Angewandte Linguistik
ILC Institute of
Language Competence

Hapaxe, Morpheme und Produktivität im COVID-19-Diskurs

Philipp Dreesen, Julia Krasselt, Klaus Rothenhäusler

Online Tagung DISKURSMORPHOLOGIE
18./19.03.2021

Aufbau des Vortrags

1 Einleitung

{Diskurs}{morphologie}

- Fragestellung: Kann man die Produktivität von Morphemen in Diskursen berechnen?
- Grundlegende Idee: Produktivität als Hinweis auf Grad der Diskursdynamik
- Methodologische Idee: Komplementärer Fokus auf nicht-lexikalische Einheiten, z.B. {un-}, {-heit}
- Verfahren: Berechnung von Produktivitätsmassen nach Harald Baayen
- Hintergrund: SNF-Projekt der Angewandten Diskurslinguistik zum Schweizer COVID-19-Diskurs

2.1 Analyse: Vorgehen

Produktivität und Korpusvergleich

- Kontrastive Untersuchung
  - Datengrundlage: Wort des Jahres (WDJ) Korpus 2020
    - Medienkorpus: alle grossen überregionalen und einige regionale Publikationen
    - webbasiert
    - enthält Texte der letzten fünf Jahre
  - Vergleich zwischen 2019 und 2020

- datengeleitetes Vorgehen
  - keine Vorauswahl
  - alle morphologischen Kategorien werden betrachtet
  - Identifikation von signifikanten Unterschieden

2.2 Analyse: Daten und Annotation

Korpuskennzahlen

Korpus	Dokumente	Tokens
WDJ_2019	178.000	91 Mio
WDJ_2020	160.000	78 Mio

- *"variable-corpus"*-Ansatz nach Gaeta und Ricca (2006)
  - identische Tokenanzahl pro Kategorie aus beiden Korpora

Morphologische Analyse

- Finite State Transducer: Zmorge (Sennrich/Kunz 2014)
  - Aktualität durch Automatische Generierung aus Wiktionary (Stand 1.2.2021)
  - Ambige Ausgabe

2.2 Analyse: Daten und Annotation

Heuristische Disambiguierung

```
> Infizierte
<CAP>infizier<~>en<+V><3><Sg><Past><Subj>
<CAP>infizier<~>en<+V><3><Sg><Past><Ind>
<CAP>infizier<~>en<+V><1><Sg><Past><Subj>
<CAP>infizier<~>en<+V><1><Sg><Past><Ind>
<CAP>infiziert<+ADJ><Pos><NoGend><Acc><Pl><St>
<CAP>infiziert<+ADJ><Pos><NoGend><Nom><Pl><St>
<CAP>infiziert<+ADJ><Pos><Fem><Acc><Sg>
<CAP>infiziert<+ADJ><Pos><Fem><Nom><Sg>
<CAP>infiziert<+ADJ><Pos><Masc><Nom><Sg><Wk>
Infiziert<~>e<+NN><NoGend><Acc><Pl><St>
Infiziert<~>e<+NN><NoGend><Nom><Pl><St>
Infiziert<~>e<+NN><Fem><Acc><Sg>
Infiziert<~>e<+NN><Fem><Nom><Sg>
```

2.2 Analyse: Daten und Annotation

Heuristische Disambiguierung **(Entfernung Flexionsmerkmale)**

```
> Infizierte
<CAP>infizier<~>en<+V><3><Sg><Past><Subj>
<CAP>infizier<~>en<+V><3><Sg><Past><Ind>
<CAP>infizier<~>en<+V><1><Sg><Past><Subj>
<CAP>infizier<~>en<+V><1><Sg><Past><Ind>
<CAP>infiziert<+ADJ><Pos><NoGend><Acc><Pl><St>
<CAP>infiziert<+ADJ><Pos><NoGend><Nom><Pl><St>
<CAP>infiziert<+ADJ><Pos><Fem><Acc><Sg>
<CAP>infiziert<+ADJ><Pos><Fem><Nom><Sg>
<CAP>infiziert<+ADJ><Pos><Masc><Nom><Sg><Wk>
Infiziert<~>e<+NN><NoGend><Acc><Pl><St>
Infiziert<~>e<+NN><NoGend><Nom><Pl><St>
Infiziert<~>e<+NN><Fem><Acc><Sg>
Infiziert<~>e<+NN><Fem><Nom><Sg>
```

2.2 Analyse: Daten und Annotation

Heuristische Disambiguierung (**Entfernung Flexionsmerkmale**)

```
> Infizierte
<CAP>infizier<~>en<+V>
<CAP>infizier<~>en<+V>
<CAP>infizier<~>en<+V>
<CAP>infizier<~>en<+V>
<CAP>infiziert<+ADJ><Pos>
<CAP>infiziert<+ADJ><Pos>
<CAP>infiziert<+ADJ><Pos>
<CAP>infiziert<+ADJ><Pos>
<CAP>infiziert<+ADJ><Pos>
Infiziert<~>e<+NN>
Infiziert<~>e<+NN>
Infiziert<~>e<+NN>
Infiziert<~>e<+NN>
```

2.2 Analyse: Daten und Annotation

Heuristische Disambiguierung (**Entfernung Flexionsmerkmale**)

```
> Infizierte
<CAP>infizier<~>en<+V>
<CAP>infiziert<+ADJ><Pos>
Infiziert<~>e<+NN>
```

2.2 Analyse: Daten und Annotation

Heuristische Disambiguierung (**Wortart TreeTagger**)

```
> Infizierte/NN
<CAP>infizier<~>en<+V>
<CAP>infiziert<+ADJ><Pos>
Infiziert<~>e<+NN>
```

2.2 Analyse: Daten und Annotation

Heuristische Disambiguierung (**Wortart TreeTagger**)

```
> Infizierte/NN
<CAP>infizier<~>en<+V>
<CAP>infiziert<+ADJ><Pos>
Infiziert<~>e<+NN>
```

2.2 Analyse: Daten und Annotation

Heuristische Disambiguierung (**Wortart TreeTagger**)

> **Infizierte/NN**

Infiziert<~>e<+NN>

2.2 Analyse: Daten und Annotation

Heuristische Disambiguierung

```
> Infizierte/NN

Infiziert<~>e<+NN>
```

- Bleiben mehrere Analysen übrig, wird die komplexeste (=meiste Morpheme) gewählt
- Gibt es mehrere komplexeste, wird die erste gewählt

2.2 Analyse: Daten und Annotation

Morphologische Kategorien

```
> Corona-Abstandsregelung
```

{Corona}-<TRUNC>Abstand<->s<#>regel<~>ung<+NN>

- Jedes Morphem bildet eine Kategorie

2.2 Analyse: Daten und Annotation

Morphologische Kategorien

```
> Corona-Abstandsregelung
```

```
{Corona}-<TRUNC>Abstand<->s<#>regel<~>ung<+NN>
```

- Jedes Morphem bildet eine Kategorie
  - Bindestrichglieder: <TRUNC>
  - Kompositaelemente: <#>
  - Fugenelemente: <->
  - andere Morphemgrenze: <~>

2.3 Analyse: Hapaxe

Baayen (2005, u.a.): *Morphological Productivity*

- Quantifizierung von Produktivität als kontinuierliche Grösse

- Produktivität einer Kategorie in Abhängigkeit von
  a) Tokenfrequenz (z.B. Anzahl Wörter im Korpus mit «Abstand.+»)
  b) Typenfrequenz (z.B. Anzahl Typen mit «Abstand.+»)
  c) Hapaxlegomena (z.B. Einmalbelege mit «Abstand.+»)

- Neubildungen finden sich in Korpora hauptsächlich unter den Hapaxen (Baayen 2009)

- Drei Produktivitätsmaße:
  - Realisierte Produktivität
  - Expandierende Produktivität
  - Potentielle Produktivität

2.3 Analyse: Hapaxe

Baayen (2005): *Morphological Productivity*

- Realisierte Produktivität:

$$\frac{\text{Anzahl Typen einer Kategorie}}{\text{Summe der Token einer Kategorie}}$$

- Kategorien mit *mehr* Typen sind *produktiver* als Kategorien mit weniger Typen

Corona-TRUNC

	2019	2020
Realisierte Produktivität	0	0.225
Beispiele	–	Anti-Corona-Verkäuferin Corona-Türgriff Corona-Skeptiker-Szene Corona-Gefahrenzone Corona-Beratung Corona-Gefühlslage EU-Corona-Zahlen Corona-Stewards Contra-Corona-Schutzmasken Corona-Verdacht ...

2.3 Analyse: Hapaxe

Not<#>

Baayen (2005): *Morphological Productivity*

- Expandierende Produktivität: Beitrag, den eine Kategorie zur Wachstumsrate des Vokabulars leistet
- *Hapax conditioned productivity*

$$\frac{\text{Anzahl Hapaxe einer Kategorie}}{\text{Anzahl Hapaxe im Korpus}}$$

- Wachstumsrate eines Vokabulars operationalisierbar durch Hapaxe

	2019	2020
Hapaxe mit Not<#>	292	299
Hapaxe gesamt	1’122’719	777’688
Expandierende Produktivität	3.8-04	2.6-04
Beispiele (Hapaxe)	Notausweichsystem Notabwurf Notausrüstung Notbündnis	Notfallvorhalteleistung Notbetreuungsangebote Notbetreuungsdienst Notbetriebsprogramm Notbetriebsregime Noteinkauf Noteinkommen Notfallbettenauslastung Notfalleingriffen Notfallfond ....

2.3 Analyse: Hapaxe

tritt<->

Baayen (2005): *Morphological Productivity*

- Potentielle Produktivität: Wachstum einer Kategorie
- *category conditioned productivity*:

$$\frac{\text{Anzahl Hapaxe einer Kategorie}}{\text{Tokenfrequenz der Kategorie}}$$

  - ➢ Hohe Tokenfrequenz & wenige Hapaxe: geringe potentielle Produktivität
  - ➢ Hohe Tokenfrequenz & viele Hapaxe: hohe Produktivität

	2019	2020
Hapaxe mit tritt<->	102	165
Tokenfrequenz tritt<->	4446	1576
Potentielle Produktvität	0.10	0.06
Beispiele (Hapaxe)	Austrittsgedanke Austrittsbegehren Austrittsgelder	Rücktrittsrhythmus Austrittsjubel Austrittsbefürworter Übertrittsberechtigung Grenzübertrittsversuche Kino-Eintrittspreis

3 Ergebnisse

Exemplarische Belege zu statistisch signifikanter Produktivität

Morphem	Realisierte P.	Potenzielle P.	Expandierend P.	Hochfrequente Token 2020
{19}	+	+	+	*Covid-19-Patienten, Covid-19-Erkrankung, Covid-19-Pandemie, Covid-19-Fälle, Covid-19-Gesetz*
{betrieb}	+	-	+	*Spielbetrieb, Verkehrsbetriebe, Trainingsbetrieb, Schulbetrieb, Flugbetrieb, Gastronomiebetriebe*

3 Ergebnisse

Exemplarische Belege zu statistisch signifikanter Produktivität

Morphem	Realisierte P.	Potenzielle P.	Expandierend P.	Hochfrequente Token 2020
{zahl}	-	-	+	Fallzahlen, Infektionszahlen, zahlen, Steuerzahler
{steck}	-	-	+	steckt, angesteckt, Ansteckung, ansteckend, einstecken
{Ein}	-	-	+	Einwohner, Einschränkungen, Einführung, Einhaltung, Einschätzung, Einkommen, Eindämmung
{Fasnacht}	-	-	+	Fasnachtsgesellschaft, Fasnachtsumzug, Fasnachtszeit, Fasnachtsveranstaltungen, Fasnachtswagen
{BAG}	-	-	+	BAG-Zahlen, BAG-Direktor, BAG-Sprecher, BAG-Empfehlungen, BAG-Vorgaben, BAG-Richtlinien, BAG-Vertreter, BAG-Vorschriften, BAG-Liste

4 Diskussion und Ausblick

Datengeleitetes Vorgehen als exploratives Vorgehen
Datengeleitetes Vorgehen als Korrektiv (Triangulation)

Herausforderung Datenaufbereitung: Morphemannotation und Disambiguierung
Herausforderung Datenanalyse: Statische Fragen u.a. zu Korpusvergleich
Herausforderung Erkenntnisinteresse: Fokussierung auf Morphemtypen
Herausforderung Interpretation: allgemeine Sprachentwicklung, Diskursdynamiken etc.

5 Literatur

Baayen, R. H. (2009). Corpus linguistics in morphology: morphological productivity. In Lüdeling, A., and Kyto, M. (Eds.) *Corpus Linguistics. An international handbook*. Mouton De Gruyter, Berlin, 900-919.

Gaeta, L. & Ricca, D. (2006). Productivity in Italian word formation: A variable-corpus approach. Linguistics. 44. 57-89.

Sennrich, R. & Kunz, B. (2014). Zmorge: A German Morphological Lexicon Extracted from Wiktionary. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014).

Lupica Spagnolo, M. (2013) Morphologische Produktivität in deutschsprachigen Texten nicht nativer Autoren. Eine korpuslinguistische Analyse. In Zeitschrift für germanistische Linguistik, 41(3). doi: 10.1515/zgl-2013-0021.

Back-up: Hapaxkurve

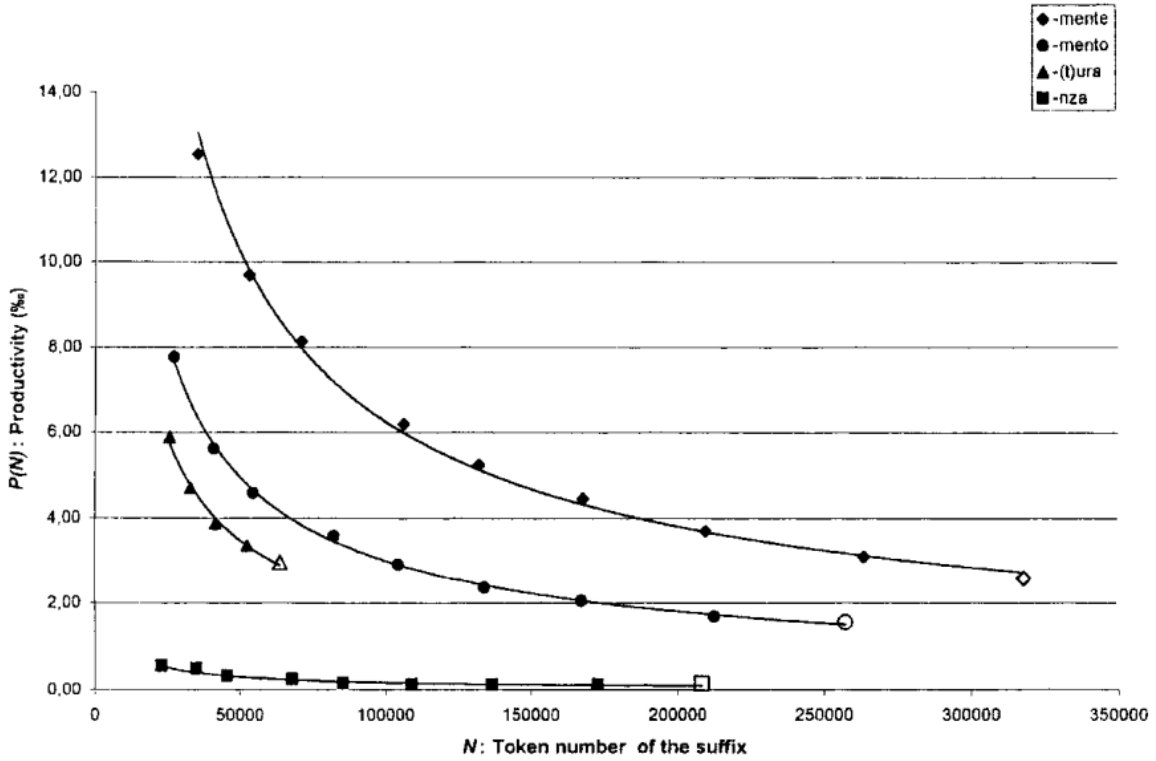

Figure 2. *Productivity as a function of* N

Aus Gaeta & Rica (2006), S. 63

Back-up: Vokabularwachstum im Covid-Korpus