Big (Language) Data – From research strategies to proof-of-concept and implementation projects in linguistics research

Gerhard Budin
Centre for Translation Studies
University of Vienna
LEARN Workshop Vienna
6th of April, 2016
Introduction & Overview

• What is Big (Language) Data
• The “ZTW Cloud” – a digital humanities research infrastructure
• Research data management strategies
• Case study: FWF-SFB “DiÖ German in Austria”
• Conclusions
Big Data + linguistics = Big Language Data

- Big Data has become a crucial concept for linguistics research due to its statistical paradigm of the analysis of language usage in computational corpus linguistics and computational translation studies.
- "More data are good data" – for training tools, testing algorithms for language data analysis, for linguistic corpus analysis
- Thus Big Language Data is becoming a cross-disciplinary (computer science – linguistics – digital humanities) methodology that is increasingly applied in relevant research projects
ZTW Cloud: a digital humanities research infrastructure

• ZTW = Zentrum für Translationswissenschaft = Centre for Translation Studies at the University of Vienna

• The ZTW Cloud (located in the HPSC Arsenal) is a research infrastructure for language and multimedia data consisting of
  – A HADOOP server cluster with about 100 TB storage and processing capacity mainly used for:
    • Machine translation research on MOSES
    • Text Data mining and corpus research (Cloudera, ProTerm, etc.)
    • CAT & Terminology research (SDL Suite, Multiterm, etc.)
    • Multimedia conference interpreting corpus research
  – A conference interpreting facility (20 booths in 2 rooms) and other digital interpreting tools
  – 3 media labs with 80 student multimedia workplaces
Research data management Strategies

• The ZTW Cloud is embedded in (and in fact the core of) the CLARIN Core Facility at the University of Vienna

• ...which in turn is embedded in the Austrian CLARIAH (CLARIN+DARIAH) digital humanities research infrastructure network (Vienna, Graz, Innsbruck, et al.)

• Thus the CLARIN and DARIAH data management best practices focusing on the needs of linguistics and other humanities disciplines (metadata standards, data curation principles, corpus annotation standards, etc.) are strictly followed (for CLARIN based on TEI and ISO)

• In addition, the generic research data management principles of LERU, LEARN, and other initiatives at EU level and at national levels are also followed
Case study: FWF-SFB “DiÖ German in Austria”

• -> A showcase for the genesis of a large research project (in this case in linguistics research) – from research questions to a project that is meant to be a proof-of-concept and an implementation of a systematic research data management strategy

• SFB = Spezialforschungsbereich funded by FWF (Fond zur Förderung der Wissenschaftlichen Forschung)

• F 60 DiÖ: A consortium consisting of UniVie (coord. with 3 departments from 2 faculties), Uni Graz, Uni Salzburg, Austrian Academy of Sciences, and other members)

• The project started on January 1st 2016 and lasts in its first phase for 4 years with an option for prolongation for another 4 years
Long Term Goals

Comprehensive, interdisciplinary and multidimensional linguistic analyses on DiÖ and its varieties

Set up of a digital research infrastructure on DiÖ (sustainability)
Research Program: Task Clusters

A: Coordination

B: Variation and Change of German in Austria – Perspectives of Variationist Linguistics

C: German and other Languages in Austria – Perspectives of Language Contact

D: ‘German in the Minds’ – Language Attitudes and Perception

E: Collaborative Online Research Platform – ‘German in Austria’
Examples of research methods

Data Processing ...
(segmented) recorded data,
transcriptions and annotations (EXMERaLDA)

→

orthographic transcriptions plus:
– conversation analytic transcription
– phonetic transcription

vowels/diphthongs, liquids)

Data Documentation:
– research platform (PP11)
– first ‘two-dimensional’
talking dialect atlas for
whole of Austria
(cf. SAO 1998ff.)

... and Data Analysis:

→ – morphological and syntactic analysis
→ – acoustic-phonetic analysis with STx
(selected

→ – analysis of variation and change in
apparent and real time
→ – dialectometric analysis (e.g. with
GeoLing) in stage II
(cf. Rumpf/Pickl/Elspaß/König/Schmidt 2009
...)

9
Key Methods: Locations and Informants

160 informants at 16 villages

<table>
<thead>
<tr>
<th>Groups</th>
<th>Autochthony</th>
<th>Number (gender)</th>
<th>Age (years)</th>
<th>Education type</th>
<th>Occupation type</th>
<th>Regional mobility</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Old’</td>
<td>+ high</td>
<td>1f / 1m</td>
<td>&gt; 65</td>
<td>- high</td>
<td>+ manual</td>
<td>- high</td>
</tr>
<tr>
<td>‘Young-I’</td>
<td></td>
<td>2f / 2m</td>
<td>20–30</td>
<td>+ high</td>
<td>- manual</td>
<td>+ high</td>
</tr>
<tr>
<td>‘Young-II’</td>
<td></td>
<td>2f / 2m</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Key Methods: Data Survey

Providing data of diverse ‘sections’ (varieties/styles) of an informant’s individual variation repertoire

- reading tasks
- translation tasks
- ‘experiments’ / tests
- interviews (formal)

- translation tasks
- ‘experiments’ / tests
- informal conversations
Key Methods: Data Handling and Analyses

Stage I
quantitative & qualitative analyses (phonology, morphology, syntax)

- various transcriptions
  orthographic transcriptions, phonetic transcriptions, grammatical annotations & (partly) conversation-analytic transcriptions
  (➔ ‘stylistic practice’)

- various analyses
  type-token analyses, phonetic distance measurements, analysis of co-occurrence restrictions, ‘local-sequence analyses’ a. o.

- comprehensive statistical analyses
Key Methods: Approach

Integrative Approach

Correlative Global Analyses

Conversational Local Analyses

analyses of discourse data
analyses of preferred variants/groups of variants
analyses of connections mobility and languages

Linguistic Target Levels

Stage I
› phonetic/phonological
› morphological/syntactic
› pragmatic

Stage II
› suprasegmental (prosodic)
› lexical
› pragmatic
Key Methods: Locations

Vienna
population: 1,766,746 (Statistik Austria)
migration background: 40-50% (wien.at)

Graz
population: 309,323 (Statistik Austria)
migration background: 18% (graz.at)
Key Methods: Informants

Stage I
40 informants per city
40 informants per regional surroundings of a city

---

total sample size: 160 Informants

Stage II
additional sample of informants with migration background

Selection of Informants
central criteria: local autochthony (permanent primary residence)
sample balanced by age (two groups: 20-30 and >65), gender (1:1), educational background, type of occupation
### Key Methods: Informants

Additional urban specific criteria of selection

<table>
<thead>
<tr>
<th>2 districts per city</th>
<th>20 pers.</th>
<th>10 pers. (born and raised native speaker)</th>
<th>10 pers. (native speaker inner Austrian migration)</th>
</tr>
</thead>
<tbody>
<tr>
<td>migration background in population high</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>migration background in population low</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4 agglomeration locations per city</th>
<th>20 pers.</th>
<th>10 pers. (born and raised native speaker)</th>
<th>10 pers. (native speaker inner Austrian migration)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 with preserved demographical structures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 with changing demographical structures (high influx)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Key Methods: Data Survey

Instruments of Data Collection

- analytical interviews
- questionnaires
- tests
- free everyday conversations
- conversations among friends
- long-time recordings (stage II - 12h-recordings)

Formal situation – close to intended standard

Informal situation – far from intended standard (dialect)

1. analyses regarding socio-pragmatic and individual variation → inter-speaker variation
2. analyses regarding situative-pragmatic and interactional variation → intra-speaker variation
Key Methods: Data Survey / Selection

a) **edited data** from former research projects

b) **data on the situational embedding**: material revealing the factual situation of language use at that time

(e. g. official statistical returns, annual school programmes and school reports, publications on the occasion of anniversaries and other jubilees etc.)
Key Methods: Data Survey / Data Selection

a) **edited data** from former research projects

b) **data on language contact**: material from all available primary and secondary sources concerning Slavic-German contact phenomena, e.g. language corpora, dictionaries, relevant literature and collections, popular descriptions and recurring assumptions.
### Key Methods: Data Handling and Analyses

- evaluation and annotation of written sources
- application of philological and etymological methods
- application of methods from comparative linguistics
  (feature-by-feature comparison & internal reconstruction)

<table>
<thead>
<tr>
<th>Slavic Loans in DiÖ</th>
<th>Czech Loans in DiÖ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Buchtel (B-/W-)</strong></td>
<td><strong>heidipritsch</strong></td>
</tr>
<tr>
<td>‘yeast pastry’ &lt; Czech <em>buchta</em></td>
<td>‘totally gone’ &lt; Czech <em>hajdy + pryč</em></td>
</tr>
<tr>
<td>Liwanze</td>
<td><strong>Hubitschko</strong></td>
</tr>
<tr>
<td>‘pancakes’ &lt; Czech <em>lívanec</em></td>
<td>‘peck on the cheek’ &lt; Czech <em>hubička</em></td>
</tr>
<tr>
<td><strong>Klobasse (-e/-i)</strong></td>
<td><strong>Kaluppe</strong></td>
</tr>
<tr>
<td>‘hard smoked sausage’ &lt; Czech <em>klobása</em></td>
<td>‘dilapidated, ramshackle hut’ &lt; Czech <em>chalupa</em> also in the German diminutive form <em>Kalupperl</em></td>
</tr>
<tr>
<td><strong>Kolatsche (K-/G-)</strong></td>
<td><strong>Leschak</strong></td>
</tr>
<tr>
<td>‘small yeast cake with filling’ &lt; Czech <em>koláč</em></td>
<td>‘lay-about’ &lt; Czech <em>ležák</em></td>
</tr>
<tr>
<td><strong>Oblate</strong></td>
<td><strong>nemam</strong></td>
</tr>
<tr>
<td>‘fine wafer’ &lt; Czech <em>oplatka</em>; the DiÖ form is stressed on the first syllable as in Czech</td>
<td>‘have-not’ &lt; Czech <em>nemám</em></td>
</tr>
<tr>
<td><strong>Palatschinke</strong></td>
<td><strong>petschieren</strong></td>
</tr>
<tr>
<td>‘jam-filled pancake’ &lt; Czech <em>palačinka</em> (&lt; Hungarian palacsinta &lt; Romanian <em>plăcintă</em>)</td>
<td>‘seal’ &lt; Czech <em>zapečetit</em></td>
</tr>
<tr>
<td><strong>Powidl</strong></td>
<td><strong>powidalen</strong></td>
</tr>
<tr>
<td>‘plum jam’ &lt; Czech <em>povidla</em></td>
<td>‘tell’ derived from the preterite form of Czech <em>povidat</em></td>
</tr>
<tr>
<td><strong>Haluschka</strong></td>
<td></td>
</tr>
<tr>
<td>‘chopped cabbage fried in butter &amp; served over boiled noodles’ &lt; Slovak <em>haluška</em></td>
<td></td>
</tr>
</tbody>
</table>
**Key Methods: Data Handling and Analyses**

- evaluation and annotation of written sources
- application of philological and etymological methods
- application of methods from comparative linguistics
  (feature-by-feature comparison & internal reconstruction)

<table>
<thead>
<tr>
<th>Böhmak</th>
<th>‘Czech male’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feschak</td>
<td>‘dashing young man’</td>
</tr>
<tr>
<td>Tränak</td>
<td>‘camp follower’ (&lt; French <em>train</em> and -ák)</td>
</tr>
<tr>
<td>Tschunkerl</td>
<td>‘mucky pup’ &lt; čuně ‘piglet’ + Bavarian diminutive suffix –erl</td>
</tr>
<tr>
<td>Armutschkerl</td>
<td>‘poor wretch’ with two combined diminutive suffixes, i. e. Czech -č(e)k- + Bavarian –erl</td>
</tr>
<tr>
<td>verdobrischen</td>
<td>‘squander, blow’ &lt; dobrý ‘good’</td>
</tr>
<tr>
<td>Na servus Březina!</td>
<td>in order to express unpleasant surprise</td>
</tr>
<tr>
<td>Er ist immer der Nowak</td>
<td>in the sense of ‘he is always the victim, he is always abused’</td>
</tr>
<tr>
<td>auf Lepschi gehen</td>
<td>‘enjoy oneself’ equivalent to Czech <em>jít na lepší</em></td>
</tr>
<tr>
<td>pomáli, pomáli!</td>
<td>‘not so fast!’ &lt; Moravian Czech or Slovak <em>pomaly</em> ‘slow’</td>
</tr>
</tbody>
</table>
Key Methods: Data Survey and Selection

Multivariate Data Survey and Selection: quantitative & qualitative methods focusing on attitudes & perception

I: Online Questionnaires

II: In-depth Interviews (round I) (Task Cluster B)

III: ‘Listener Judgement Tests’ & Interviews (round II)
TASK CLUSTER E

COLLABORATIVE ONLINE RESEARCH PLATFORM

‘GERMAN IN AUSTRIA’
Goals

Construction and use of a collaborative online research platform (DiÖ Virtual Research Environment) for the entire project in order to facilitate or enable

- **efficiency**: collaborative research work
- **persistency**: consistent & persistent data preservation
- **accessibility**: enhancing accessibility & usability of research data
- **extent**: linking and correlation of all data
Goals

Construction and use of a collaborative online research platform (DiÖ Virtual Research Environment) for the entire project in order to facilitate or enable

- administration: coordination of individual PPs (with PP01)
- dissemination: fast and easy publication of results
- security: single-sign-on, copyright and access management
- progress: further development of digital humanities methods
Expected Results

A comprehensive and consistent annotation framework covering

• phonology, morphology, syntax/grammar, lexis, discourse/text, multimedia, etc.
• based on international best practices in relevant R&D (Linguistic Annotation Framework etc., see CLARIN strategic standing document on interoperability and standards (TEI, ISO, EAGLES, etc.))
• adapted if needed to variationist linguistics research perspectives and to specific research questions in any of the project parts of DiÖ
DiÖ – Virtual Research Environment

**processing & annotation**
- PoS tagging, using LAF/LMF/MAF/SynAF standards due to research questions

**corpora**
- documentation describing digitising/curating

**collaborative research**
- Corpus-driven vs. -based
detailed linguistic research
- CLARIN-(AT)-Repository

**research data, results**
- documentation/meta-data
- visualisation/aggregation
- networking/linking/curation
DiÖ – VRE Functional Architecture

collaborative workingspace

- transcription → peer discussion
- annotation
- analysis
- analysis

repository

- data
- resources
- DiÖ-data
- publications
- dissemination material

DiÖ - data dissemination material
annotation collaboration corpus building

collaborative writing
Key Methods

Based on methods from computational linguistics, language technologies and digital humanities we will use data modelling, research process modelling and user modelling in order to provide the appropriate environment for the following use cases:

- data search and data access
- data collection and analysis, aggregation
- collaborative research work (e.g. annotation)
- publication and long-term preservation
- reusing research results, also for teaching
Eine Woche, weu i bin auch Fliegenfischer,

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I[v]</td>
<td>er von einem besonders schönen Urlaub?</td>
<td>Mhm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F[v]</td>
<td>meinen letzten Urlaub</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A[v]</td>
<td>itAkk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>II[v]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Oh Gott, o gott. Heuer war ich fischn in Mallnitz.  
  - Mhm 

**Data conversion into XML**

**POS tagging, lemmatising**

<table>
<thead>
<tr>
<th>t27</th>
<th>Heuer</th>
<th>ADV</th>
<th>heuer</th>
</tr>
</thead>
<tbody>
<tr>
<td>t28</td>
<td>war</td>
<td>VAFIN</td>
<td>sein</td>
</tr>
<tr>
<td>t29</td>
<td>ich</td>
<td>PPER</td>
<td>ich</td>
</tr>
<tr>
<td>t30</td>
<td>fischn</td>
<td>ADJD</td>
<td>&lt;unknown&gt;</td>
</tr>
<tr>
<td>t31</td>
<td>in</td>
<td>APPR</td>
<td>in</td>
</tr>
<tr>
<td>t32</td>
<td>Malnitz</td>
<td>NE</td>
<td>&lt;unknown&gt;</td>
</tr>
<tr>
<td>t33</td>
<td>.</td>
<td>$</td>
<td>.</td>
</tr>
<tr>
<td>t34</td>
<td>Mhm</td>
<td>NN</td>
<td>&lt;unknown&gt;</td>
</tr>
<tr>
<td>t35</td>
<td>Eine</td>
<td>ART</td>
<td>ein</td>
</tr>
<tr>
<td>t36</td>
<td>Woche</td>
<td>NN</td>
<td>Woche</td>
</tr>
<tr>
<td>t37</td>
<td>.</td>
<td>$</td>
<td>.</td>
</tr>
<tr>
<td>t38</td>
<td>Fliegenfischer</td>
<td>NN</td>
<td>&lt;unknown&gt;</td>
</tr>
</tbody>
</table>

**harmonisation in line with TEI (Text Encoding Initiative) standards**

**adaptation of the tag set and training of the POS-tagger on ‘German in Austria’, especially spoken variety**

**Meta-data & workflow management**

**storing, archiving, publishing, reusing**

**repository**
The DiÖ Data management plan – basic principles

• All research data are open access by default, designed for re-use by all researchers in follow-up research
• All research data have been generated in a specific stage of a research workflow (in a project part) and additional meta-data include relevant details
• Research data sets are to be published as such, in order to raise their visibility (and that of their creators) and citability, but also linked to publications that refer to them, both with meta-data
• Data scientists in PP 11 assist researchers in PPs 02-10 in coherent and sustainable (meta-)data management, use of research tools and annotation standards, data validation, etc.
• All research data are stored on the virtual research environment for re-use within the project consortium and by default by researchers outside the consortium. Final project results (data and publications) are stored at PHAIDRA or a related research data archival infrastructure
• Roles and responsibilities are clearly defined for the whole project and all its parts (data scientist, data curator, data collector, principal investigator, etc.)
Conclusions

• Research data management has been an integral part of the DiÖ project –
  – during its conception and planning phase
  – Clearly specified in the project application text
  – Being implemented right from the beginning of the project itself as integral part of the (emerging) virtual research environment for DiÖ
  – Embedded in and as part of the national and European research infrastructure CLARIN
  – It is not a goal by itself but specifically serves to
    • increase the quality and efficiency of the project by generating research data according to systematic research data management workflows with coherent data generation methods
    • Raise the visibility of the project, its research data, their creators as researchers, thus increasing the long-term impact of the project
Conclusions

• The systematic research data management approach in DiÖ serves as
  – its methodological proof-of-concept as a showcase for the necessity of research data policies beyond the organisational level
  – as a test lab for improving the RDM methodology in digital humanities and all other fields in particular for large data sets thus providing a showcase for „Big Language Data“

• The role of data scientist and data curator requires specific training and qualifications of researchers (in our case linguists or other humanities researchers), research librarians and archivists, IT specialists, etc.
Thank you for your attention!

Questions? Discussion?

gerhard.budin@univie.ac.at

Acknowledgements:
• Paolo Budroni and team for constant support and encouragement in research data management strategies
• DiÖ Team (Speaker: Alexandra Lenz and team as well as co-project part leaders Stephan Elspaß, Arne Ziegler, Stephan Newerkla, etc.) for their research visions and their interest in digital humanities research data management