Using corpora for diversity linguistics

BA Language Science - WP1
Europäische Sprachen: Praxis und Variation: Projektorientiertes Arbeiten

Proseminar 2 SWS ECTS-Punkte: 5
Montag, 10-12 Uhr - R. 1.27 ----- **Mondays 11-12 MS Teams**
Lecturer: Annemarie Verkerk
First class: (06-04-2020) 04-05-2020
Last class: 13-07-2020
Examination deadlines: 14-08-2020
Moodle key:

Description

In this seminar we focus on learning how to carry out linguistic research, in this case specifically cross-linguistic research using (parallel) corpora. Through several case-studies, we work on 1) digesting scientific literature to uncover the state-of-the-art on a particular topic; 2) formulating a relevant research question; 3) studying the right method to answer the research question; 4) carrying out the research, in this case annotating and searching corpora; and 5) presenting the results in both oral and written format. Over the course of the seminar we will work with several corpora and on several topics, including verbal lexical semantics (motion, perception, ingestion, communication, placement), Komolgorov complexity, pronouns, and question marking. For the final assessment, each student works further on one of the case-studies or on their own research question, presents their work orally and writes a conference abstract.

Organization

The seminar will have both theoretical and practical parts. There are weekly readings and assignments. These can be worked on individually or in groups, as long as students clearly indicate who contributed what. Towards the end, students will pick their own topic to develop further, write a research plan about, present it orally at the end of the seminar and submit a report in form of an abstract (1-2 pages).

Requirements

Active participations in class
Weekly assignments
Oral presentation and conference abstract

This course is a Proseminar, the final exam consists of a writing a conference abstract and giving a presentation about a small research project (the project need not necessarily be finished). This course is not graded, i.e. you pass or you don't. I would recommend that you participate and do the weekly assignments, as these will build up towards writing an interesting conference abstract and giving a nice presentation.
The presentations will be done in the final class (we'll plan a longer session for that class). Conference abstracts are then due 14-08-2020.

Course Schedule

Given the shortened SoSe 2020 starts on 04-05-2020, we have 11 weeks. It just so happens that this (very short) book on research methods has 11 chapters:
Walliman, Nicholas. 2011. *Research Methods: The Basics*. London: Routledge.
We will be reading a chapter from the book each week, and watching youtube lectures on research methodology so you have different types of input.
The content on using corpora comes mostly from recent papers.
The assignments will focus on both skills in research methodology and corpus linguistics, as well as content questions.
We will have weekly discussion sessions where we discuss the assignments and any questions you have regarding the videos, readings, or assignments. I will start these sessions with the most important take-away points for that week.

	date	**class**	**reading**	**assignment**
1	06-04-2020	Introduction to linguistic typology	Goldhahn et al. (2014)	1 Organizing your knowledge
2	13-04-2020	Word order typology	Levshina (2019)	2 Word order typology
3	20-04-2020	Entropy of words	Bentz et al. (2017)	3 Entropy of words
4	27-04-2020	Nouns slow down speech	Seifart et al. (2018)	4 Nouns slow down speech
5	04-05-2020	*Research basics* Diversity linguistics	*Walliman chapter 1* Bickel (2014)	5 Diversity linguistics (+ Assignment 1)
6	11-05-2020	*Research theory* Parallel corpora	*Walliman chapter 2* Cysouw & Wälchli (2007) Goldhahn et al. (2014)	6 Parallel corpora
7	18-05-2020	*Structuring the research project* Type-Token Ratio	*Walliman chapter 3* Kettunen (2014)	7 TTR in CWB across languages
8	25-05-2020	*Research ethics* Dependency lengths	*Walliman chapter 4* Futrell et al. 2015	8 Research project structure & Obtaining data
9	01-06-2020	*Finding and reviewing the literature* Zipf's law	*Walliman chapter 5* Bentz et al. 2016 & tba	9 Zotero & R Basics
10	08-06-2020	*The nature of data* Lexical typology	*Walliman chapter 6* tba	10 Zipf's law
11	15-06-2020	*Collecting and analysing secondary data* Komolgorov complexity	*Walliman chapter 7* Ehret & Szmrecsanyi (2016) Ehret (2018)	11 Lexical typology
12	22-06-2020	*Collecting primary data* Tu/vous pronouns	*Walliman chapter 8* Levshina (2017)	12 Komolgorov complexity
13	29-06-2020	*Quantitative data analysis*	*Walliman chapter 9* Gries (2010)	13 ParTy corpus annotation
14	06-07-2020	*Qualitative data analysis* Presentation skills	*Walliman chapter 10* no paper	14 ParTy corpus analysis
15	13-07-2020	*Writing the proposal and writing up the research* Research report preparation	*Walliman chapter 11* no paper	15 Research plan PRESENTATIONS

1. **06-04-2020** Introduction

Here is the optional course work for this week:
Watching: "Introduction to linguistic typology": https://youtu.be/af2T3nTsGFI
Anke Lüdeling on "Corpus Linguistics": https://www.youtube.com/watch?v=21a-lOghoK0
Reading: Goldhahn, Dirk, Uwe Quasthoff, and Gerhard Heyer. 2014. 'Corpus-Based Linguistic Typology: A Comprehensive Approach'. In *Proceedings of Konvens 2014*, Hildesheim, Germany.
Assignment: 1 Organizing your knowledge.

2. **13-04-2020** Word order typology

The optional course work for the remaining three weeks before the week of 04-05-2020 are exercises in paper-reading. There will be a paper each week, some general questions/instructions on how to read it, and specific questions to test your understanding. The papers are the latest studies in diversity linguistics using corpora. You can send me your assignment and use the forum for asking questions.

Watching: "How to read a scientific paper":
https://www.youtube.com/watch?v=5Eg_Gzz3hXY
"Word order typology": https://www.youtube.com/watch?v=IHacpFB2kK4
Reading: Levshina, Natalia. 2019. 'Token-Based Typology and Word Order Entropy: A Study Based on Universal Dependencies'. *Linguistic Typology* 23 (3): 533–72.
Assignment: 2 Word order typology

3. **20-04-2020** Entropy

Watching: "Journey into information theory: Measuring information"
https://www.youtube.com/watch?v=PtmzfpV6CDE
"Journey into information theory: Information entropy"
https://www.youtube.com/watch?v=2s3aJfRr9gE
Reading: Bentz, Christian, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i-Cancho. 2017. 'The Entropy of Words—Learnability and Expressivity across More Than 1000 Languages', April, 1–34.
Assignment: 3 Entropy of words

4. **27-04-2020** Nouns slow down speech

Watching: "Conversation analysis": https://www.youtube.com/watch?v=amAofYfkmAw
(to have a sense of the kind of data Seifart et al. are using, and how it is frequently analyzed.)
Reading: Seifart, Frank, Jan Strunk, Swintha Danielsen, Iren Hartmann, Brigitte Pakendorf, Søren Wichmann, Alena Witzlack-Makarevich, Nivja H de Jong, and Balthasar Bickel. 2018. 'Nouns Slow down Speech across Structurally and Culturally Diverse Languages'. *Proceedings of the National Academy of Sciences* 137 (May): 201800708–6.
Assignment: 4 Nouns slow down speech

5. **04-05-2020** Research basics and diversity linguistics

Watching: "The Fight To Save The Dying Languages Of Alaska":
https://www.youtube.com/watch?v=Xn7mkEsxybw

"Languages Matter!" https://www.youtube.com/watch?v=Q-XozG0RSCo
Reading:

1. Walliman (2011). Research methods. Chapter 1: Research basics.
2. Bickel, Balthasar. 2014. "Linguistic Diversity and Universals." In *The Cambridge Handbook of Linguistic Anthropology*, edited by Nicholas J. Enfield, Paul Kockelman, and Jack Sidnell, 101–24. Cambridge: Cambridge University Press.

Assignment: 5 Diversity linguistics
Please do assignment 1 if you haven't done it yet

6. **11-05-2020** Research theory & parallel corpora

Watching: "Corpus-based linguistic typology": https://youtu.be/5ew6NFZ2CuA
Reading:

1. Walliman (2011). Research methods. Chapter 2: Research theory.
2. Wälchli, Bernhard. 2007. "Advantages and Disadvantages of Using Parallel Texts in Typological Investigations." *STUF-Sprachtypologie Und Universalienforschung* 60 (2): 118–134.
3. Goldhahn, Dirk, Uwe Quasthoff, and Gerhard Heyer. 2014. 'Corpus-Based Linguistic Typology: A Comprehensive Approach'. In *Proceedings of Konvens 2014*, Hildesheim, Germany.

Assignment: 6 Parallel corpora

7. **18-05-2020** Corpus-based cross-linguistic measures

Watching: nothing to watch
Reading:

1. Walliman 2011. Research methods. Chapter 3: Structuring the research project
2. Chapter 2 from Sinnemäki, Kaius. 2011. *Language Universals and Linguistic Complexity*. PhD Thesis, University of Helsinki. SKIP section 2.1, 2.4 and 2.5, i.e. reading from page 10 to page 32.
3. Kettunen, Kimmo. 2014. 'Can Type-Token Ratio Be Used to Show Morphological Complexity of Languages?' *Journal of Quantitative Linguistics* 21 (3): 223–45.

Assignment: 7 Corpus-based cross-linguistic measures

8. **25-05-2020** Dependency grammar

Watching: "Dependency Parsing" from 6.30 to 45.30:
https://www.youtube.com/watch?v=PVShkZgXznc
Reading:

1. Walliman 2011. Research methods. Chapter 4: Research ethics.
2. Futrell, Richard, Kyle Mahowald, and Edward Gibson. 2015. 'Large-Scale Evidence of Dependency Length Minimization in 37 Languages'. *Proc Natl Acad Sci U S A*, August, 201502134. https://doi.org/10.1073/pnas.1502134112.

Assignment: 8 Structure of a research project, obtaining data, and dependency grammar

9. **01-06-2020** Zipf's law

Watching: "Zipf's law": https://www.youtube.com/watch?v=aVmf8MKev5M
Reading:

1. Walliman 2011. Research methods. Chapter 5: Finding and reviewing the literature.

2. Piantadosi, Steven T. 2014. 'Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions'. *Psychonomic Bulletin & Review* 21 (5): 1112–1130.
3. Bentz, Christian & Ferrer-i-Cancho, Ramon (2016). Zipf's law of abbreviation as a language universal. In Bentz, Christian, Jager, Gerhard & Yanovich, Igor (eds.) Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics. Tübingen: University of Tubingen.

Assignment: 9 Zipf's law, Zotero & R Basics

10. 08-06-2020 Zipf's law & lexical typology

Watching: "A course in Cognitive Linguistics: Color": https://www.youtube.com/watch?v=ggfMQ0Zuv6o
(This isn't very close to the topic of lexical typology, but it's the closest thing I could find and anyway, it's fun.)

Reading:

1. Walliman 2011. Research methods. Chapter 6: The nature of data.
2. from last week: Bentz, Christian & Ferrer-i-Cancho, Ramon (2016). Zipf's law of abbreviation as a language universal. In Bentz, Christian, Jager, Gerhard & Yanovich, Igor (eds.) Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics. Tübingen: University of Tubingen.
3. Slobin, Dan I. 2005. 'Relating Narrative Events in Translation'. In *Perspectives on Language and Language Development: Essays in Honor of Ruth A. Berman*, edited by Dorit Diskin Ravid and Hava Bat-Zeev Shyldkrot, 115–29. Dordrecht: Kluwer.

Assignment: 10 Zipf's law & lexical typology

11. 15-06-2020 Lexical typology & Kolmogorov complexity

Watching: "Kolmogorov complexity with Python": https://www.youtube.com/watch?v=KyB13PD-UME

Reading:

1. Walliman 2011. Research methods. Chapter 7: Collecting and analyzing secondary data.
2. Ehret, Katharina, Benedikt Szmrecsanyi, Raffaella Baechler, and Guido Seiler. (2016) "An Information-Theoretic Approach to Assess Linguistic Complexity." In *Complexity, Isolation, and Variation*, 71–94. Berlin: De Gruyter.

Assignment: 11 Lexical typology in CWB & Kolmogorov complexity

12. 22-06-2020 Kolmogorov complexity & Tu/vous pronouns

Watching: "Why "No Problem" Can Seem Rude: Phatic Expressions": https://www.youtube.com/watch?v=eGnH0KAXhCw
"Politeness in Linguistics: An overview": https://www.youtube.com/watch?v=e-B-kJi0Rek

Reading:

1. Walliman 2011. Research methods. Chapter 8: Collecting primary data.
2. Levshina, Natalia. 2017. "A Multivariate Study of T/V Forms in European Languages Based on a Parallel Corpus of Film Subtitles." *Research in Language* 15 (2): 153–72.

Assignment: 12 Komolgorov complexity & Tu/vous pronouns

13. 29-06-2020 Tu/vous pronouns & statistics

Watching: Quantitative data analysis & statistics
Reading:

1. Walliman 2011. Research methods. Chapter 9: Quantitative data analysis.
2. Gries, Stefan Th, Aquilinio Sánchez Pérez, and Moisés Almala Sánchez. 2010. 'Useful Statistics for Corpus Linguistics'. In *A Mosaic of Corpus Linguistics: Selected Approaches*, 269–91. Frankfurt am Main: Peter Lang.

Assignment: 13 Tu/vous pronouns: ParTy corpus annotation

14. 06-07-2020 Presentation skills

Watching: "The surprising secret to speaking with confidence":
https://www.youtube.com/watch?v=a2MR5XbJtXU
„Mark Zuckerberg presentation skills breakdown":
https://www.youtube.com/watch?v=XqopUmIjN4I
Reading:
Walliman 2011. Research methods. Chapter 10: Qualitative data analysis.
nothing else
Assignment: 14 Tu/vous pronouns, part III: analysis / Presentation skills

15. 13-07-2020 Writing skills, PRESENTATION DAY

Watching: Writing skills
Reading:
Walliman 2011. Research methods. Chapter 11: Writing the proposal and writing up the research.
nothing else
Assignment: 15 Research plan & commiserations about writing