Using Twitter for linguistic purposes: three case studies

Ines Rehbein

DGfS 2014

January 17, 2015
Outline

1. Introduction

2. Case study I: Regional variation

3. Case study II: Variation in discourse

4. Case study III: Word order variation

5. Conclusions
Web Data as a Challenge for Theoretical Linguistics and Corpus Design

- **Advantages**
  - large amount of data
  - easily accessible
  - already digitised

- **Challenges**
  - legal issues
  - privacy issues
  - no (reliable) meta data
  - representativity
  - replicability
  - no reliable linguistic annotations

(also see Kilgariff, 2003; Kilgariff and Grefenstette, 2003; Lüdeling, Evert and Baroni, 2007)
Web Data as a Challenge for Theoretical Linguistics and Corpus Design

Topic of this talk

Can we use web data in linguistic research, and if so, what for?

- Three case studies on using Twitter microtext for studying
  1. regional variation
  2. variation in discourse
  3. word order variation
Outline

1. Introduction
2. Case study I: Regional variation
3. Case study II: Variation in discourse
4. Case study III: Word order variation
5. Conclusions
Case study I: Regional Variation

Can we use Twitter microtext for studying regional variation?

- Limitations
  - no reliable metadata (age, social group, demographic info)
  - only 1% of all German tweets come with geoposition information (Scheffler 2013)
  - representativity

- Which research questions can we investigate with this type of data?
Case study I: Regional Variation

Related work

- O’Connor et al. (2010): A Mixture Model of Demographic Lexical Variation – how demographic social factors influence lexical choice

- Eisenstein (2012): Phonological variation in Twitter microtext
  - Consonant cluster reduction in tweets shows the same sensitivity to context as in spoken language: reduced form more frequent when the subsequent segment starts with a vowel

- Herdağdelen (2012): Twitter n-gram corpus with demographic metadata used to compute gender bias of verb phrases (Herdağdelen & Baroni 2011)

- Doyle (2014): Mapping Language Variation and Dialects through Twitter
Case study I: Regional Variation

This work

- **Goal:** building a corpus reflecting regional variation in German

- **Approach:**
  - collect tweets from 48 different locations in Germany
    - Twitter Search API (geoposition parameters)
    - time period of 6 months
  - Geoposition parameter only approximates the user environment (no guarantee that the user has been socialised in that region)

- **Corpus with 7,311,960 tweets (105,074,399 tokens)**

**Hypothesis**

By collecting a large amount of data we will be able to observe regional differences between the different regional subcorpora
Case study I: Regional Variation

Do the different subcorpora reflect regional variation?

How to test?

- Principal Components Analysis (PCA) based on the frequency of dialectal markers in the different subcorpora:
  - reduce complexity of high-dimensional data
  - detect underlying patterns, based on the variance in the data
  - dialectal markers: dat, isch, misch, ned, schee, wa
  - using the 16 largest subcorpora

- Hierarchical clustering of the first four components of the PCA
Case study I: Regional Variation

Conclusions

- Proof of concept: different subcorpora display different characteristics which are correlated with their regional distribution – what now?

- What we can do:
  - test whether a particular lexical item / phrase / phonetic spelling is more prominent in a particular subcorpus
  - use the data for quantitative studies, averaging over individual users

- What we shouldn’t do:
  - make claims based on
    - individual users
    - individual data points
Outline

1 Introduction

2 Case study I: Regional variation

3 Case study II: Variation in discourse

4 Case study III: Word order variation

5 Conclusions
Case study II: Variation in discourse

- Twitter as a communication form (not a register) (Dürscheid 2003)
  - Different registers: news, ads, public announcements, private communication, ...

- Can we use Twitter for investigating variation in discourse as a convenient stand-in for spoken language?
  - Pro: large amounts of data, no transcription needed
  - Con: tweets are very short, not enough context for analysis

Goal:
Building a register-specific Twitter corpus with private communication in context
Case study II: Variation in discourse

- Twitter: several possibilities for identifying and sorting tweets
  - Hashtag: augment tweets with one or more semantic tags or keywords
  - Twitter metadata: who posted the tweet to whom?

<table>
<thead>
<tr>
<th>total</th>
<th>tweet authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;7 mio tweets</td>
<td>191,589</td>
</tr>
<tr>
<td>tweets directed to other user(s)</td>
<td>59,787</td>
</tr>
<tr>
<td>tweets with no recipient</td>
<td>178,630</td>
</tr>
</tbody>
</table>

- Tweets without a recipient
  - news
  - ads
  - job announcements
  - automatically created tweets
  - ...

Rehbein (DGfS 2014)  Twitter for linguistics  January 17, 2015  14 / 31
Case study II: Variation in discourse

JSON metadata

```json
{
    "created_at": "Tue Mar 26 15:02:00 +0000 2013",
    "id": 316565706151956480, "id_str": "316565706151956480", "text": "@kirschkopf Ja, weil!", "source": "<a href="http://www.metrotwit.com" rel="nofollow">...
```

Rehbein (DGfS 2014)
Case study II: Variation in discourse

Reconstruct the discourse between two or more users

Procedure:

- Identify private communication with help of the `reply_to_status_id` field
- From those, select frequently tweeting authors
- Extract all tweets from the corpus sent by one of these authors
- Extract tweet ids, use as seed for creating the discourse-in-context corpus

Iterative process:

1. extract first tweet id from the list
2. download the tweet
3. extract `reply_to_status_id`, add to list
4. continue until you reach the first tweet of the thread
Case study II: Variation in discourse

<table>
<thead>
<tr>
<th>Seed:</th>
<th>25 twitter users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus:</td>
<td>11,130 tweets posted by 799 users</td>
</tr>
</tbody>
</table>

Example

@A schlechte Laune?
   “Bad mood?”

@B Ein bisschen, ja. :l
   “A bit, yes.”

@A mhm. Magst du mir sagen warum? :)
   “Mhm. You wanna tell me, why?”

@B Weil wegen Eltern und nervig und so halt.
   “Because of my parents and everything is annoying and so on.”
Case study II: Variation in discourse
Comparing private tweets to spoken language

- Comparison of type-token ratio (TTR) in the discourse-in-context corpus and in spoken language
  - TüBa-D/S: students role-playing business partners, domain: scheduling
  - KiDKo: spontaneous, highly informal dialogues between adolescents, domain: unrestricted

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Tokens</th>
<th>Types</th>
<th>Hapax Leg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TüBa-D/S</td>
<td>138.121</td>
<td>4.199 (3.975)</td>
<td>1.824 (1.717)</td>
</tr>
<tr>
<td>KiDKo</td>
<td>138.121</td>
<td>6.339 (5.429)</td>
<td>2.222 (1.827)</td>
</tr>
<tr>
<td>Twitter</td>
<td>138.121</td>
<td>18.791 (14.489)</td>
<td>12.521 (8.985)</td>
</tr>
</tbody>
</table>

Table: excluding punctuation, user names, hashtags, emoticons, numbers

TüBa-D/S (Stegmann et al. 2000), KiDKo (Rehbein et al. 2013)
Case study II: Variation in discourse

Possible reasons for larger vocabulary size in Twitter

- Domain effects: open domain vs. restricted domain (schedule making)
- More different users (speakers) → more variety

Noise
- Tokenisation:
  - sieseien (they were), sindkaum (are rarely), siehtdas (sees this), ...
- Non-canonical spelling:
  - isch/icke → ich (I), ichs → ich es (I it), ned → nicht (not), ...
- Foreign material:
  - nice, sorry, mom, clean, closed, come, cup, other, word, workspaces, ...

Fair comparison not feasible without manual correction
⇒ comparative corpus-based studies (sentence length/TTR/...) unreliable!
Case study II: Variation in discourse
Comparing tweets with spoken language

- Can we use Twitter as a convenient stand-in for spoken language?
- Markers of orality in different corpora
  - INT: interjections
  - BS: backchannel signals
  - QU: question tags
  - FIL: filler

<table>
<thead>
<tr>
<th>Corpus</th>
<th>ITJ</th>
<th>QU</th>
<th>BS</th>
<th>FIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>1.157</td>
<td>162</td>
<td>91</td>
<td>24</td>
</tr>
<tr>
<td>TüBa-D/S</td>
<td>909</td>
<td>310</td>
<td>16</td>
<td>na</td>
</tr>
<tr>
<td>KiDKo</td>
<td>3.934</td>
<td>235</td>
<td>534</td>
<td>802</td>
</tr>
</tbody>
</table>
Can we use Twitter as a convenient stand-in for spoken language?

Markers of orality in different corpora

- **INT**: interjections
- **QU**: question tags
- **BS**: backchannel signals
- **FIL**: filler

<table>
<thead>
<tr>
<th>Corpus</th>
<th>ITJ</th>
<th>QU</th>
<th>BS</th>
<th>FIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>1.157</td>
<td>162</td>
<td>91</td>
<td>24</td>
</tr>
<tr>
<td>TüBa-D/S</td>
<td>909</td>
<td>310</td>
<td>16</td>
<td>na</td>
</tr>
<tr>
<td>KiDKo</td>
<td>3.934</td>
<td>235</td>
<td>534</td>
<td>802</td>
</tr>
</tbody>
</table>
Case study II: Variation in discourse
Comparing tweets with spoken language

Can we use Twitter as a convenient stand-in for spoken language?

Markers of orality in different corpora
  - INT: interjections
  - QU: question tags
  - BS: backchannel signals
  - FIL: filler

<table>
<thead>
<tr>
<th>Corpus</th>
<th>ITJ</th>
<th>QU</th>
<th>BS</th>
<th>FIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>1.157</td>
<td>162</td>
<td>91</td>
<td>24</td>
</tr>
<tr>
<td>TüBa-D/S</td>
<td>909</td>
<td>310</td>
<td>16</td>
<td>na</td>
</tr>
<tr>
<td>KiDKo</td>
<td>3.934</td>
<td>235</td>
<td>534</td>
<td>802</td>
</tr>
</tbody>
</table>
Case study II: Variation in discourse
Comparing tweets with spoken language

- Can we use Twitter as a convenient stand-in for spoken language?
- Markers of orality in different corpora
  - INT: interjections
  - QU: question tags
  - BS: backchannel signals
  - FIL: filler

<table>
<thead>
<tr>
<th>Corpus</th>
<th>ITJ</th>
<th>QU</th>
<th>BS</th>
<th>FIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>1.157</td>
<td>162</td>
<td>91</td>
<td>24</td>
</tr>
<tr>
<td>TüBa-D/S</td>
<td>909</td>
<td>310</td>
<td>16</td>
<td>na</td>
</tr>
<tr>
<td>KiDKo</td>
<td>3.934</td>
<td>235</td>
<td>534</td>
<td>802</td>
</tr>
</tbody>
</table>
Can we use Twitter as a convenient stand-in for spoken language?

Markers of orality in different corpora

- **INT**: interjections
- **BS**: backchannel signals
- **QU**: question tags
- **FIL**: filler

<table>
<thead>
<tr>
<th>Corpus</th>
<th>ITJ</th>
<th>QU</th>
<th>BS</th>
<th>FIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>1.157</td>
<td>162</td>
<td>91</td>
<td>24</td>
</tr>
<tr>
<td>TüBa-D/S</td>
<td>909</td>
<td>310</td>
<td>16</td>
<td>na</td>
</tr>
<tr>
<td>KiDKo</td>
<td>3.934</td>
<td>235</td>
<td>534</td>
<td>802</td>
</tr>
</tbody>
</table>
Case study II: Variation in discourse

Conclusions

- Possible to build a corpus with private communication in context
  - Highly interactive data, many features of orality
  - Also differences to real spoken language
  - But also crucial differences between spoken language corpora

Most serious limitation:
Not possible to do comparative corpus studies based on quantitative measures such as TTR, sentence length, ...
Outline

1. Introduction
2. Case study I: Regional variation
3. Case study II: Variation in discourse
4. Case study III: Word order variation
5. Conclusions
Case study III: Word order variation

- Tweets as supplementary data for investigating non-frequent non-canonical phenomena in spoken language
- Test case: weil-V2 (because with verb second word order)
- Ongoing discussion, see, e.g. Antomo & Steinbach (2010), Reis (2013)
Case study III: Word order variation
Creating a corpus with *weil* clauses

- Collect instances of *weil* via Twitter
  - Twython (Tatjana Scheffler), access to Twitter API via Tweepy
  - Download period: 11 days

- *weil* corpus:
  - 1,045,164 tokens
  - 51,768 tweets
  - avg. tweet length: 20.2 tokens
  - (avg. tweet length in the discourse-in-context corpus: 12.4 tokens)

Twython: [http://www.ling.uni-potsdam.de/~scheffler/twitter/](http://www.ling.uni-potsdam.de/~scheffler/twitter/)
Case study III: Word order variation

Data analysis

- Identifying instances of *weil-V2*
  - use POS tags, search for *weil X V\textsubscript{finite}*
  - POS tags are noisy $\rightarrow$ need to validate results!
  - (we might miss some instances due to erroneous POS)
Case study III: Word order variation

Data analysis II

- Compare distribution with TüBa-D/S and KiDKo

<table>
<thead>
<tr>
<th></th>
<th>TüBa-D/S</th>
<th>KiDKo</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>weil</td>
<td>403</td>
<td>342</td>
<td>51,839</td>
</tr>
<tr>
<td>weil-V2</td>
<td>143</td>
<td>149</td>
<td>(\approx 1,094)</td>
</tr>
<tr>
<td>weil (PPER) (V_{\text{fin}})</td>
<td>37.1</td>
<td>42.9</td>
<td>38.3</td>
</tr>
<tr>
<td>weil (ADV) (V_{\text{fin}})</td>
<td>22.4</td>
<td>11.4</td>
<td>21.4</td>
</tr>
<tr>
<td>weil (PIS/PDS) (V_{\text{fin}})</td>
<td>11.9</td>
<td>12.1</td>
<td>13.6</td>
</tr>
<tr>
<td>weil (NP) (V_{\text{fin}})</td>
<td>11.2</td>
<td>11.4</td>
<td>19.5</td>
</tr>
<tr>
<td>weil (PP) (V_{\text{fin}})</td>
<td>11.2</td>
<td>5.4</td>
<td>6.9</td>
</tr>
<tr>
<td>weil (VVINF) (V_{\text{fin}})</td>
<td>0.7</td>
<td>0.0</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Case study III: Word order variation

Data analysis II

- Compare distribution with TüBa-D/S and KiDKo

<table>
<thead>
<tr>
<th></th>
<th>TüBa-D/S</th>
<th>KiDKo</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>weil</td>
<td>403 %</td>
<td>342 %</td>
<td>51,839 %</td>
</tr>
<tr>
<td>weil-V2</td>
<td>143 35.5</td>
<td>149 43.6</td>
<td>≈ 1,094 2.1</td>
</tr>
<tr>
<td>weil PPER V&lt;sub&gt;fin&lt;/sub&gt;</td>
<td>37.1</td>
<td>42.9</td>
<td>38.3</td>
</tr>
<tr>
<td>weil ADV V&lt;sub&gt;fin&lt;/sub&gt;</td>
<td>22.4</td>
<td>11.4</td>
<td>21.4</td>
</tr>
<tr>
<td>weil PIS/PDS V&lt;sub&gt;fin&lt;/sub&gt;</td>
<td>11.9</td>
<td>12.1</td>
<td>13.6</td>
</tr>
<tr>
<td>weil NP V&lt;sub&gt;fin&lt;/sub&gt;</td>
<td>11.2</td>
<td>11.4</td>
<td>19.5</td>
</tr>
<tr>
<td>weil PP V&lt;sub&gt;fin&lt;/sub&gt;</td>
<td>11.2</td>
<td>5.4</td>
<td>6.9</td>
</tr>
<tr>
<td>weil VVINF V&lt;sub&gt;fin&lt;/sub&gt;</td>
<td>0.7</td>
<td>0.0</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Case study III: Word order variation

Data analysis II

- Compare distribution with TüBa-D/S and KiDKo

<table>
<thead>
<tr>
<th></th>
<th>TüBa-D/S</th>
<th>KiDKo</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>weil</td>
<td>403 %</td>
<td>342 %</td>
<td>51,839 %</td>
</tr>
<tr>
<td>weil-V2</td>
<td>143 %</td>
<td>149 %</td>
<td>≈ 1,094 %</td>
</tr>
<tr>
<td>weil PPER V_{fin}</td>
<td>37.1</td>
<td>42.9</td>
<td>38.3</td>
</tr>
<tr>
<td>weil ADV V_{fin}</td>
<td>22.4</td>
<td>11.4</td>
<td>21.4</td>
</tr>
<tr>
<td>weil PIS/PDS V_{fin}</td>
<td>11.9</td>
<td>12.1</td>
<td>13.6</td>
</tr>
<tr>
<td>weil NP V_{fin}</td>
<td>11.2</td>
<td>11.4</td>
<td>19.5</td>
</tr>
<tr>
<td>weil PP V_{fin}</td>
<td>11.2</td>
<td>5.4</td>
<td>6.9</td>
</tr>
<tr>
<td>weil VVINF V_{fin}</td>
<td>0.7</td>
<td>0.0</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Case study III: Word order variation

Data analysis II

- Compare distribution with TüBa-D/S and KiDKo

<table>
<thead>
<tr>
<th></th>
<th>TüBa-D/S</th>
<th>KiDKo</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>weil</td>
<td>403</td>
<td>342</td>
<td>51,839</td>
</tr>
<tr>
<td>weil-V2</td>
<td>143</td>
<td>149</td>
<td>≈ 1,094</td>
</tr>
<tr>
<td>weil PPER V_{fin}</td>
<td>37.1</td>
<td>42.9</td>
<td>38.3</td>
</tr>
<tr>
<td>weil ADV V_{fin}</td>
<td>22.4</td>
<td>11.4</td>
<td>21.4</td>
</tr>
<tr>
<td>weil PIS/PDS V_{fin}</td>
<td>11.9</td>
<td>12.1</td>
<td>13.6</td>
</tr>
<tr>
<td>weil NP V_{fin}</td>
<td>11.2</td>
<td>11.4</td>
<td>19.5</td>
</tr>
<tr>
<td>weil PP V_{fin}</td>
<td>11.2</td>
<td>5.4</td>
<td>6.9</td>
</tr>
<tr>
<td>weil VVINF V_{fin}</td>
<td>0.7</td>
<td>0.0</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Case study III: Word order variation

Conclusions

- Easy to extract additional instances for *weil-V2* with similar properties as the spoken instances
- Larger samples might help us to detect patterns in the data
- Questions we can’t answer:
  - How many instances of *weil-V2* are in the Twitter corpus?
  - Ratio of *weil-V2* in equally sized samples from different corpora?
Outline

1. Introduction
2. Case study I: Regional variation
3. Case study II: Variation in discourse
4. Case study III: Word order variation
5. Conclusions
Conclusions

- Twitter data can be used to
  - extract register-specific corpora
  - extract instances for investigating rare linguistic phenomena
- Data can be used for
  - quantitative studies where we average over many tweets/users
- Data should not be used for
  - making claims based on individual users/tweets

Results from comparative corpus-based studies (based on sentence length/TTR/syntactic complexity, ...) are not reliable

- What we need in the future:
  - more reliable preprocessing/linguistic annotations (at least for some types of studies)
Thank You!

Questions?
References