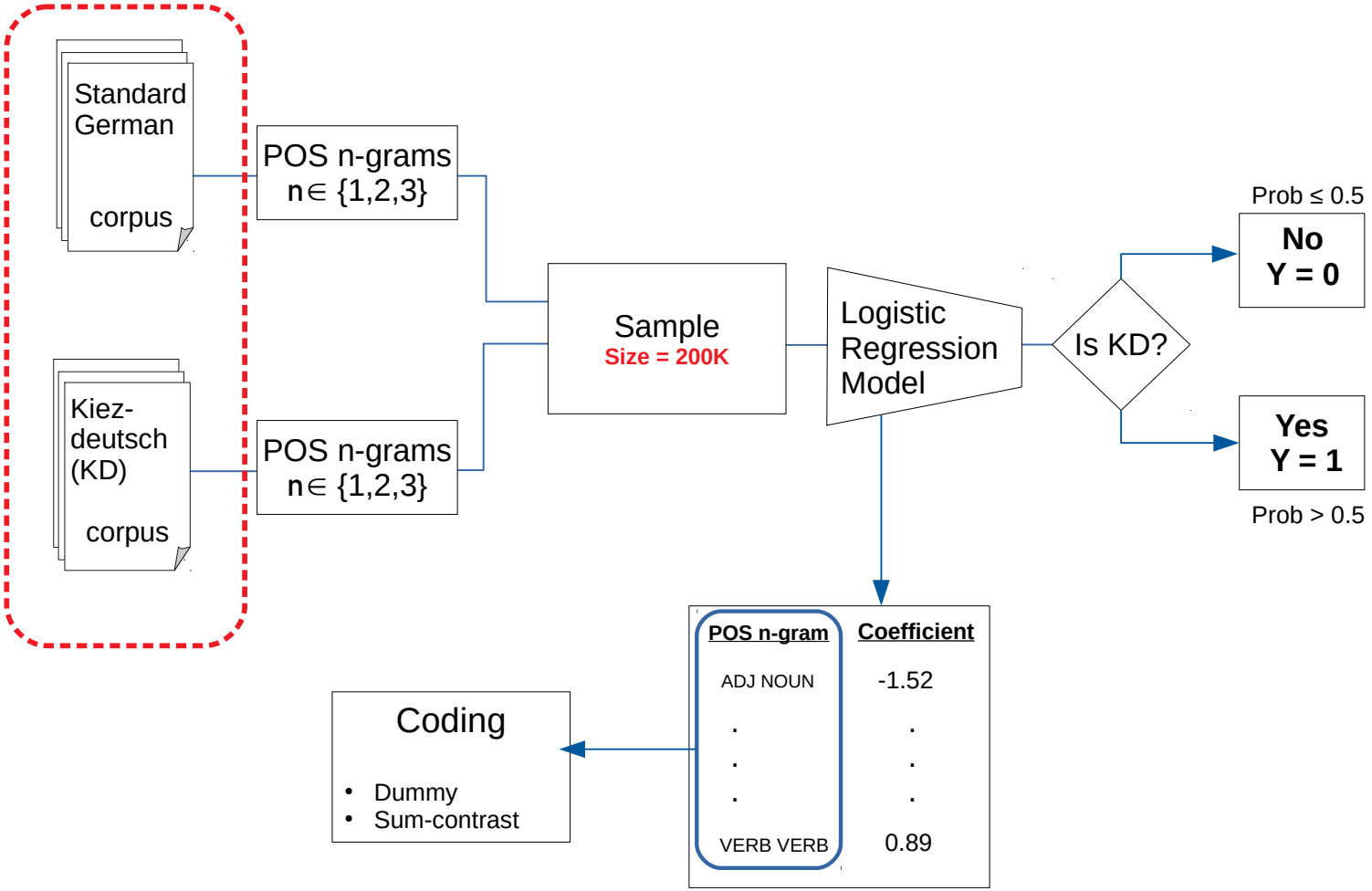
Logistic Regression 3

- A logistic regression model is a type of Generalized Linear Model (GLM).
- GLM extends the linear model by allowing non-normal distributions for the outcome.
- The Generalized Linear Mixed Model (GLMM) extends GLM to account for factors that affect the outcome but are not directly studied (e.g., subjects in an experiment).

Logistic Regression 3

- A logistic regression model is a type of Generalized Linear Model (GLM).
- GLM extends the linear model by allowing non-normal distributions for the outcome.
- The Generalized Linear Mixed Model (GLMM) extends GLM to account for factors that affect the outcome but are not directly studied (e.g., subjects in an experiment).
- **This thesis $\rightarrow$ GLM and GLMM.**

Methodology - Data



Kiez**D**eutsch **Ko**rpus (KiDKo)

- Collection of spontaneous peer-group dialog between teenagers from multi-ethnic and mono-ethnic communities in Berlin.

Kiez**D**eutsch **Ko**rpus (KiDKo)

- Collection of spontaneous peer-group dialog between teenagers from multi-ethnic and mono-ethnic communities in Berlin.
- **Speakers:** 2+ per conversation, and are teen-aged students (14-17).

Kiez**D**eutsch **Ko**rpus (KiDKo)

- Collection of spontaneous peer-group dialog between teenagers from multi-ethnic and mono-ethnic communities in Berlin.
- **Speakers:** 2+ per conversation, and are teen-aged students (14-17).
- **POS:** Automatically tagged using Stuttgart-Tübingen-TagSet (STTS).

Kiez**D**eutsch **Ko**rpus (KiDKo)

- Collection of spontaneous peer-group dialog between teenagers from multi-ethnic and mono-ethnic communities in Berlin.
- **Speakers:** 2+ per conversation, and are teen-aged students (14-17).
- **POS:** Automatically tagged using Stuttgart-Tübingen-TagSet (STTS).
- **Sub-corpora:** 1 multi-ethnic community (KiDKo-Mu) and 1 for mono-ethnic (KiDKo-Mo).

Kiez**D**eutsch **Ko**rpus (KiDKo)

- Collection of spontaneous peer-group dialog between teenagers from multi-ethnic and mono-ethnic communities in Berlin.
- **Speakers:** 2+ per conversation, and are teen-aged students (14-17).
- **POS:** Automatically tagged using Stuttgart-Tübingen-TagSet (STTS).
- **Sub-corpora:** 1 multi-ethnic community (KiDKo-Mu) and 1 for mono-ethnic (KiDKo-Mo).
- **This thesis** $\rightarrow$ **KiDKo-Mu.**

German RAdio INterviews (GRAIN)

- Non-static collection of interviews broadcast weekly on German public radio.

German RAdio INterviews (GRAIN)

- Non-static collection of interviews broadcast weekly on German public radio.
- **Speakers:** 2 adults per interview (1 host, 1 guest). Guests appear in their professional capacity (e.g., director, council chairman)

German RAdio INterviews (GRAIN)

- Non-static collection of interviews broadcast weekly on German public radio.
- **Speakers:** 2 adults per interview (1 host, 1 guest). Guests appear in their professional capacity (e.g., director, council chairman)
- **POS:** Tagged using STTS.

German RAdio INterviews (GRAIN)

- Non-static collection of interviews broadcast weekly on German public radio.
- **Speakers:** 2 adults per interview (1 host, 1 guest). Guests appear in their professional capacity (e.g., director, council chairman)
- **POS:** Tagged using STTS.
- **Sub-corpora:** silver standard set (automatically annotated) and gold standard set (manually annotated).

German RAdio INterviews (GRAIN)

- Non-static collection of interviews broadcast weekly on German public radio.
- **Speakers:** 2 adults per interview (1 host, 1 guest). Guests appear in their professional capacity (e.g., director, council chairman)
- **POS:** Tagged using STTS.
- **Sub-corpora:** silver standard set (automatically annotated) and gold standard set (manually annotated).
- **This thesis $\rightarrow$ Silver standard set.**

Key Statistics

Corpus	Original Size	Processed Size	Speakers
KiDKo (Mu)	359,000	230,000	201
GRAIN (Silver set)	221,000	220,000	124

Table: Key statistics of the corpora used in this thesis. Numbers are approximates.

Data Processing

Processing included the following steps:

- Assign unique IDs to speakers in both corpora.

Data Processing

Processing included the following steps:

- Assign unique IDs to speakers in both corpora.
- Cleanup: remove punctuation (e.g., !,.), speech disfluencies (e.g., pauses, hesitation, repeated words) and non-words (e.g., uninterpretable).

Data Processing

Processing included the following steps:

- Assign unique IDs to speakers in both corpora.
- Cleanup: remove punctuation (e.g., !,.), speech disfluencies (e.g., pauses, hesitation, repeated words) and non-words (e.g., uninterpretable).
- Get sentence-level information (e.g., full sentence, length).

Data Processing

Processing included the following steps:

- Assign unique IDs to speakers in both corpora.
- Cleanup: remove punctuation (e.g., !,.), speech disfluencies (e.g., pauses, hesitation, repeated words) and non-words (e.g., uninterpretable).
- Get sentence-level information (e.g., full sentence, length).
- Map fine-grained POS tags from STTS to coarse-grained universal dependency (UD) tags (e.g., NE,NN $\rightarrow$ NOUN).

Data Processing

Processing included the following steps:

- Assign unique IDs to speakers in both corpora.
- Cleanup: remove punctuation (e.g., !,.), speech disfluencies (e.g., pauses, hesitation, repeated words) and non-words (e.g., uninterpretable).
- Get sentence-level information (e.g., full sentence, length).
- Map fine-grained POS tags from STTS to coarse-grained universal dependency (UD) tags (e.g., NE,NN $\rightarrow$ NOUN).
- Lemmatize KiDKo tokens (e.g., habe, hast $\rightarrow$ haben)

Data Exploration

Data exploration revealed the following:

- Both corpora follow a Zipfian distribution.

Data Exploration

Data exploration revealed the following:

- Both corpora follow a Zipfian distribution.
- KiDKo has more sentences, but they are shorter.

Data Exploration

Data exploration revealed the following:

- Both corpora follow a Zipfian distribution.
- KiDKo has more sentences, but they are shorter.
- KiDKo has much more particles, verbs, and pronoun.

Data Exploration

Data exploration revealed the following:

- Both corpora follow a Zipfian distribution.
- KiDKo has more sentences, but they are shorter.
- KiDKo has much more particles, verbs, and pronoun.
- GRAIN has much more determiners, nouns and adpositions.

Data Exploration

Data exploration revealed the following:

- Both corpora follow a Zipfian distribution.
- KiDKo has more sentences, but they are shorter.
- KiDKo has much more particles, verbs, and pronoun.
- GRAIN has much more determiners, nouns and adpositions.
- Some speech disfluencies in KiDKo are not tagged as such.
  Example: repeated words tagged according to their POS.

Experiment 1

- **Contribution:** Build GLM and GLMMs to find which POS n-grams are most predictive of Kiezdeutsch in the dataset.

Experiment 1

- **Contribution:** Build GLM and GLMMs to find which POS n-grams are most predictive of Kiezdeutsch in the dataset.
- **Models:** 22 GLMs & GLMMs for main experiment, 12 models for additional experiments (e.g., test granularity & interaction).

Experiment 1

- **Contribution:** Build GLM and GLMMs to find which POS n-grams are most predictive of Kiezdeutsch in the dataset.
- **Models:** 22 GLMs & GLMMs for main experiment, 12 models for additional experiments (e.g., test granularity & interaction).
- We discuss the results of the POS n-grams GLMs with sum-contrast coding.

Experiment 1 - POS Unigram Results

- **Most predictive:** Particles (e.g., Ja, nicht), numerals (e.g., zwei, 2008) and pronouns (e.g., ich, du).
  $\rightarrow$ GLM supports directive particles (lassma) and particle 'so' phenomena.

Experiment 1 - POS Unigram Results

- **Most predictive:** Particles (e.g., Ja, nicht), numerals (e.g., zwei, 2008) and pronouns (e.g., ich, du).
  $\rightarrow$ GLM supports directive particles (lassma) and particle 'so' phenomena.
- Particles also highlight backchannel responses (e.g., 'Ja' & 'Hm-hm') which are important in conversations among bilingual speakers.

Experiment 1 - POS Unigram Results

- **Most predictive:** Particles (e.g., Ja, nicht), numerals (e.g., zwei, 2008) and pronouns (e.g., ich, du).
  → GLM supports directive particles (lassma) and particle 'so' phenomena.
- Particles also highlight backchannel responses (e.g., 'Ja' & 'Hm-hm') which are important in conversations among bilingual speakers.
- **Least predictive:** Determiners (die, der), adpositions (in, auf), and nouns (Deutschland, Alter).
  → GLM supports bare NPs phenomenon.

Experiment 1 - POS Bigram Results

- **Most predictive:** "PRT PRT" (e.g., Ja ja), "VERB ADV" (e.g., Lass mal), "PRT NOUN" (e.g., nicht Training), and "PRT NUM" (e.g., nicht 360).
  $\rightarrow$ GLM supports directive particles (lassma) and particle 'so' phenomena.

Experiment 1 - POS Bigram Results

- **Most predictive:** "PRT PRT" (e.g., Ja ja), "VERB ADV" (e.g., Lass mal), "PRT NOUN" (e.g., nicht Training), and "PRT NUM" (e.g., nicht 360).
  $\rightarrow$ GLM supports directive particles (lassma) and particle 'so' phenomena.
- Particle 'nicht' may indicate increased use of negation in Kiezdeutsch.

Experiment 1 - POS Bigram Results

- **Most predictive:** "PRT PRT" (e.g., Ja ja), "VERB ADV" (e.g., Lass mal), "PRT NOUN" (e.g., nicht Training), and "PRT NUM" (e.g., nicht 360).
  → GLM supports directive particles (lassma) and particle 'so' phenomena.
- Particle 'nicht' may indicate increased use of negation in Kiezdeutsch.
- **Least predictive:** "ADP DET" (e.g., in das, von der), "DET NOUN" (e.g., die Grünen), "DET VERB" (e.g., die sollte) , and "NOU DET" (e.g., den Kandidaten).
  → GLM supports bare NPs phenomenon.

Experiment 1 - POS Bigram Results

- **Most predictive:** "PRT PRT" (e.g., Ja ja), "VERB ADV" (e.g., Lass mal), "PRT NOUN" (e.g., nicht Training), and "PRT NUM" (e.g., nicht 360).
  → GLM supports directive particles (lassma) and particle 'so' phenomena.
- Particle 'nicht' may indicate increased use of negation in Kiezdeutsch.
- **Least predictive:** "ADP DET" (e.g., in das, von der), "DET NOUN" (e.g., die Grünen), "DET VERB" (e.g., die sollte) , and "NOU DET" (e.g., den Kandidaten).
  → GLM supports bare NPs phenomenon.
- "DET VERB" and "NOU DET" may indicate decreased use of relative clauses.

Experiment 1 - POS Trigram Results

- Quasi complete separation detected for several POS triples like “PRT PRT PRT” (e.g., Ja ja ja), “DET VERB ADP” (e.g., der war im).

Experiment 1 - POS Trigram Results

- Quasi complete separation detected for several POS triples like “PRT PRT PRT” (e.g., Ja ja ja), “DET VERB ADP” (e.g., der war im).
- **Most predictive:** “PRT NOUN VERB” (e.g., nicht Shisha rauchen), “PRT NOUN ADV’ (e.g., nicht Schluss so), “PRT PRT NOUN” (e.g., nicht eh Samstag).
  $\rightarrow$ may indicate increased use of negation in Kiezdeutsch.

Experiment 1 - POS Trigram Results

- Quasi complete separation detected for several POS triples like “PRT PRT PRT” (e.g., Ja ja ja), “DET VERB ADP” (e.g., der war im).
- **Most predictive:** “PRT NOUN VERB” (e.g., nicht Shisha rauchen), “PRT NOUN ADV’ (e.g., nicht Schluss so), “PRT PRT NOUN” (e.g., nicht eh Samstag).
  $\rightarrow$ may indicate increased use of negation in Kiezdeutsch.
- **Least predictive:** “DET ADP NOUN” (e.g., den im Jahre), “DET ADV ADJ” (e.g., die ganz klare), “NOUN DET ADV” (e.g., Präsidentschaft die jetzt).
  $\rightarrow$ GLM supports bare NPs phenomenon.

Experiment 1 - POS Trigram Results

- Quasi complete separation detected for several POS triples like “PRT PRT PRT” (e.g., Ja ja ja), “DET VERB ADP” (e.g., der war im).
- **Most predictive:** “PRT NOUN VERB” (e.g., nicht Shisha rauchen), “PRT NOUN ADV’ (e.g., nicht Schluss so), “PRT PRT NOUN” (e.g., nicht eh Samstag).
  $\rightarrow$ may indicate increased use of negation in Kiezdeutsch.
- **Least predictive:** “DET ADP NOUN” (e.g., den im Jahre), “DET ADV ADJ” (e.g., die ganz klare), “NOUN DET ADV” (e.g., Präsidentschaft die jetzt).
  $\rightarrow$ GLM supports bare NPs phenomenon.
- “NOUN DET ADV” may indicate decreased use of relative clauses in Kiezdeutsch.

Experiment 2

- **Contribution:** Add positional information then run GLMs from Experiment 1 to find which POS n-grams are most predictive of Kiezdeutsch in the dataset.

Experiment 2

- **Contribution:** Add positional information then run GLMs from Experiment 1 to find which POS n-grams are most predictive of Kiezdeutsch in the dataset.
- **Models:** 6 GLMs were tested on a sample of size 100,000.

Experiment 2 - Positional Information

Positional information was added in two ways:

- **Sentence Markers:** introduce 2 POS tags to mark sentence boundaries, SOS (start of sentence) and EOS (end of sentence).

Experiment 2 - Positional Information

Positional information was added in two ways:

- **Sentence Markers:** introduce 2 POS tags to mark sentence boundaries, SOS (start of sentence) and EOS (end of sentence).
- **Augmented POS tags:** add affix to each POS tag to indicate its position in the sentence.

Experiment 2 - Positional Information 2

(4) a. <SOS> Was machst du <EOS>
SOS PRON VERB PRON EOS
'What are you doing?'

b. Was machst du
SOS_PRON VERB_MID PRON_EOS
'What are you doing?'

Experiment 2 - Results 1

Sentence markers results:

- Results for all POS n-gram models in line with experiment 1.
  $\rightarrow$ support for bare NPs, directive particles and particle 'so'.

Experiment 2 - Results 1

Sentence markers results:

- Results for all POS n-gram models in line with experiment 1.
  → support for bare NPs, directive particles and particle 'so'.
- POS bigram "SOS VERB" (e.g., <SOS> Sehe), POS trigrams "SOS VERB NOUN" (e.g., <SOS> War Deutscher), "SOS VERB ADV" (e.g., <SOS> Habe doch), "SOS VERB ADP" (e.g., <SOS> Ist bei) are some of the most predictive of Kiezdeutsch in the data.
  → support verb-first declaratives phenomenon.

Experiment 2 - Results 2

Augmented POS tags results:

- Results for POS unigram model in line with experiment 1.
  $\rightarrow$ support for bare NPs, directive particles and particle 'so'.

Experiment 2 - Results 2

Augmented POS tags results:

- Results for POS unigram model in line with experiment 1.
  → support for bare NPs, directive particles and particle 'so'.
- POS bigram and trigram models suffered from data sparsity and separation.
  → data and models are insufficient to produce significant results.

Conclusion

- This thesis filled a gap in the knowledge by performing a large-scale logistic regression analysis of Kiezdeutsch w.r.t. standard German.

Conclusion

- This thesis filled a gap in the knowledge by performing a large-scale logistic regression analysis of Kiezdeutsch w.r.t. standard German.
- Findings support well-known phenomena: bare NPs, directive particles, particle 'so', and verb-first.

Conclusion

- This thesis filled a gap in the knowledge by performing a large-scale logistic regression analysis of Kiezdeutsch w.r.t. standard German.
- Findings support well-known phenomena: bare NPs, directive particles, particle 'so', and verb-first.
- Findings suggest possible phenomena: increased use of negation, decreased use of relative clauses.
  $\rightarrow$ further research is recommended to determine if these are Kiezdeutsch phenomena or latent properties of our corpora.

Conclusion

- This thesis filled a gap in the knowledge by performing a large-scale logistic regression analysis of Kiezdeutsch w.r.t. standard German.
- Findings support well-known phenomena: bare NPs, directive particles, particle 'so', and verb-first.
- Findings suggest possible phenomena: increased use of negation, decreased use of relative clauses.
  $\rightarrow$ further research is recommended to determine if these are Kiezdeutsch phenomena or latent properties of our corpora.
- Adding positional information improves representation of syntactic phenomena given enough data.

Conclusion

Thank you for listening.
Questions?

Coding Categorical Variables

Flavor	**C1**	**C2**	**C3**
Vanilla	0	0	0
Chocolate	**1**	0	0
Lemon	0	**1**	0
Other	0	0	**1**

Table: Dummy coding for the variable ice cream flavor with 4 groups. Vanilla is the reference level.

Coding Categorical Variables

Flavor	C1	C2	C3
Vanilla	0	0	0
Chocolate	**1**	0	0
Lemon	0	**1**	0
Other	0	0	**1**

Table: Dummy coding for the variable ice cream flavor with 4 groups. Vanilla is the reference level.

Flavor	C1	C2	C3
Vanilla	**0.75**	-0.25	-0.25
Chocolate	-0.25	**0.75**	-0.25
Lemon	-0.25	-0.25	**0.75**
Other	-0.25	-0.25	-0.25

Table: Sum contrast coding for the variable ice cream flavor with 4 groups. The grand mean is the reference level.

Logistic Regression

- A logistic regression model is a type of Generalized Linear Model (GLM) which uses the logit (log-odds) for the link function f(Y).
- It uses the logistic/sigmoid function to calculate the probability of the outcome.

$$f(Y) = \alpha + \beta X$$

Logistic Regression

- A logistic regression model is a type of Generalized Linear Model (GLM) which uses the logit (log-odds) for the link function f(Y).
- It uses the logistic/sigmoid function to calculate the probability of the outcome.

$$f(Y) = \alpha + \beta X$$

$$f(Y) = \log\left[\frac{p}{(1-p)}\right]$$

where $p = P(Y = 1)$

$$P(Y = 1) = \frac{1}{1+e^{-\theta}}$$

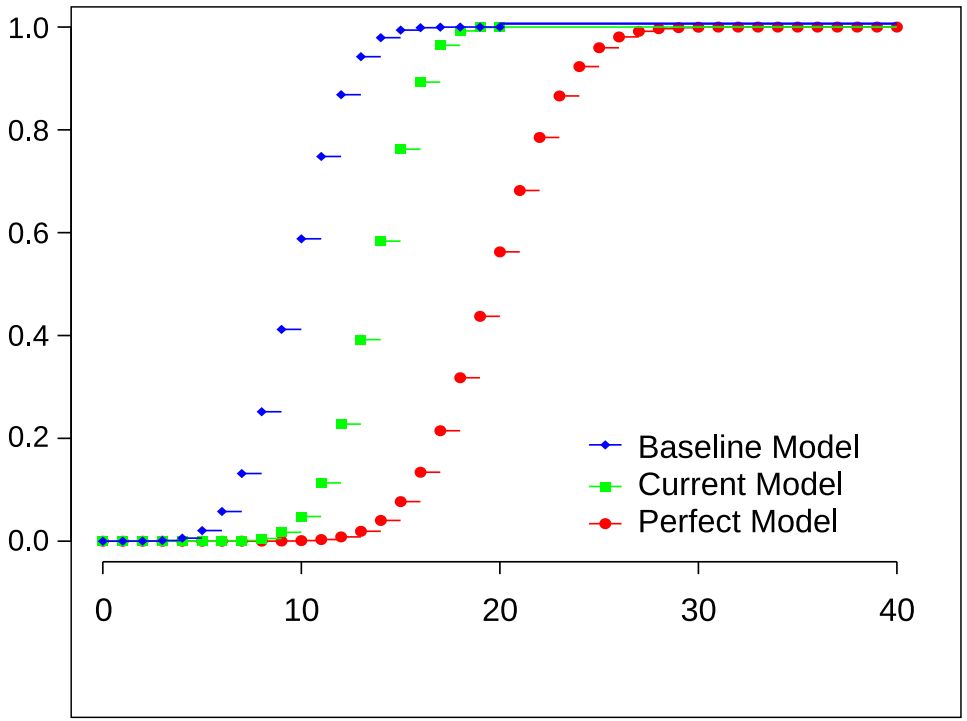
$\theta = \alpha, \beta$ (model parameters)

Comparing Models

- Analysis of Variance (ANOVA) using likelihood ratio test (LRT) reveals if the more complex model is significantly better at capturing the data than the simpler model.

Comparing Models

- Analysis of Variance (ANOVA) using likelihood ratio test (LRT) reveals if the more complex model is significantly better at capturing the data than the simpler model.



References I


Freywald, U., Mayr, K., Özçelik, T., and Wiese, H. (2011).
Kiezdeutsch as a multiethnolect.
Ethnic styles of speaking in European metropolitan areas, pages 45–73.


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


Heinz, B. (2002).
Backchannel responses as strategic responses in bilingual speakers' conversations.
Journal of Pragmatics, 35(7):1113–1142.


Jannedy, S. (2010).
The Usage and Distribution of so in Spontaneous Berlin Kiezdeutsch.
ZASPiL Nr. 52–September 2010 Papers from the Linguistics Laboratory, 43.


Kilgarriff, A. (2001).
Comparing corpora.
International journal of corpus linguistics, 6(1):97–133.

References II


Petrov, S., Das, D., and McDonald, R. (2012).
A Universal Part-of-Speech Tagset.
In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).


Rehbein, I. and Schalowski, S. (2013).
STTS goes Kiez–Experiments on Annotating and Tagging Urban Youth Language.
The Journal for Language Technology and Computational Linguistics (JLCL).


Rehbein, I., Schalowski, S., and Wiese, H. (2014).
The KiezDeutsch Korpus (KiDKo) Release 1.0.
In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.


Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999).
Guidelines für das tagging deutscher textcorpora mit STTS.
Universität Stuttgart, Universität Tübingen.

References III


Stevenson, P., Horner, K., Langer, N., and Reershemius, G. (2017).
The German-speaking world: A practical introduction to sociolinguistic issues.
Routledge, 3 edition.


Wiese, H. (2009).
Grammatical innovation in multiethnic urban europe: New linguistic practices among adolescents.
Lingua, 119(5):782–806.


Wiese, H., Freywald, U., and Mayr, K. (2009).
Kiezdeutsch as a Test Case for the Interaction Between Grammar and Information Structure.
ISIS 12, Working Papers of the SFB 632 "Information Structure".


Wiese, H., Freywald, U., Schalowski, S., and Mayr, K. (2012).
Das KiezDeutsch-Korpus. Spontansprachliche Daten Jugendlicher aus urbanen Wohngebieten.
Deutsche Sprache, 40:97–1236.