“The Challenges and Joys of Analysing Ongoing Language Change in Web-based Corpora: A Case Study”

Anne Krause

Leipzig University / Research Training Group GRK DFG 1624 “Frequency Effects in Language”, University of Freiburg

10th Web as Corpus Workshop (WAC-X), Humboldt University Berlin

12/08/2016

Outline

1. Project description
2. Challenges
   a. Text type selection
   b. Crawling
   c. POS-Annotation
   d. Authorship confirmation
   e. Annotation for frequency variables
   f. Annotation for further variables
   g. Meta information
3. Joys
   a. Results of Corpus Study
   b. Design and Results of Experimental Study
4. Conclusion

1a. German strong verbs with vowel gradation

(the example of *geben* ‘to give’)

Present		**Indicative**	**Imperative**
Singular	**1st**	ich gebe	
	2nd	du gibst	gib!
	3rd	er/ sie/ es gibt	
Plural	**1st**	wir geben	
	2nd	ihr gebt	
	3rd	sie geben	

1a. German strong verbs with vowel gradation

(the example of *geben* 'to give')

Present		**Indicative**	**Imperative**
Singular	**1st**	ich gebe	
	2nd	du gibst	**gib!**
	3rd	er/ sie/ es gibt	
Plural	**1st**	wir geben	
	2nd	ihr gebt	
	3rd	sie geben	

1a. Change in the Imperative Singular

established i-stem	analogical e-stem	
gib!	gebe! / geb!	‘to give’
tritt!	trete! / tret!	‘to kick/ to step’
iss!	esse! / ess!	‘to eat’
befiehl!	befehle! / befehl!	‘to command’
milk!	melke! /melk!	‘to milk’

1b. Hypothesis

Conserving Effect:

“high frequency forms with alternations resist analogical leveling: while English *weep* / *wept*, *creep* / *crept* and *leap* / *leapt* have a tendency to regularize to *weeped*, *creeped* and *leaped* respectively, the high frequency verbs with the same pattern, *keep* / *kept*, *sleep* / *slept* show no such tendency (Hooper 1976, Bybee 1985)”

(Bybee und Thompson 1997: 380)

2. Challenges

a. Text type selection

b. Crawling

c. POS-Annotation

d. Authorship confirmation

e. Annotation for frequency variables

f. Annotation for further variables

g. Meta information

2a. Text type selection

Challenge:

- in German, imperative singular forms only occur in conversation between very familiar speakers (Duden Grammar 2009: 548-550)
- many speakers rather use indicative, infinitive or modal constructions

➢ very low number of instances in even large corpora of both spoken (e.g. DGD database) and written (e.g. DeReKo) German

- high numbers in existing web-based corpora (e.g. DeWaC) but no meta information about texts (esp. timestamp) or authors

2a. Text type selection

My solution:

Walkthroughs

- guides for variety of video games
- step-by-step instructions

➢ variety of verbs used

- written by gamers for other gamers

➢ imperative singular use

Alternative solution:

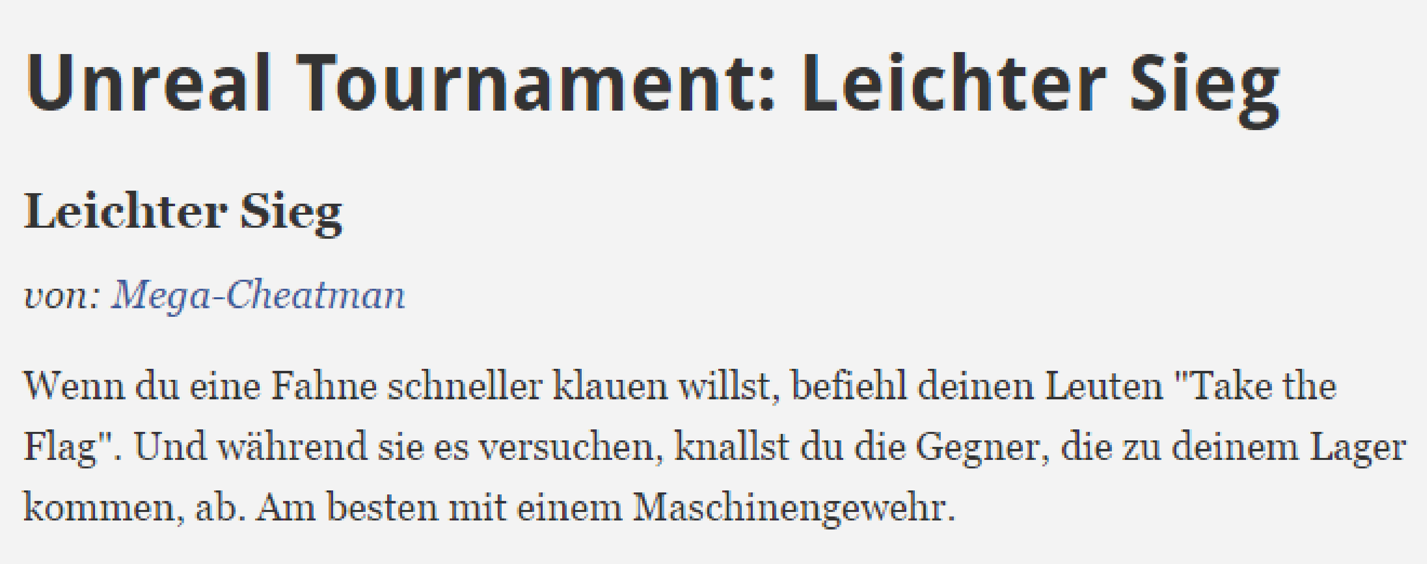
- pilot search for target word/ construction in existing corpora
- identifies text types/ textual domains with high yield rate

➢ corpus of these texts/ websites can be compiled

2b. Crawling

Challenge:

- walkthrough texts on *spieletipps.de* have timestamp
- it is not displayed with the text but on the author's profile page



Unreal Tournament: Leichter Sieg

Leichter Sieg

von: Mega-Cheatman

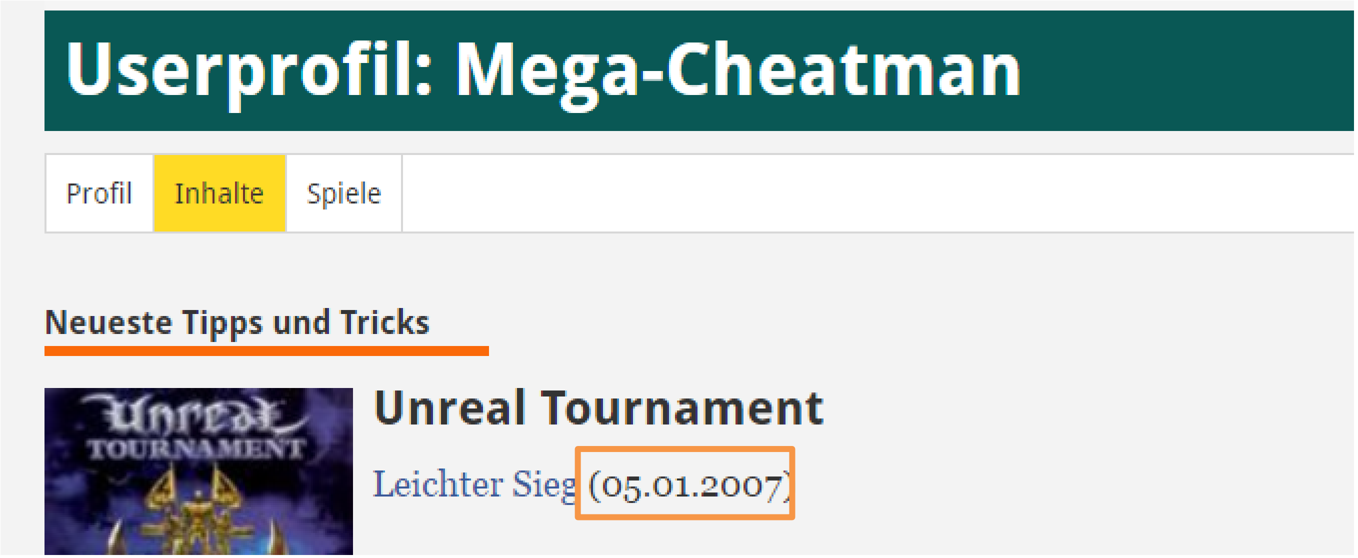
Wenn du eine Fahne schneller klauen willst, befiehl deinen Leuten "Take the Flag". Und während sie es versuchen, knallst du die Gegner, die zu deinem Lager kommen, ab. Am besten mit einem Maschinengewehr.

http://www.spieletipps.de/tipps-1297-unreal-tournament-tipps-tricks/4/#1

2b. Crawling

Challenge:

- walkthrough texts have timestamp
- it is not displayed with the text but on the author’s profile page



Userprofil: Mega-Cheatman

Profil | Inhalte | Spiele

Neueste Tipps und Tricks

Unreal Tournament

Leichter Sieg (05.01.2007)

http://www.spieletipps.de/profil/mega-cheatman/#Content

2b. Crawling

My solution:

- extrapolation of posting year from
  - release year of game
  - registration year of author on website
- manual correction for exact timestamp if available

Alternative solutions:

- close inspection of websites

➢ adjust crawler

➢ (repeat crawl)

2c. POS-Annotation

Challenge:

- POS-Annotation should speed up corpus search
- TreeTagger, STTS Tagset (vvimp vs. vvfin)
- analogical imperative variants tagged as vvfin

Present		**Indicative**	**Imperative**
Singular	**1st**	ich geb(e)	
	2nd	du gibst	gib/ geb(e)
	3rd	er/ sie/ es gibt	

2c. POS-Annotation

My solution:

- manual correction of POS-tags (analogical imperative singular variants)

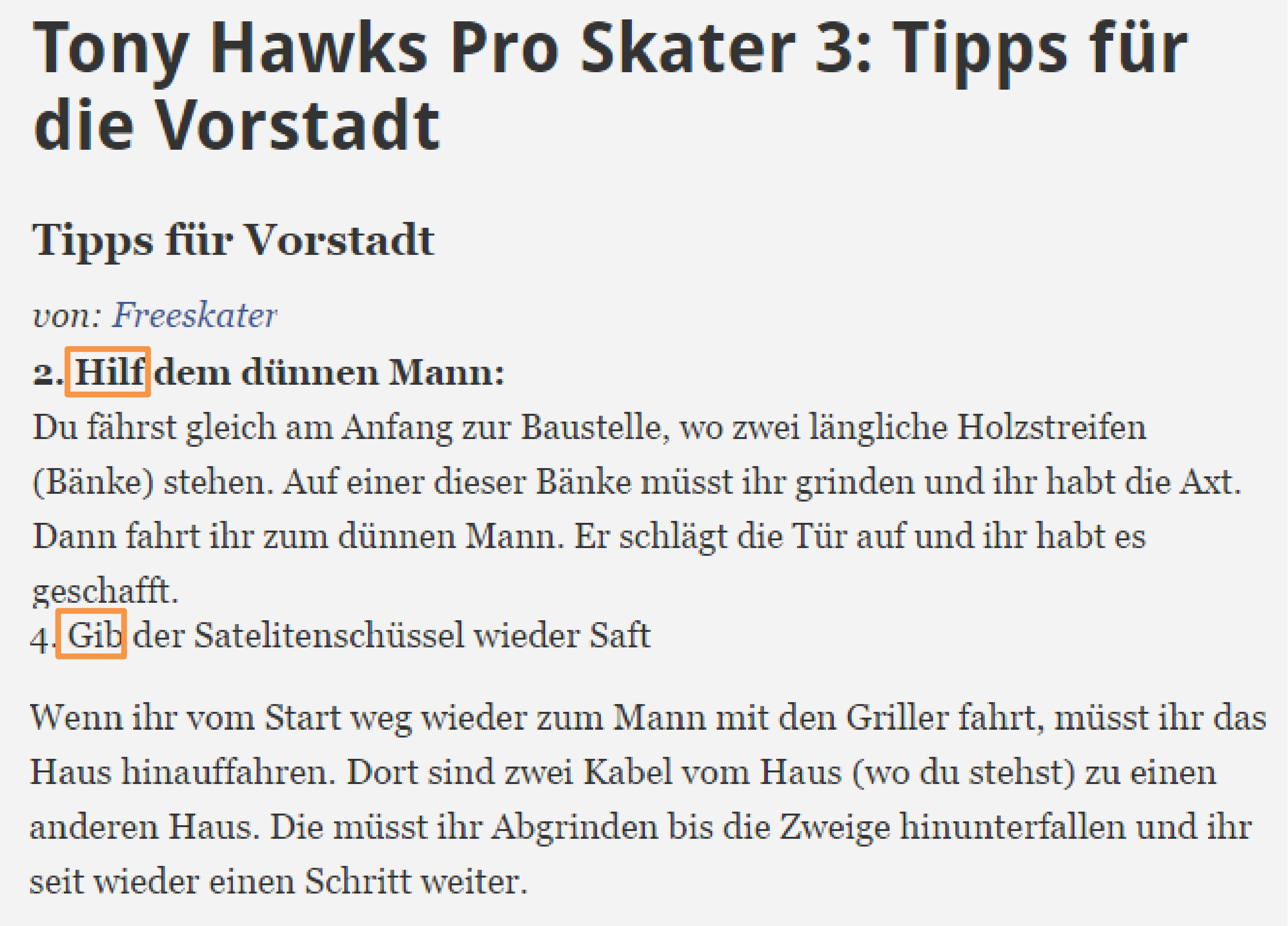
Alternative solution:

- train tagger on target forms/ constructions

2d. Authorship confirmation

Challenge:

- some imperative singular forms may have been quoted from inside a game



Tony Hawks Pro Skater 3: Tipps für die Vorstadt

Tipps für Vorstadt

von: Freeskater

2. Hilf dem dünnen Mann:

Du fährst gleich am Anfang zur Baustelle, wo zwei längliche Holzstreifen (Bänke) stehen. Auf einer dieser Bänke müsst ihr grinden und ihr habt die Axt. Dann fahrt ihr zum dünnen Mann. Er schlägt die Tür auf und ihr habt es geschafft.

4. Gib der Satelitenschüssel wieder Saft

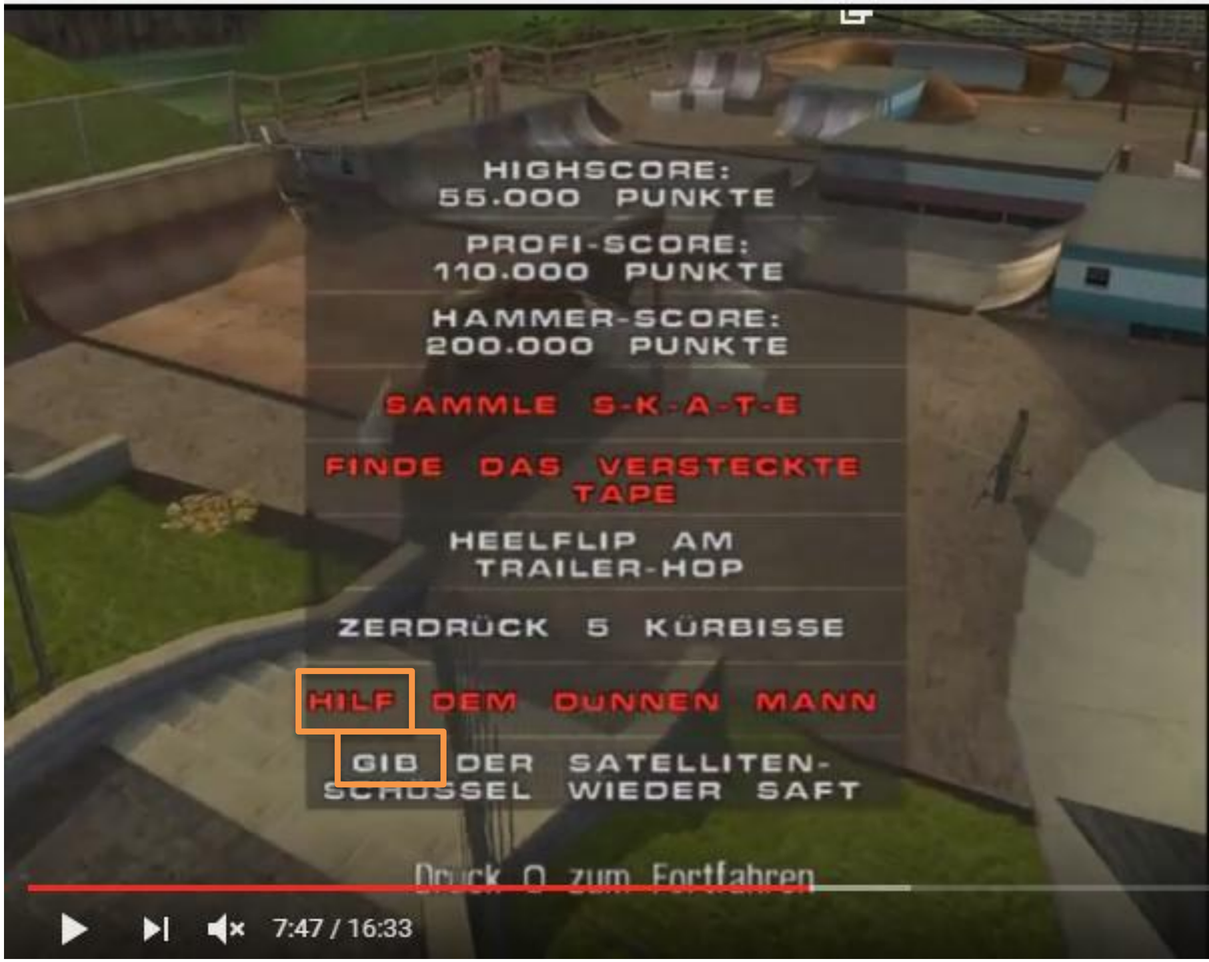
Wenn ihr vom Start weg wieder zum Mann mit den Griller fahrt, müsst ihr das Haus hinauffahren. Dort sind zwei Kabel vom Haus (wo du stehst) zu einen anderen Haus. Die müsst ihr Abgrinden bis die Zweige hinunterfallen und ihr seit wieder einen Schritt weiter.

http://www.spieletipps.de/tipps-7722-tony-hawks-pro-skater-3-tipps-fuer-vorstadt/#1

2d. Authorship confirmation

My solution:

- consulting other walkthroughs, lists of tasks and achievements and Let’s Play videos (youtube.com)



HIGHSCORE:
55.000 PUNKTE

PROFI-SCORE:
110.000 PUNKTE

HAMMER-SCORE:
200.000 PUNKTE

SAMMLE S-K-A-T-E

FINDE DAS VERSTECKTE TAPE

HEELFLIP AM TRAILER-HOP

ZERDRÜCK 5 KÜRBISSE

HILF DEM DUNNEN MANN

GIB DER SATELLITEN-SCHÜSSEL WIEDER SAFT

Drück O zum Fortfahren

https://www.youtube.com/watch?v=ZnfonJ24yf4

2d. Authorship confirmation

My solution:

- consulting other walkthroughs, lists of tasks and achievements and Let's Play videos (youtube.com)

Alternative solution:

- creation of reference corpus

2e. Annotation for frequency variables

Challenge:

- basic assumption: verb token frequency determines stem vowel choice (Conserving Effect)
- frequency distributions in walkthroughs are skewed
- example: essen ‘to eat’ is a strong verb with vowel gradation, but avatars in games do not usually eat a lot

➢ values for frequency variables could not be taken from walkthrough corpus

2e. Annotation for frequency variables

My solution:

- consultation of
  - frequency dictionaries (Jones and Tschirner 2006; Ruoff 1990),
  - reference corpora (DeReWo; Projekt Wortschatz Universität Leipzig) and
  - dictionary of German (Duden online)

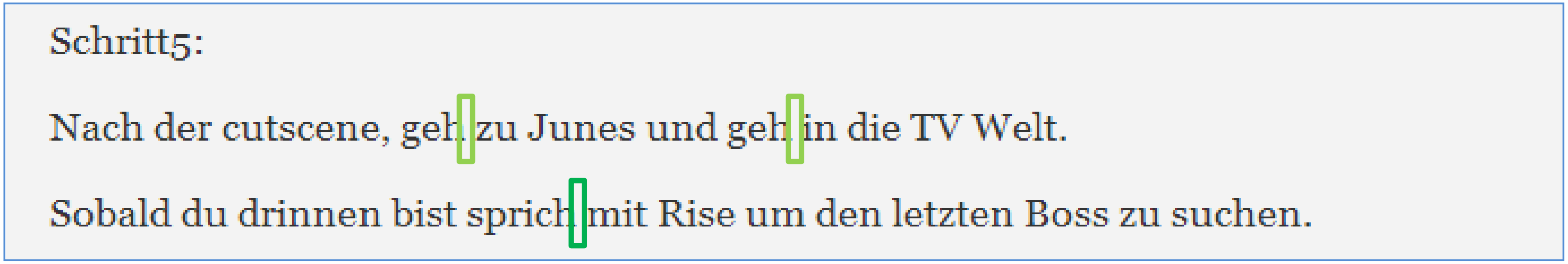
Alternative solution:

- depending on corpus and research question
  - drawing frequencies from research corpus
  - or from reference corpora/ other sources (more than one)

2f. Annotation for further variables

Challenge:

- close reading of some corpus texts revealed potential persistence effect (Szmrecsanyi 2005, 2006)

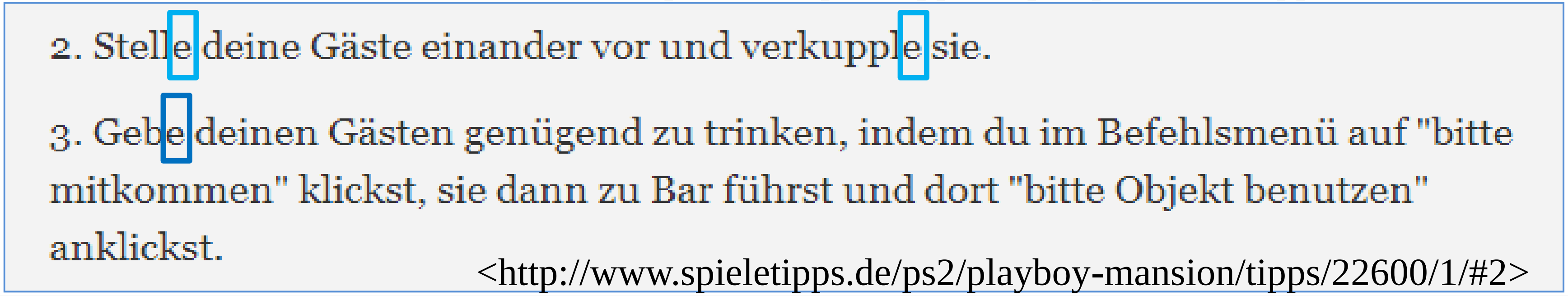


Schritt5:

Nach der cutscene, geh zu Junes und geh in die TV Welt.

Sobald du drinnen bist sprich mit Rise um den letzten Boss zu suchen.

<http://www.spieletipps.de/ps2/persona-4/tipps/33796/2/#2>



2. Stelle deine Gäste einander vor und verkupple sie.

3. Gebe deinen Gästen genügend zu trinken, indem du im Befehlsmenü auf "bitte mitkommen" klickst, sie dann zu Bar führst und dort "bitte Objekt benutzen" anklickst.

<http://www.spieletipps.de/ps2/playboy-mansion/tipps/22600/1/#2>

2f. Annotation for further variables

My solution:

- manual annotation of observations for
  - presence
  - verb class and
  - suffixation

  of imperative singular forms
- 20 words of left context

Alternative solution:

- close inspection of a sample of target texts
- adjust crawler or search interface

2g. Meta information

Challenge:

- authors of walkthroughs provide little personal information in their profile pages
- ➢ annotation for sociolinguistic variables is scarce:
- ➢ of all 1939 instances in the final dataset
    - ➢ 21.3 % have an annotation for speaker age
    - ➢ 13.4 % for speaker gender and
    - ➢ 6.8 % for speaker's residence

2g. Meta information

My solution:

- identify trends in corpus study
- collect further data in experimental study

Alternative solution:

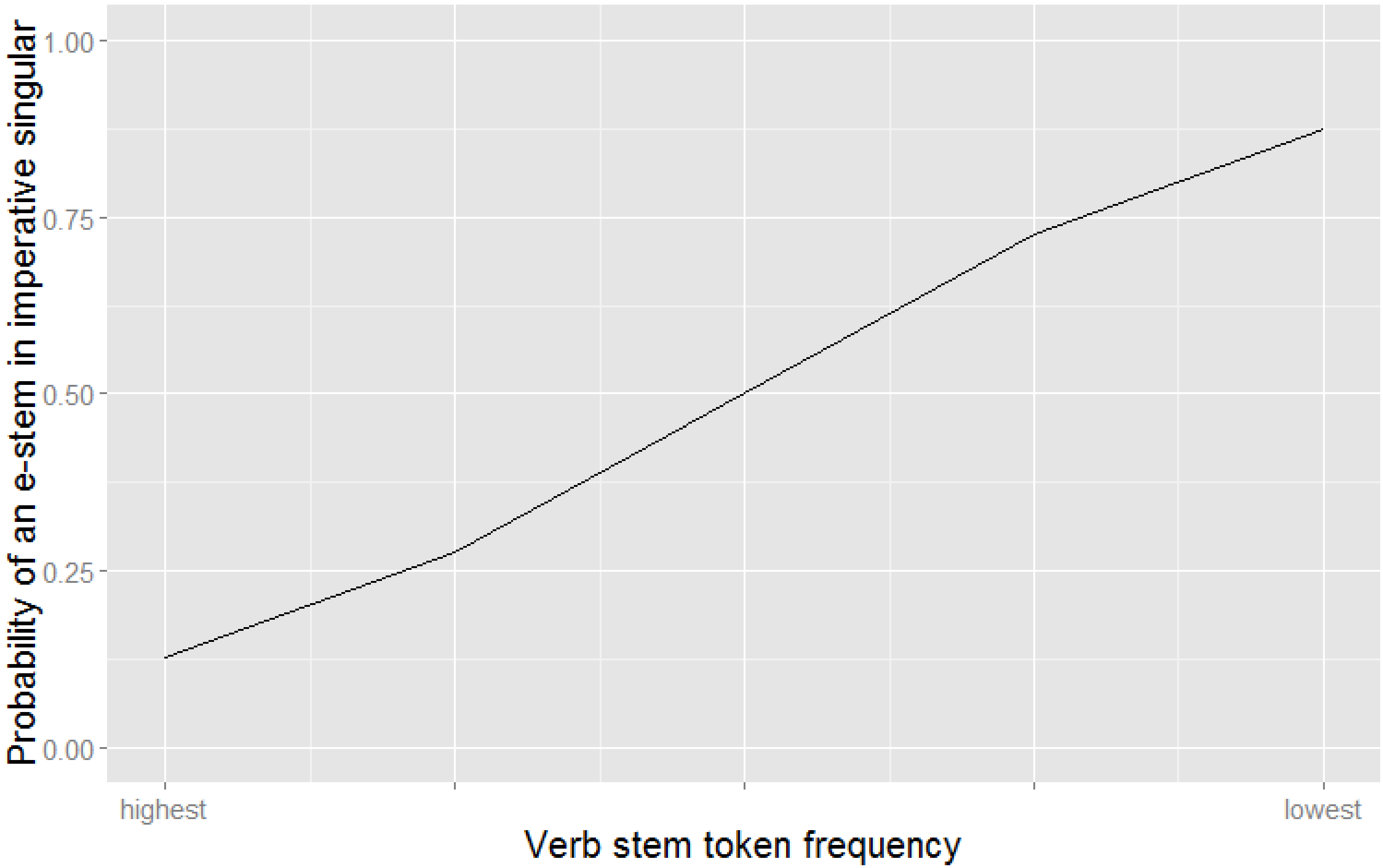
- collect more data
  - from additional corpus
  - from interviews
  - from experiment
- for specific research questions

3. Joys

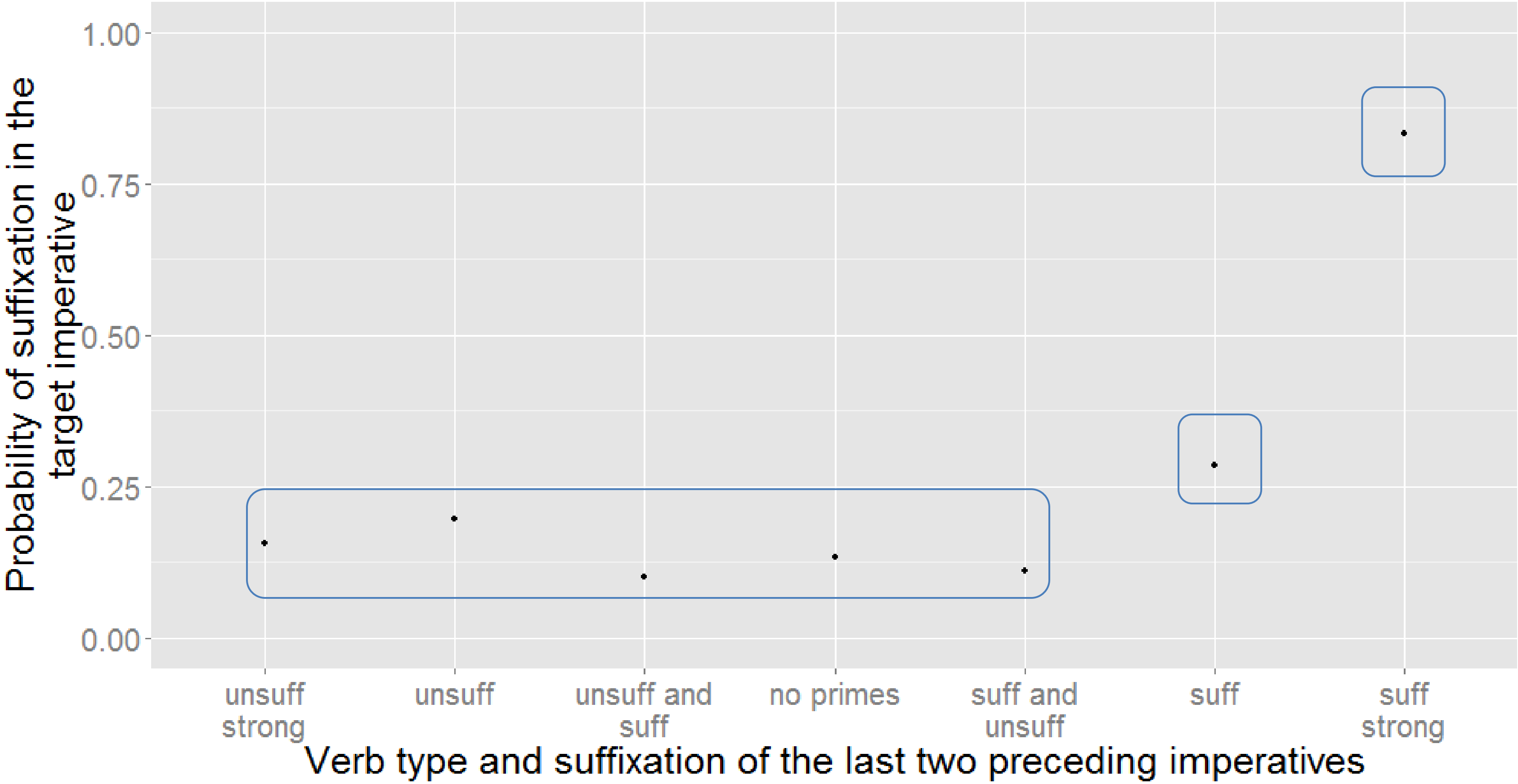
a. Results of Corpus Study

b. Design and Results of Experimental Study

3a. Results of Corpus Study - Verb token frequency effect



3a. Results of Corpus Study - Persistence Effect



3b. Design of Experimental Study

- Stimuli
  - 30 stimulus sentences and 30 fillers in alternation

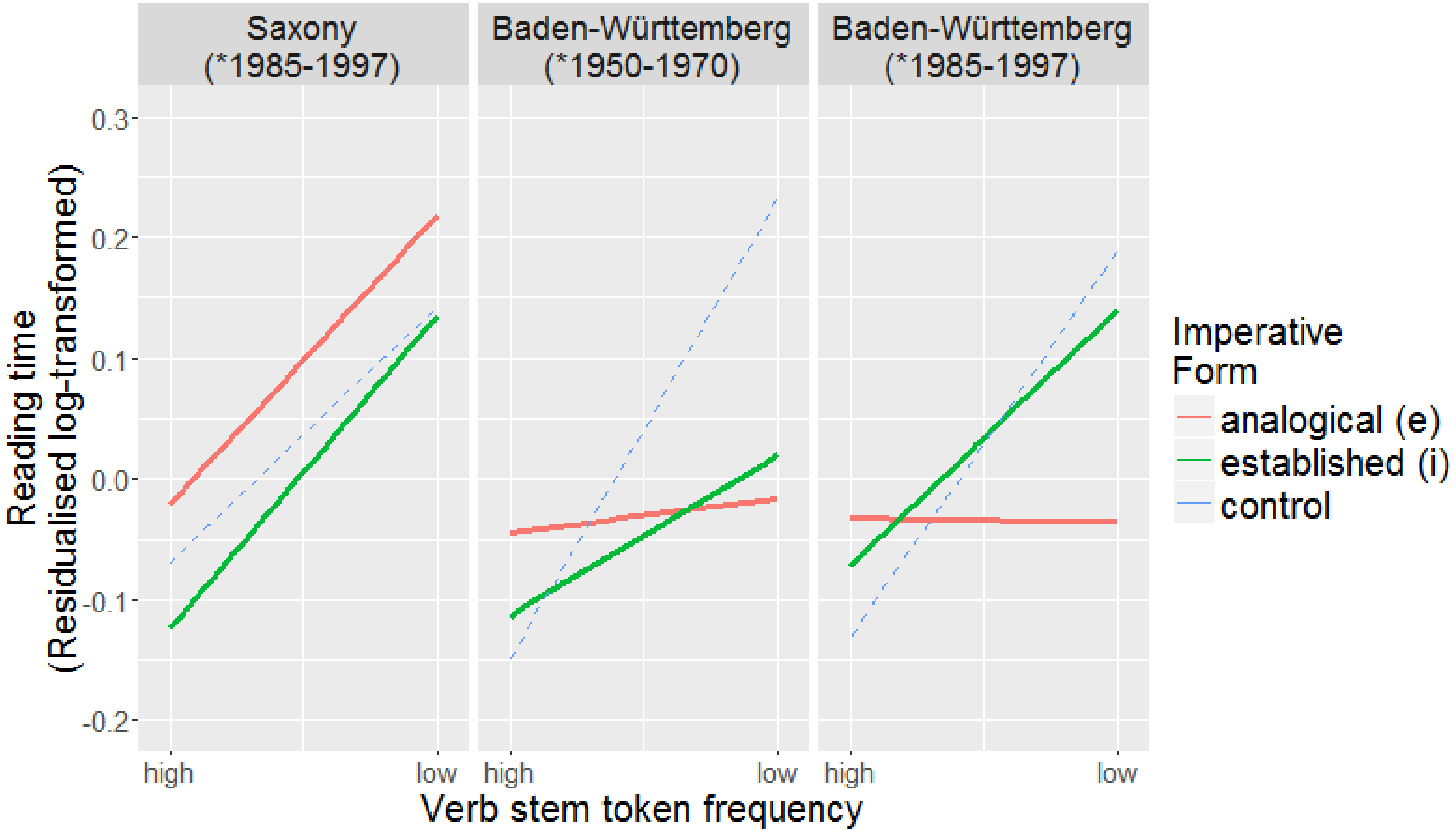
- Participants
  - from Baden-Württemberg:

    31 aged 18-30

    27 aged 45-65
  - from Saxony:

    28 aged 18-30

3b. Results of Experimental Study



4. Conclusion

- challenges lure in all stages of corpus compilation, annotation and search
- time-consuming manual correction/ annotation may be avoided by
  - careful inspection of websites and texts
  - training tagger
  - compilation of reference corpus
- further studies can yield additional and/or converging evidence
- results are worth the effort

Thank you for your attention!

anne.krause@uni-leipzig.de

Acknowledgements:

DFG German Research Foundation

Research Training Group GRK DFG 1624 “Frequency Effects in Language”, University of Freiburg

Prof. Dr. Peter Auer, University of Freiburg (supervisor)

Prof. Dr. Christian Mair, University of Freiburg (supervisor)

Prof. Dr. Doris Schönefeld, Leipzig University

References:

Bybee, Joan und Thompson, Sandra. 1997. “Three Frequency Effects in Syntax”, *Proceedings of the Twenty-Third Annual Meeting of the Berkeley Linguistics Society*, 378-388.

Duden online. 2013. <http://www.duden.de/woerterbuch>. - Duden-Corpus

Jones, Randall L., und Erwin Tschirner. 2006. *A Frequency Dictionary of German: Core Vocabulary for Learners*. London: Routledge. - Herder/BYU-Corpus

Korpusbasierte Wortformenliste (bzw. Wortgrundformenliste) DEREWO, <v-xxx>, mit Benutzerdokumentation, <http://www.ids-mannheim.de/derewo>, © Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland, 2013.

Ruoff, Arno. 1990. *Häufigkeitswörterbuch gesprochener Sprache*. Tübingen: Niemeyer.

Szmrecsanyi, Benedikt. 2005. “Language users as creatures of habit: A corpus-based analysis of persistence in spoken English” *Corpus Linguistics and Linguistic Theory* 11: 113-150.

Szmrecsanyi, Benedikt. 2006. *Morphosyntactic Persistence in Spoken English: A Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis*. Berlin: Mouton De Gruyter.

Wortschatz-Portal Universität Leipzig. 1998-2014. <http://wortschatz.uni-leipzig.de>.

www.spieletipps.de - walkthroughs

www.youtube.com - Let’s Play Videos