Natürliche Sprachverarbeitung für Dialekte

Wie und warum?

Verena Blaschke
MaiNLP-Gruppe, Ludwig-Maximilians-Universität München

Digital-Humanities-Kolloquium, Universität Stuttgart
29. April 2026

Natural Language Processing

… aber *welche* Sprachen?

- Viele Sprecher:innen, große Datenmengen, Standardisierung

Aber ist das repräsentativ für unsere Sprachverwendung?

- Minderheitensprachen, nicht standardisierte Varietäten, …
- Schwierig für maschinelles Lernen!
  (wenige & heterogene Daten)

Natürliche Sprachverarbeitung – welche Sprachen?

Wer von Euch/Ihnen spricht…

- einen (Nichtstandard-)Dialekt oder eine Regionalsprache?
- eine Sprache, die in wenigen/keinen Sprachverarbeitungssystemen verfügbar ist?

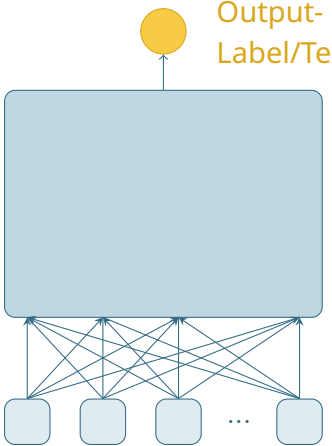
Wer hat Erfahrung mit…

- Sprachwissenschaft?
- Dialektologie?
- NLP?

Heute

- Überlegungen und Tipps zur Verarbeitung von Daten aus Dialekten und anderen ressourcenarmen Sprachvarietäten (low-resource languages, LRLs)
- Forschungsbeispiele + weitere Literaturhinweise
- Verständnisfragen gerne sofort,
  Diskussionsfragen im Anschluss

Übersicht – Herausforderungen & Herangehensweisen



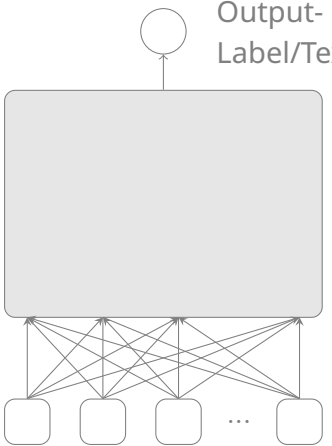
Welche NLP-Systeme & warum?

Nichtstandardsprachliche Daten modellieren

Dialekt-/LRL-Daten

- Sprachvariation
- Herausforderungen

Übersicht – Herausforderungen & Herangehensweisen



Welche NLP-Systeme & warum?

Nichtstandardsprachliche Daten modellieren

Dialekt-/LRL-Daten

- Sprachvariation
- Herausforderungen

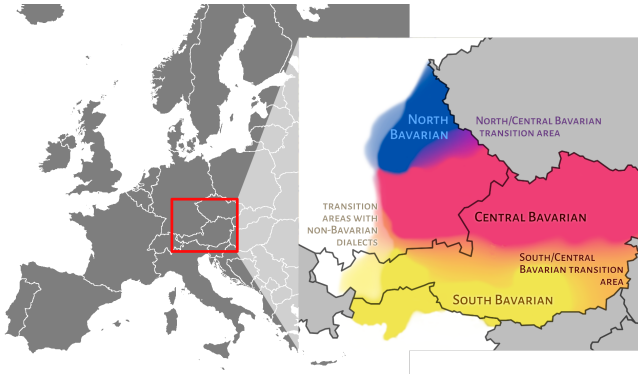
“Dialekte?”

Viele verschiedene Definitionen in der Linguistik, NLP, Alltagssprache…

- Jegliche Sprachvarietät, die von einer (geografisch) abgegrenzten Sprechergruppe gesprochen wird
- Sprachvarietäten auf Landesebene (Deutsch in DE, AT, …)
- Akzente
- …

“Dialekte?”

- **Nicht standardisiert**
- **Eng mit einer Standardsprache verwandt**
- Oft: Kontinuum Standard–Dialekte
- Often: dialektale Untergruppen



Sprachliche Unterschiede

Unterschiede zur Standardsprache

- Aussprache ($\rightarrow$ Schreibung)
- Wortschatz
- Grammatik: Morphologie, Satzbau
- Verwendungssituation – Dialektsprecher:innen beherrschen auch die Standardsprache

[Deutsch]	Sie	haben	keine	Beine	
[Bairisch]	Se	hom	koane	Haxn	ned
	De	ham	koane	Haxn	_
	Dei	hobm	koane	Haxn	_

Wann verwendet man Dialekte?

- Gesprochene Sprache
- Informelle geschriebene Kontexte (soziale Medien, Textnachrichten)
- Teils auch Literatur, Poesie; Dialektwikis

Losses da gud gehn in Albuquerque und viel Schbass bei de Konferenz!

Warum Dialekt-NLP?

- Annotationen für variationslinguistische Studien
- Maschinelles Lernen mit geringen Datenmengen und heterogenen Daten
- Anwendungen im echten Leben, die robuster mit nichtstandardsprachlichen Daten umgehen sollen
- (später mehr dazu!)

Daten

Herausforderungen in Bezug auf Dialekt- (und LRL-)Korpora

- Verfügbarkeit
- Qualität
- Schriftliche Sprachrepräsentationen

A Survey of Corpora for Germanic Low-Resource Languages and Dialects

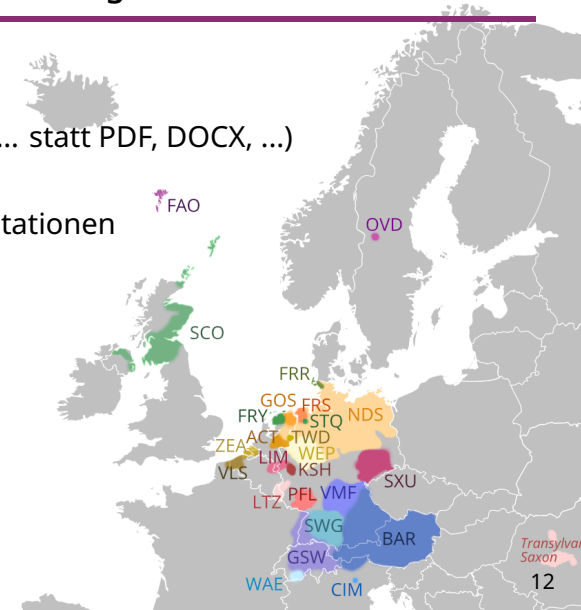
Verena Blaschke **Hinrich Schütze** **Barbara Plank**

Datensätze für ressourcenarme germanische Varietäten

- Verfügbar für Forschung
- Computerfreundliche Formate (XML, TSV, TXT, … statt PDF, DOCX, …)
- Ganze Sätze
- Mit oder ohne NLP-Annotationen
- Hohe Datenqualität

➔ 100+ Datensätze für 35 germanische Dialekte + “kleine” Sprachen

github.com/mainlp/germanic-lrl-corpora



Wie kann man Datensätze finden?

- Veröffentlichungen (ACL Anthology, arXiv, via Google Scholar / Semantic Scholar) + viel Stöbern…
- Google Dataset Search `datasetsearch.research.google.com`
- Datenrepositorien
  - Zenodo `zenodo.org`
  - European Language Grid `live.european-language-grid.eu`
  - CLARIN Virtual Language Observatory `vlo.clarin.eu`
  - OpenSLR `openslr.org`
  - Text+ `text-plus.org`
  - OLAC `www.language-archives.org`
  - ORTOLANG `www.ortolang.fr/market/corpora`
  - Hamburg Centre for Language Corpora (HZSK)
  - OPUS `opus.nlpl.eu`

Annotationen

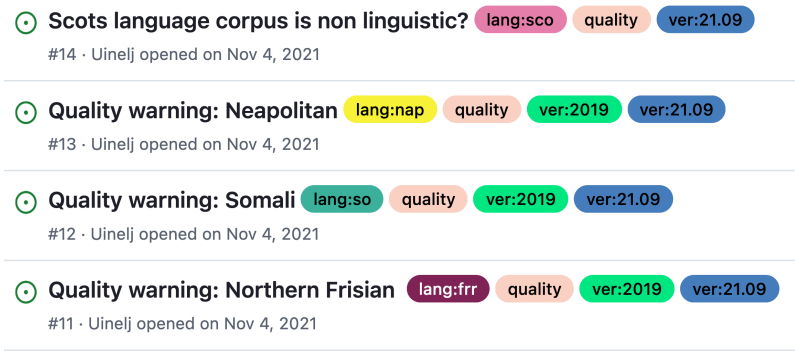
Inwieweit sind die Sprachdaten für NLP-Aufgaben annotiert?

- Morphosyntax (Wortklassen, Satzbau-Annotationen)
- Geolocation, Dialektgruppe
- Paraphrasen, Übersetzungen, Satzthemen, Absichten
  - Selten (aber werden langsam beliebter!)
- Hauptsächlich: kuratiert (elizitiert, verschriftliche Interviews, aus Büchern, manuell nachgeprüfte Internetdaten, ...), aber ohne Annotationen für NLP-Aufgaben
  - ... aber manchmal auch nicht kuratiert (z.B. aus Webcrawls)

Datenqualität: Unkuratierte Daten

Unkuratierte Daten in ressourcenarmen Sprachen weisen oft geringe Qualität auf – falsche Sprache, schlechte Datenaufbereitung (Kreutzer+, TACL 2022; Abadji+, LREC 2022)

OSCAR-Korpus (seitdem verbessert)



Datenqualität: Unkuratierte Daten

Unkuratierte Daten in ressourcenarmen Sprachen weisen oft geringe Qualität auf – falsche Sprache, schlechte Datenaufbereitung

(Kreutzer+, TACL 2022; Abadji+, LREC 2022)

“Westflämischer” QED-OPUS-Korpus

```
<w id="33.28">07,</w>
<w id="33.29">624&amp;amp;</w>
<w id="33.30">lt;</w>
<w id="33.31">br</w>
<w id="33.32">/</w>
<w id="33.33">&amp;amp;</w>
<w id="33.34">gt;</w>
<w id="33.35">Καλά</w>
<w id="33.36">,</w>
```

Datenqualität: Varietäten mit geringem Prestige

Shock an aw: US teenager wrote huge slice of Scots Wikipedia

Nineteen-year-old says he is 'devastated' after being accused of cultural vandalism

Sprache [Am Gwëntext werkeln]

Das ist alles kein Bairisch, z. B. *Da ehemalige Generoi Gerhard Graf vo Schwerin wurde am 24. Mai 1950 Konrad Adenauers "Beroda in technischn Frong da Sicherheit" zua geheima Voabereitung des Aufbaus westdeitscha Streitkräfte.* oder *Am 22. Mai 1956 drod de mid grousa Mehrheit beschlossene Wehrvafassung (Eagänzung des Grundgesetzs Art. 87a GG) in Kroft, am 1. Obril folgte des Gsetz üba de Rechtsstäiung des Soidotn und am 21. Juli des Wehrpflichtgesetz.* usw. --Holder (dischkrian) 08:03, 16. Jen. 2019 (CET)

Das das alles kein Bairisch ist, würde ich nicht sagen, Holder. Aber man muss es verbessern. Vor allem muss man den Genitiv ersetzen und das Präteritum, und einige Wörter. Ich werde da mithelfen. --Muadabuali (dischkrian) 21:59, 16. Jen. 2019 (CET)

Brooks/Hern, The Guardian, 2020
bar.wikipedia.org/wiki/Dischkrian:Bundeswehr

Vorwiegend gesprochene Varietäten & Verschriftlichungen

- Normalisierung (eng verwandte Standardsprache)

Etter litt godsnakk kom tre av kyrne …	NB Tale
[After some coaxing, three of the cows came …]	*Norwegisch*
können sie ihre jugendzeit beschreiben	ArchiMob
[Can you describe your youth?]	*Schweizerdeutsch*

Vorwiegend gesprochene Varietäten & Verschriftlichungen

- Normalisierung (eng verwandte Standardsprache)
- Phone[m/t]ische Verschriftlichungen

ˈˈ{t@4 lˈit gˈˈu:snAkk kˈOm t4ˈe: ˈA:v CˈyːnˈG@ …
Etter litt godsnakk kom tre av kyrne … NB Tale
[After some coaxing, three of the cows came …] *Norwegisch*

chönd sii iri jugendziit beschriibe
können sie ihre jugendzeit beschreiben ArchiMob
[Can you describe your youth?] *Schweizerdeutsch*

Vorwiegend gesprochene Varietäten & Verschriftlichungen

- Normalisierung (eng verwandte Standardsprache)
- Phone[m/t]ische Verschriftlichungen
- (Mehr oder weniger verbreitete) Orthographien

Nu leyt em de böyse vynd disse nacht … UD LSDC
[Now, this night, the wicked enemy let them…] *Niederdeutsch*

Vorwiegend gesprochene Varietäten & Verschriftlichungen

- Normalisierung (eng verwandte Standardsprache)
- Phone[m/t]ische Verschriftlichungen
- (Mehr oder weniger verbreitete) Orthographien
- Ad-hoc-Verschriftlichungen

Nu leit em de baise Find düse Nacht …
Nu leyt em de böyse vynd disse nacht … UD LSDC
[Now, this night, the wicked enemy let them…] *Niederdeutsch*

Vorwiegend gesprochene Varietäten & Verschriftlichungen

- Normalisierung (eng verwandte Standardsprache)
- Phone[m/t]ische Verschriftlichungen
- (Mehr oder weniger verbreitete) Orthographien
- Ad-hoc-Verschriftlichungen

$\rightarrow$ Ein NLP-System, das gut für eine bestimmte Art der schriftlichen Darstellung funktioniert, funktioniert nicht unbedingt auch gut für andere!

Vorwiegend gesprochene Varietäten & Verschriftlichungen

Sprecher:innen können sich auch in ihren Meinungen bzgl. Verschriftlichung unterscheiden!

Beispiele aus Wikis in Dialekten/Regionalsprachen:

De Brukers vun de Wikipedia op Plattdüütsch hebbt utmaakt, dat se de **Sass-Schrievwies** na dat Wöörbook vun Johannes Sass (kiek ok ünner Wikipedia:Wöörböker) bruken doot.

Jrundsätzlich [der Quälltäx ändere]

Jeder schriev, wie em de Fingere jewaaße sin.

Datenknappheit & -überschneidungen

Bairische Wikipedia

- In den Daten, auf denen das mBERT-Sprachmodell trainiert wurde*
- Im bairischen Eigennamenerkennungs-Datensatz (BarNER)
- In der bairischen Syntax-Baumbank (MaiBaam)

Schweizerdeutsche Datensätze mit Wortklassen-Annotationen

- Überschneidungen zwischen Datensätzen: NOAH und UZH Universal Dependencies

*großes Problem bei der Evaluierung von LLMs wie ChatGPT: wurden die Sprachmodelle bereits auf den Testdaten (samt Lösungen) trainiert?

Empfehlungen

Blaschke, Schütze & Plank (NoDaLiDa 2023)

"A survey of corpora for Germanic low-resource languages and dialects"

… bei der *Verwendung* von Dialekt-/LRL-Datensätzen

- Immer die Qualität prüfen!
  - Offensichtliche Probleme? In der richtigen Sprache? Vermutlich von tatsächlichen Sprecher:innen?
- Ist die schriftliche Darstellung für meine Zwecke geeignet?
- Überschneiden sich die Trainings- und Evaluierungsdaten?
- Nach Daten aus verschiedenen Fachrichtungen schauen (NLP, Sprachwissenschaft, …)

Empfehlungen

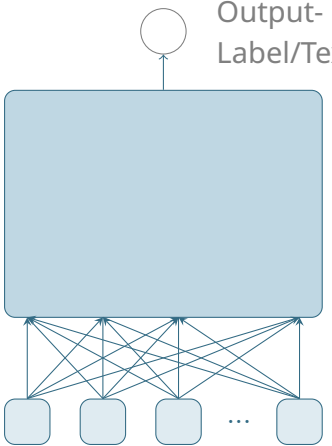
Blaschke, Schütze & Plank (NoDaLiDa 2023)

"A survey of corpora for Germanic low-resource languages and dialects"

… bei der *Erstellung* von Dialekt-/LRL-Datensätzen

- Die Art der schriftlichen Darstellung dokumentieren
- Metadaten festhalten (Größe des Datensatzes, Datenquellen, Annotationsverfahren, Lizenz und Zugriffsbedingungen)
- Archive für Langzeitarchivierung verwenden (CLARIN, LRE Map, Zenodo, …)

Übersicht



Welche NLP-Systeme & warum?

Nichtstandardsprachliche Daten modellieren

Dialect data

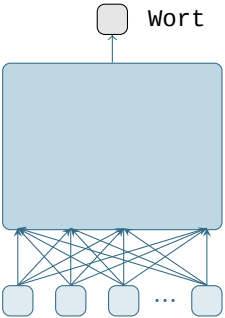
Sprachliche Unterschiede

Unterschiede zur Standardsprache

- **Aussprache ($\rightarrow$ Schreibung)**
- Wortschatz
- Morphology
- Syntax
- Verwendungssituation

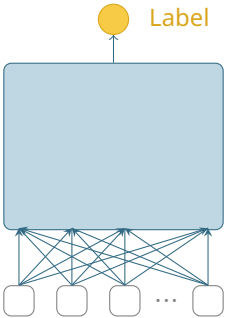
Transfer über Dialekte hinweg

✗ Pretraining



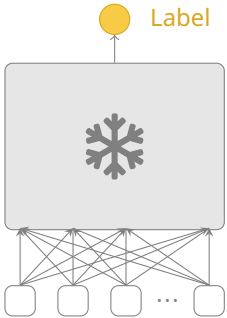
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit,

✗ Finetuning



Standardsprachlicher Inputtext für eine NLP-Aufgabe

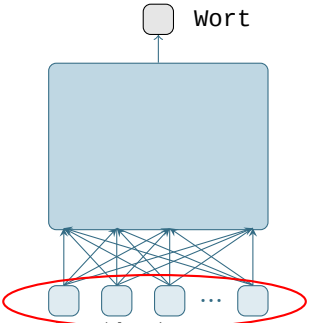
✓ Transfer



Text in einem verwandten Dialekt

Transfer über Dialekte hinweg

✗ Pretraining



Kodierung des Eingabetexts

Text wird in häufig vorkommende Buchstabenfolgen unterteilt
– “subword tokens” –
diese werden in Zahlenvektoren konvertiert

Mundartschreibweisen + Tokenisierung

Einteilung in Subword-Tokens mit GBERT

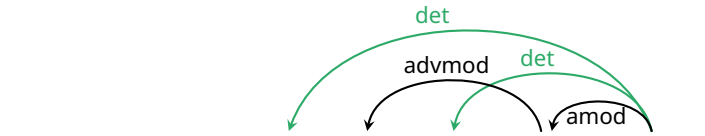
Die	Lammer	hat	ein	recht		sauberes	Wasser
`Die`	`Lamm` `-er`	`hat`	`ein`	`recht`		`sauber` `-es`	`Wasser`
D’	Lomma	hod	a	rechd	a	sauwas	Wossa
`D` `'`	`Lom` `-ma`	`ho` `-d`	`a`	`rech` `-d`	`a`	`sau` `-was`	`Wo` `-ssa`

Unterschiedliche Tokens
→ unterschiedliche Repräsentationen im Sprachmodell

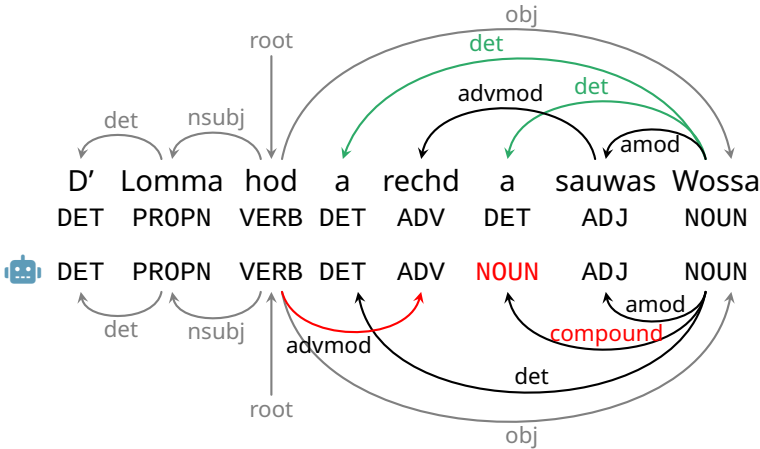
Die Tokenisierung bei ChatGPT & Co funktioniert ähnlich

Satz via `bar.wikipedia.org/wiki/Låmma`
GBERT: Chan+, COLING 2020, “German’s next language model”

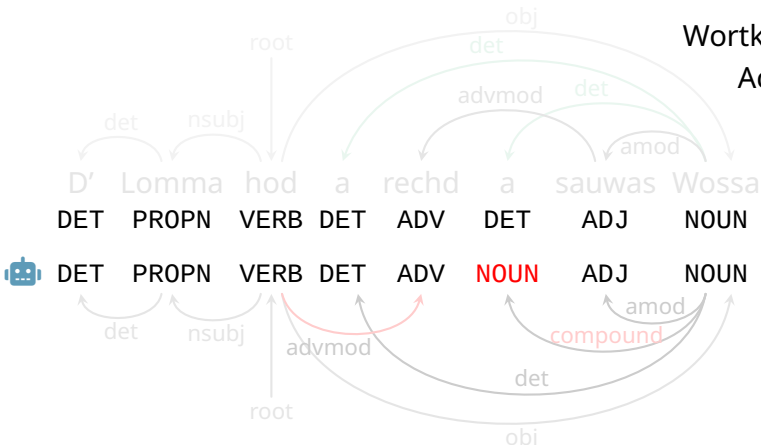
Mangelnde Robustheit von Sprachmodellen



Mangelnde Robustheit von Sprachmodellen



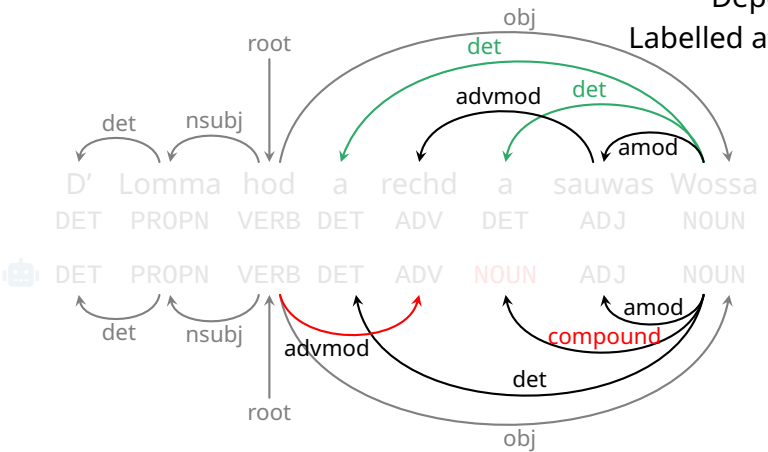
Mangelnde Robustheit von Sprachmodellen



Wortklassen:
Accuracy

Mangelnde Robustheit von Sprachmodellen

Dependenzen:
Labelled attachment score



Automatische Wortklassenbestimmung & Parsing

Blaschke, Kovačić, Peng, Schütze & Plank (LREC–COLING 2024) "MaiBaam"

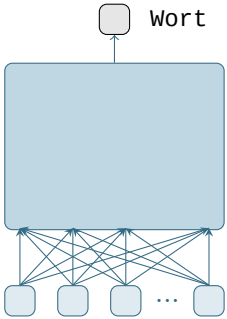
Trainingsdaten: Hochdeutsch, Testdaten: Hochdeutsch vs. Bairisch

Modell	Testsprache	Acc (%)	LAS (%)	Inputrepräsentationen
Stanza	DEU	95.9	83.7	
GBERT	DEU	96.8	83.1	
UDPipe	DEU	96.5	84.9	
Stanza	BAR	40.9	23.1	Ganze Wörter
GBERT	BAR	57.4	30.1	Subword-Tokens
UDPipe	BAR	80.5	67.3	Subword-Tok. + Buchstaben

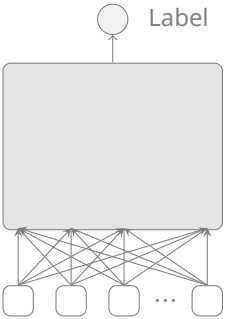
Acc = accuracy (Wortklassen); LAS = labelled attachment score
(Die Modelle haben auch noch weitere Unterschiede)

Robustere Sprachmodelle?

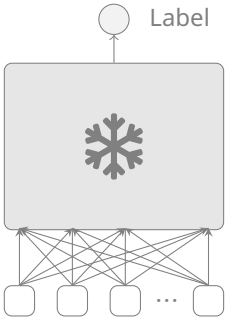
Pretraining



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit,

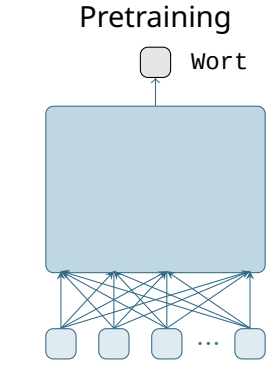


Standardsprachlicher Inputtext für eine NLP-Aufgabe



Text in einem verwandten Dialekt

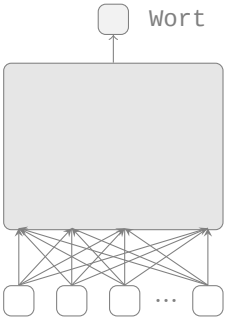
Robustere Sprachmodelle?



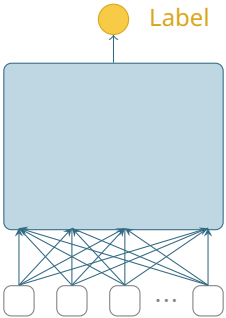
- Trainingsdaten diverser machen
  - Gezielt nach Daten in LRLs suchen
  - Künstliche Daten generieren, die den LRL-Daten ähneln
- Andere Arten von Inputrepräsentationen statt Subword-Tokens (z.B. Buchstaben)
  - Funktioniert oft besser bei Dialektdaten, aber teils schlechter bei Standardsprachdaten (+ mehr Rechenleistung nötig)
- Modell muss erst trainiert werden, bevor man weiß, ob es besser ist

Robustere Sprachmodelle?

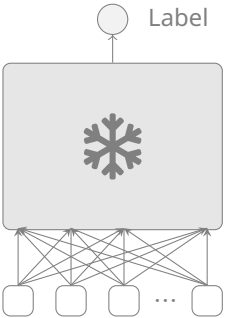
Finetuning



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit,



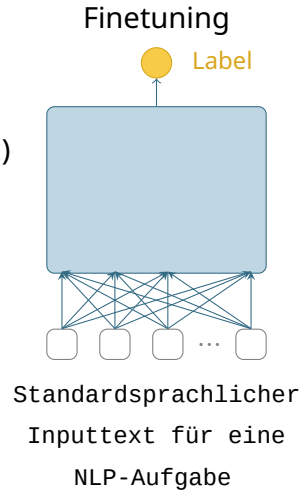
Standardsprachlicher Inputtext für eine NLP-Aufgabe



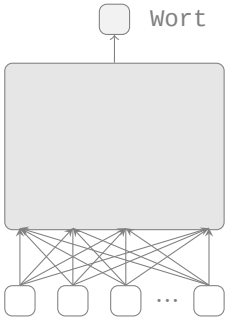
Text in einem verwandten Dialekt

Robustere Sprachmodelle?

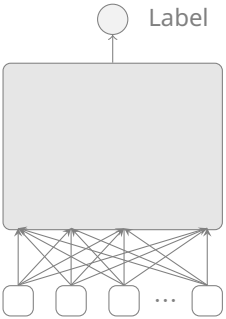
- Finetuning-Daten im Ziel-Dialekt erstellen (aufwändig!)
  Vorhandene Finetuning-Daten dem Dialekt ähnlicher machen (z.B. mit Hilfe eines Wörterbuchs)



Robustere Sprachmodelle?

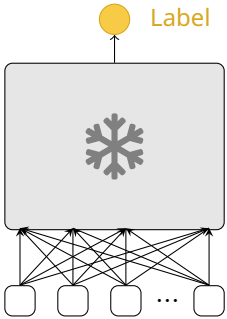


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit,



Standardsprachlicher Inputtext für eine NLP-Aufgabe

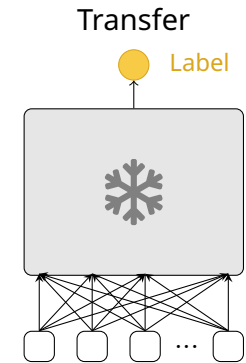
Transfer



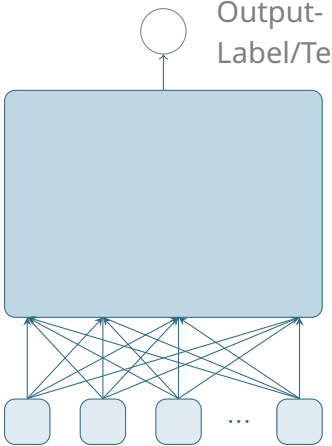
Text in einem verwandten Dialekt

Robustere Sprachmodelle?

- Dialektdaten den standardsprachlichen Daten ähnlicher machen (Normalisierung)
  - Kann schwieriger sein, als man erwartet
  - Informationen gehen verloren (Satzbau, Wortwahl, Verwendungskontext, …)



Übersicht

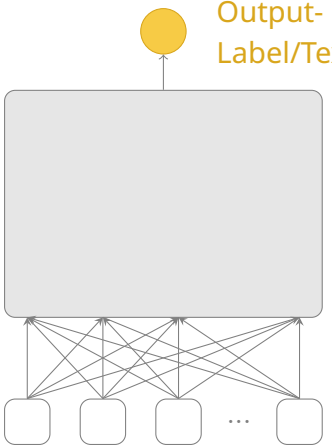


Welche NLP-Systeme & warum?

Nichtstandardsprachliche Daten modellieren

Dialect data

Übersicht



Welche NLP-Systeme & warum?

Nichtstandardsprachliche Daten modellieren

Dialect data

Sprachliche Unterschiede

Unterschiede zur Standardsprache

- Aussprache ($\rightarrow$ Schreibung)
- Wortschatz
- Morphology
- Syntax
- **Verwendungssituation**

Warum Dialekt-NLP?

Warum, wenn Dialektsprecher:innen auch die Standardsprache sprechen/verstehen?

- Sprachwissenschaft
- Maschinelles Lernen
- Anwendungen
  - Industrie-Perspektive (Untertitelung, sprachgesteuerte Navis, Meinungen zum Produkt auf sozialen Medien analysieren, ...)
  - Meinungen der Sprecher:innen

What Do Dialect Speakers Want?
A Survey of Attitudes Towards Language Technology for German Dialects

Verena Blaschke **Christoph Purschke** **Hinrich Schütze** **Barbara Plank**

Motivation

Sprachtechnologien – NLP-Anwendungen

- Maschinelle Übersetzung (machine translation, MT)
- (Text-basierte) Chatbots
- (Audio-basierte) virtelle Assistenten
- Automatische Verschriftlichung
  (automatic speech recognition, ASR)
- Audio-Synthese (text-to-speech, TTS)
- Suchmaschinen
- Rechtschreibkorrektur

Teils schon entsprechende Forschung zu deutschen Dialekten

Forschungsfragen

1. Welche (hypothetischen) Dialekt-Technologien finden Sprecher:innen besonders nützlich?
2. Hängt das davon ab, …
   - ob der Dialekt die Eingabe- oder Ausgabesprache darstellt?
   - ob Text oder gesprochene Sprache verwendet wird?
3. Wie hängen die Ergebnisse mit soziolinguistischen Faktoren zusammen?

Fragebogen

- Zielgruppe:
  Sprecher:innen deutscher Dialekte oder von verwandten Regionalsprachen
- 3 Wochen
- Über Bekannte, soziale Medien, Email-Verteiler, Heimatpflege- und Dialektvereine

Fragen

- Teil I: zum Dialekt / zur Regionalsprache
- Teil II: zu Einstellungen gegenüber Sprachtechnologien für den Dialekt / die Regionalsprache

Fragebogen

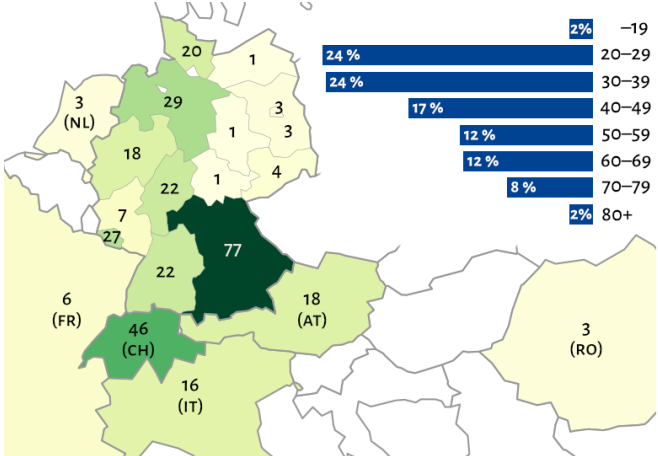
G310

20. Stimmen Sie den folgenden Aussagen zu?

Es sollte Transkriptionsprogramme geben, …	Ja, unbedingt	Eher ja	Weder noch	Eher nein	Nein, das halte ich nicht für sinnvoll	Das kann ich nicht bewerten
... die Audioaufnahmen in meinem Dialekt als geschriebenes Hochdeutsch wiedergeben.	○	○	○	○	○	○
... die Audioaufnahmen in meinem Dialekt als geschriebenen Dialekt wiedergeben.	○	○	○	○	○	○

Dialekthintergrund und -einstellungen

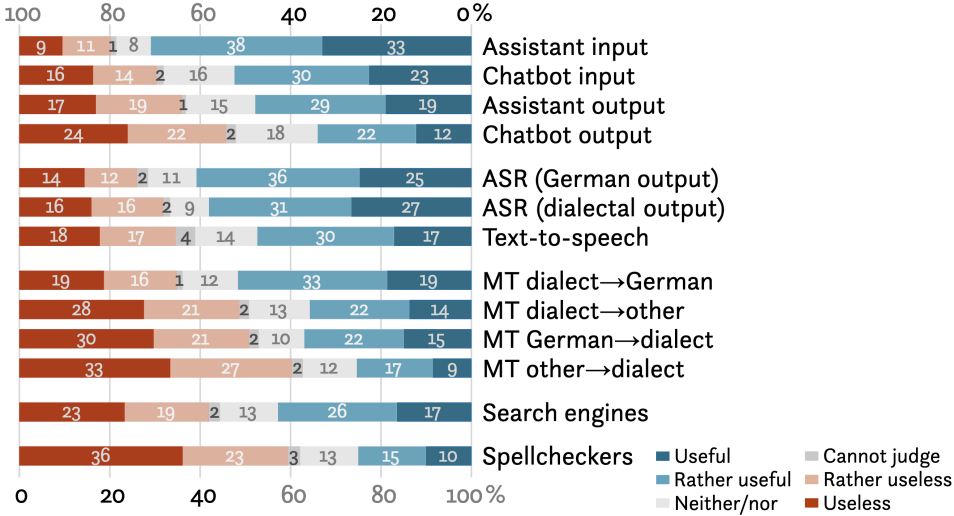
327 Gewährspersonen, die einen deutschen Dialekt (bzw. Regionalsprache) sprechen und den Fragebogen ganz ausgefüllt haben



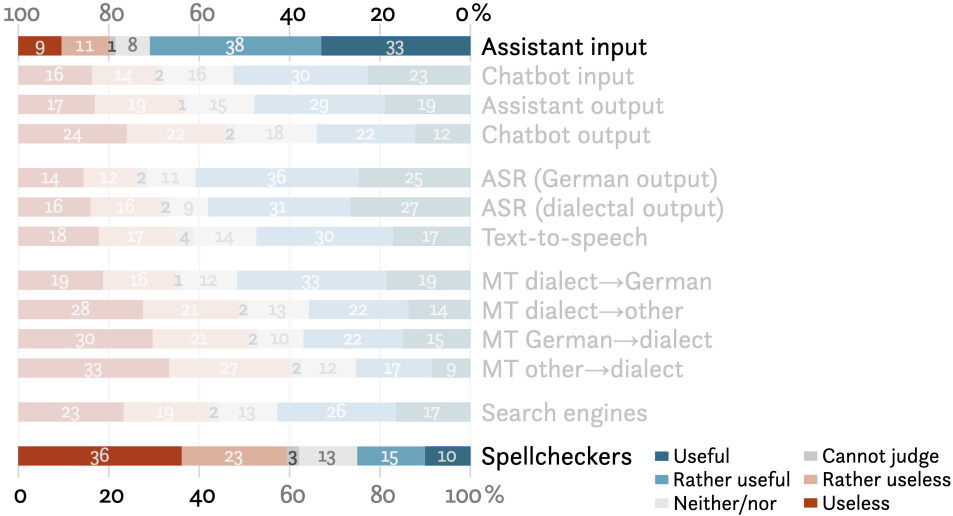
Dialekthintergrund und -einstellungen

- 52 % sprechen täglich Dialekt
- 65 % gegen eine Orthographie für ihren Dialekt
- 66 % schreiben ihren Dialekt (zumindest ab und zu)
- 35 % setzen sich für den Erhalt ihres Dialekts ein
  - im Dialektpflegeverein (13 %), Lehrkräfte, Dialektolog:innen, …
  - Verwendung des Dialekts in der Öffentlichkeit, mit Kindern
- 14 % kennen schon Sprachtechnologien für ihren Dialekt

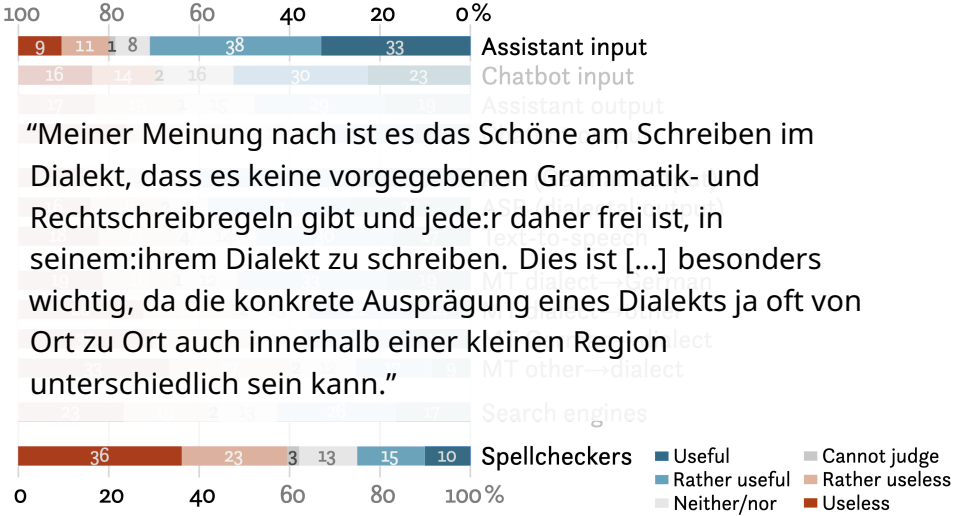
Meinungen: nützliche/unnütze Dialekt-Technologien



Meinungen: nützliche/unnütze Dialekt-Technologien

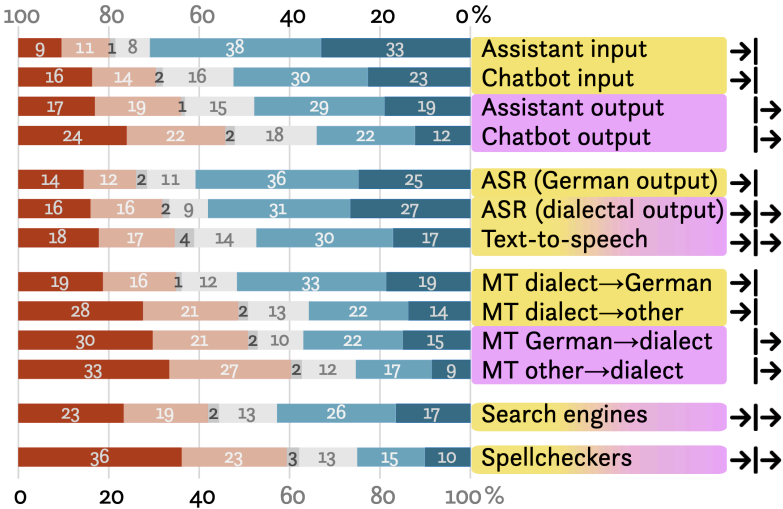


Meinungen: nützliche/unnütze Dialekt-Technologien

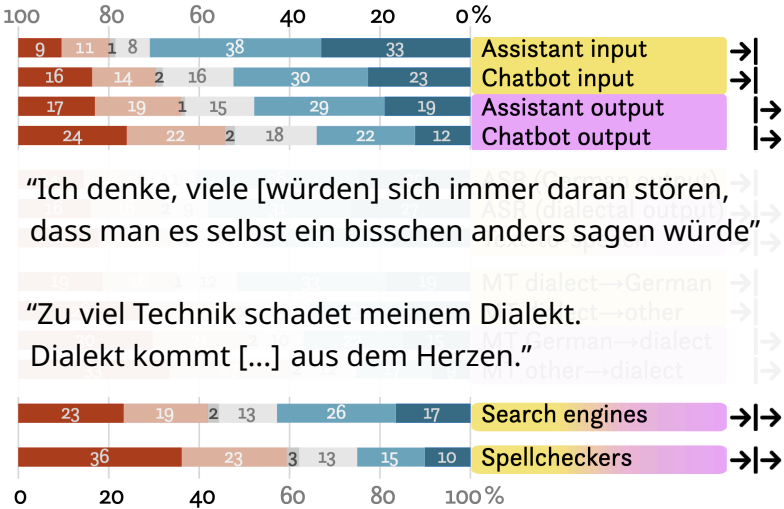


"Meiner Meinung nach ist es das Schöne am Schreiben im Dialekt, dass es keine vorgegebenen Grammatik- und Rechtschreibregeln gibt und jede:r daher frei ist, in seinem:ihrem Dialekt zu schreiben. Dies ist [...] besonders wichtig, da die konkrete Ausprägung eines Dialekts ja oft von Ort zu Ort auch innerhalb einer kleinen Region unterschiedlich sein kann."

Ein- oder Ausgabesprache?



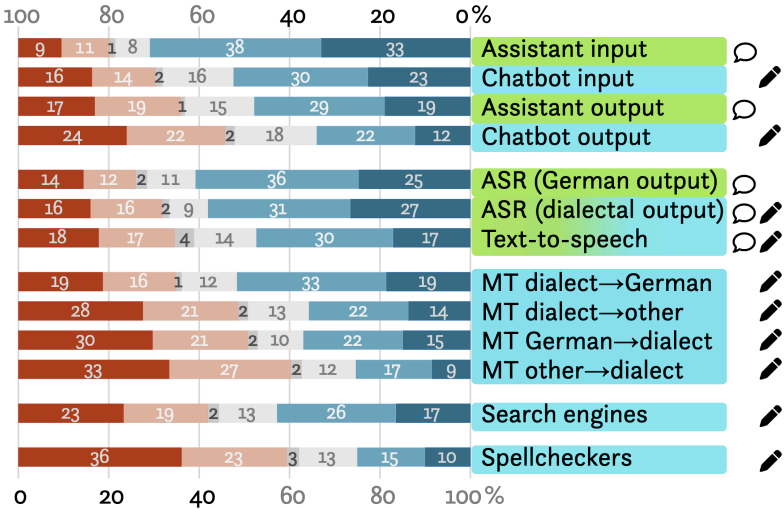
Ein- oder Ausgabesprache?



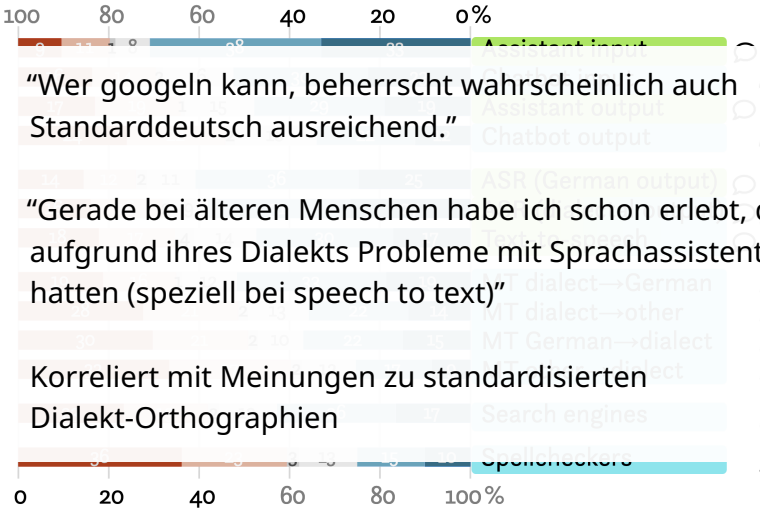
“Ich denke, viele [würden] sich immer daran stören, dass man es selbst ein bisschen anders sagen würde”

“Zu viel Technik schadet meinem Dialekt. Dialekt kommt [...] aus dem Herzen.”

Gesprochen oder geschrieben?



Gesprochen oder geschrieben?



“Wer googeln kann, beherrscht wahrscheinlich auch Standarddeutsch ausreichend.”

“Gerade bei älteren Menschen habe ich schon erlebt, dass sie aufgrund ihres Dialekts Probleme mit Sprachassistenten hatten (speziell bei speech to text)”

Korreliert mit Meinungen zu standardisierten Dialekt-Orthographien

Soziolinguistische Faktoren

“Sprachaktivist:innen” (Bemühungen zum Spracherhalt)

- Besonders positive Einstellungen gegenüber Dialekt-Technologien

❗ Das Entfernen der Antworten der Aktivist:innen hat kaum Auswirkungen auf die Rangfolge der bevorzugten Sprachtechnologien

(Sozio)linguistische Faktoren

Alter

- Sehr wenige statistisch signifikante Korrelationen!
- Jüngere Gewährspersonen: besonders interessiert an den insgesamt beliebtesten Sprachtechnologien

Dialekt“stärke”

- Gewährspersonen mit besonders starken Dialekten sind besonders daran interessiert, dass dialektaler Output genau ihrem Dialekt entspricht
- Ansonsten kaum Effekte

Soziolinguistiche Faktoren (Region/Sprache)

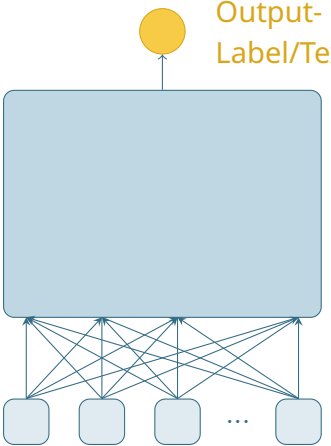
- Niederdeutsch/Plattdeutsch
  - Regionalsprache
  - Sprachlich weiter entfernt
  - Maßnahmen zum Spracherhalt
  - 👍 Technologien allgemein
  - 👍 Orthographien + Rechtschreibkorr.
- Mittel-/Süddeu. + Österreich
  - Teils durch Regiolekte ersetzt
- Schweizerdeutsch
  - Hohes Prestige
  - Starke Diglossie
  - 👎 Orthographien + Rechtschreibkorr.
  - 👍 Gesprochener Dialekt-Input

Takeaways

Blaschke, Purschke, Schütze & Plank (ACL 2024)
“What do dialect speakers want?”

- Interesse an Sprachtechnologien, die mit dialektalem **Input** umgehen können, insbesondere wenn er **gesprochen** ist
- Soziolinguistische Faktoren erklären Teile der Antworten (aber Meinungen können von den demographischen Untergruppen abweichen)
- Wichtig, auf die Wünsche & Bedürfnisse der relevanten Sprechergruppen einzugehen!

Übersicht – Herausforderungen & Herangehensweisen



Welche Dialekt-NLP-Systeme? Für wen & warum?

- Perspektiven der Sprecher:innen

Nichtstandardsprachliche Daten modellieren

- Kreative Ansätze

Datenverfügbarkeit & -qualität

- Auf welche Sprachvarietäten wird in Forschung & Produkten Rücksicht genommen?
- Wie zuverlässig und verallgemeinerbar sind die Daten?

Weitere Literatur – Dialekt-NLP

Allgemein:

- Natural language processing for similar languages, varieties, and dialects: A survey (Zampieri+, Natural Language Engineering 2020)
- Quantifying the Dialect Gap and its Correlates Across Languages (Kantharuban+, EMNLP Findings 2023)
- DIALECTBENCH: An NLP Benchmark for Dialects, Varieties, and Closely-Related Languages (Faisal+, ACL 2024)
- Variation is the Norm: Embracing Sociolinguistics in NLP (Lutgen+, LREC 2026)

Konkrete Ansätze:

- Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise (Aepli/Sennrich, ACL Findings 2022)
- Multi-VALUE: A Framework for Cross-Dialectal English NLP (Ziems+, ACL 2023)

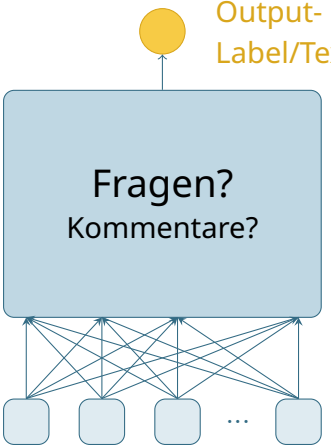
Weitere Literatur – NLP für ressourcenarme Sprachen

- NLP systems for low resource languages – hype vs. reality (Panel discussion, PML4DC @ ICLR 2023)
- Language Varieties of Italy: Technology Challenges and Opportunities (Ramponi, TACL 2024)
- Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets (Kreutzer+, TACL 2022)
- Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models (Ahia+, EMNLP 2023)
- A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios (Hedderich+, NAACL 2021)
- The State and Fate of Linguistic Diversity and Inclusion in the NLP World (Joshi+, ACL 2020)
- Evaluating the Diversity, Equity, and Inclusion of NLP Technology: A Case Study for Indian Languages (Khanuja+, EACL Findings 2023)

Weitere Literatur – Technologien & Sprachgemeinschaften

- Not always about you: Prioritizing community needs when developing endangered language technology (Liu+, ACL 2022)
- What a Creole Wants, What a Creole Needs (Lent+, LREC 2022)
- Local Languages, Third Spaces, and other High-Resource Scenarios (Bird, ACL 2022)
- Ethical Considerations for Machine Translation of Indigenous Languages: Giving a Voice to the Speakers (Mager+, ACL 2023)
- Language Technologies as If People Mattered: Centering Communities in Language Technology Development (Markl+, LREC-COLING 2024)
- My LLM might Mimic AAE [African American English] – But When Should It? (Sandoval+, NAACL 2025)

Fragen & Kommentare



Welche Dialekt-NLP-Systeme? Für wen & warum?

- Perspektiven der Sprecher:innen

Nichtstandardsprachliche Daten modellieren

- Kreative Ansätze

Datenverfügbarkeit & -qualität

- Auf welche Sprachvarietäten wird in Forschung & Produkten Rücksicht genommen?
- Wie zuverlässig und verallgemeinerbar sind die Daten?