Building a corpus of Austrian online forum postings on reported COVID-19 cases

Katharina Korecky-Kröll
University of Vienna, Department of German Studies
katharina.korecky-kroell@univie.ac.at

Materials

- Forum of Austrian online newspaper *derStandard.at*, article „Aktuelle Zahlen zum Coronavirus“
- Article and forum created on March 16th, 2020, posting activity still ongoing
- Until October 30th, 2021, 13:30: **244.847 forum postings**

Main aim

- Analysis of **diminutives** used for mitigation or irony (e.g., *ein bisserl* ‚a bit-DIM‘) and of **expressive compounds** (e.g., *sau+deppert* ‚sow+stupid‘) used for intensification in relation to the daily COVID-19 numbers in Austria

Open questions

- How can nonstandard forms (dialect forms, spelling or grammar errors, numbers, abbreviations, URLs etc.) be normalized best in the *derStandard.at* corpus?
- Which tools should be used for PoS and morphological tagging?

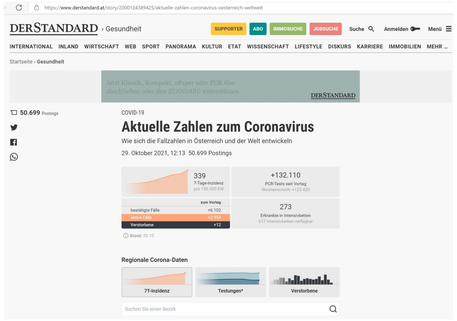


Figure 1: The online article

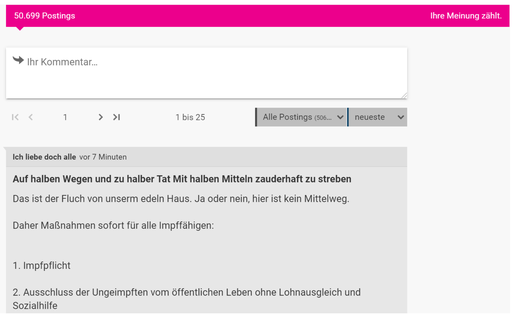


Figure 2: The forum

```
Wenn jemand ein so eigenartiges Posting absondert kann ich nicht anders. Leider!Das es Ihnen nicht passt ist eh klar.
PS: Die EMA scheint völlig irrelevant zu sein.
mir ist ganz persönlich das gimpfte und das ungimpfte gleichzeitig aufgegangen, wie ich gelesen habe, dass die wohlhabenden
Is schau guad ;)
So wie Raucher, Motorradfahrer, Alkis, Fettleibige und Extremsportler gell?
Zu Italien, Spanien und Deutschland  gesellen sich noch diese zwei Länder.
Und daran zeigt sich, was für eine verlogene Egoismus-Debatte das alles ist:Gebts mir die Impfung, mir mir mir, dann kann ich
Aus dem Archiv. https://www.derstandard.at/story/1329870225221/grippewelle-und-noroviren-wiens-spitaeler-kaempfen-mi
```

Figure 3: The postings in CSV format

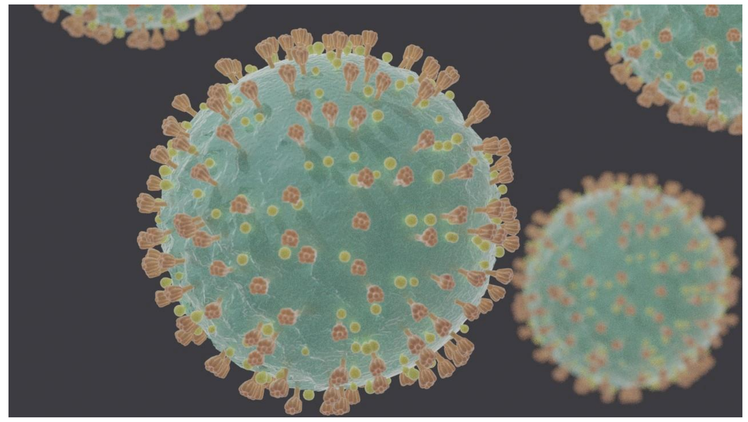
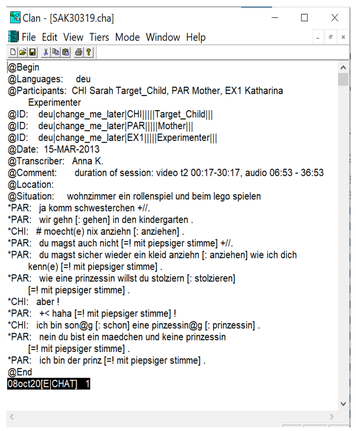


Figure 4: The topic Coronavirus SARS-CoV-2, Source: https://commons.wikimedia.org/wiki/File:Coronavirus_SARS-CoV-2.jpg, Author: Felipe Esquivel Reed (CC BY-SA 4.0): No changes were made to the original file

The methodological challenge: Tagging PoS and morphological categories

Previous work: Manually standardized transcription and semi-automatic tagging of oral child/adult data with CHILDES/CLAN (from Korecky-Kröll 2017) → but maybe not the best tool for this heterogeneous written cor

Transcribed file



Checking and tagging

- Formal CHECK of the transcript (CLAN command)
- Standardization check of nonstandard word forms with FREQ (visual check of word list)
- MOR tagging (1): creating a list of all new word forms that are not yet part of the lexicon file lex.cut
- MOR tagging (2): manual PoS and morphological tagging of word forms → copying to lex.cut → save
- MOR tagging (3): run MOR command → new version of the transcript with morphological coding tier

Disambiguation of ambiguous word forms

- Press ESC+2 to jump to the next ambiguous form
- Select the correct tag from the list of potential tags (here INF = infinitive)

```
*CHI:  # moecht(e) nix [: nichts] anziehn [: anziehen] .
%mor:  V:mod|moeg-CON:PRET:1S PRO:indef|nichts
       an#V:S|zieh-INF^an#V:S|zieh-3P^an#V:S|zieh-1P
22jul16|C|CHAT| 23 : *CHI: # moecht(e) nix [: nichts] anziehn [: anziehen] .
 an#V:S|zieh-INF
 an#V:S|zieh-3P
 an#V:S|zieh-1P
```

Tagged file (with PoS and MOR tagging)

```
*CHI:  # moecht(e) nix anziehn [: anziehen] .
%mor:  V:mod|moeg-CON:PRET:1S PRO:indef|nichts an#V:S|zieh-INF .
```

→ CLAN provides various built-in commands for corpus analyses (e.g., FREQ, MLU, VOCD, KWAL, COMBO,...), see CLAN Manual

Pilot analysis

- Based on 194.116 postings from March 16th, 2020, to April, 19th, 2021 (already in CSV format)
- **Research question**: Are there **correlations** between the daily **COVID-19 numbers** in Austria and the daily **number of postings** in the forum of this article?

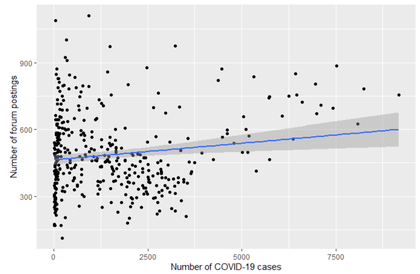


Figure 5: Correlation between COVID-19 cases in Austria and number of forum postings

Results

- Weak positive correlation between **number of COVID-19 cases** in Austria and **number of postings** (r = 0.148, df = 398, p = 0.003**), see Fig. 5
- Very weak positive correlation between **COVID-19 7-day incidence rate** in Austria and **number of postings** (r = 0.098, df = 398, p = 0.0498*)
- Overall tendency of **decreasing forum activity** in the later months of the pandemic (blue bars in Fig. 6).

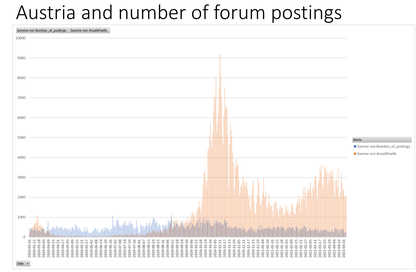


Figure 6: Absolute numbers of COVID-19 cases (in red) and forum postings (in blue)

References

Bülow, L., A. Diehr, D. Pfurtscheller & S. Thome. (in prep). *Corona-Diskurse in und über Österreich*. (Special Issue in *Wiener Linguistische Gazette*).
Essam, B.A. & M.S. Abdo (2021). How Do Arab Tweeters Perceive the COVID-19 Pandemic? *Journal of Psycholinguistic Research* 50, 507–521 (2021). https://doi.org/10.1007/s10936-020-09715-6.
Korecky-Kröll, K. (2017). Kodierung und Analyse mit CHILDES: Erfahrungen mit kindersprachlichen Spontansprachkorpora und erste Arbeiten zu einem rein erwachsenensprachlichen Spontansprachkorpus. In: C. Resch & W.U. Dressler. eds. *Digitale Methoden der Korpusforschung in Österreich*. Vienna: Austrian Academy of Sciences Press, 85-113.
MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. 3rd edition*. Mahwah: NJ: Erlbaum.
Ng, R., T.Y.J. Chow & W. Yang (2021). Culture Linked to Increasing Ageism During COVID-19: Evidence From a 10-Billion-Word Corpus Across 20 Countries, *The Journals of Gerontology: Series B*, gbab057, https://doi.org/10.1093/geronb/gbab057
Schweinberger M, Haugh M, Hames S (2021). Analysing discourse around COVID-19 in the Australian Twittersphere: A real-time corpus-based analysis. *Big Data & Society* 8/1. doi:10.1177/20539517211021437

Acknowledgements

Thanks are due to the very helpful team of *derStandard.at* for sending me the entire previous postings in CSV format for research purposes.