COST WG3- Workflow

Preparing for Bolzano meeting

Carole Tiberius
Printed dictionaries

• Planning 30%
• Writing 50%
• Publishing 20%

(Landau 1984: 227)
Computer-lexicographical process

6 phases:

– Preparation
– Data acquisition
– Computerisation
– Data processing
– Data analysis
– Preparation for online release

(Wiegand 1999; Müller-Spitzer 2003; Klosa 2011; forthcoming)
<table>
<thead>
<tr>
<th>Phase</th>
<th>Tasks (selection)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preparation</td>
<td>Dictionary sketch, Organisational plan (Financing, Workflow, Timeplan, Personnel), Pilot study, Dictionary plan (Style Guide, Sample entries, plans for the user interface and access routes, plans for the technical support, plans for user research)</td>
</tr>
<tr>
<td>Data acquisition</td>
<td>Acquisition of sources (Corpora, reference dictionaries, etc.) and other data (e.g. images, videos, sounds)</td>
</tr>
<tr>
<td>Computerisation</td>
<td>Preparation of corpus (Tagging, Lemmatisation), development/acquisition corpus software, development/acquisition DWS, development of web application, acquisition of hardware and software.</td>
</tr>
<tr>
<td>Data processing</td>
<td>Headword and frequency lists, analysis of corpus data, modelling of data structure</td>
</tr>
<tr>
<td>Data analysis</td>
<td>Editing of dictionary article, definition of hyperlinks, linking of dictionary articles and corpus data, integration of multimedia</td>
</tr>
<tr>
<td>Preparation for online release</td>
<td>Proofreading, Testing of user interface (hyperlinks, multimedia, search options, etc.) compiling of help texts.</td>
</tr>
</tbody>
</table>
The ANW
Lexicographical Process

Carole Tiberius & Tanneke Schoonheim
Trier 2012
The Algemeen Nederlands Woordenboek

- Online scholarly dictionary of contemporary standard Dutch in the Netherlands and Flanders
- General (mainly written) language
- 1970 onwards
- Corpus-based: ANW corpus
- Semasiological and onomasiological
- Users: from laymen to professionals
- First results on internet: 2009
- Project duration: 2001 - 2018
- Dictionary under construction
Phases in the Lexicographical Process

• Preparation
• Data acquisition
• Computerisation
• Data processing
• Data analysis
• Preparation for online release

(Klosa, forthcoming)
Phase of preparation

• Organisational Plan
  – Personnel (11fte – 10 fte lexicographic staff; 1 fte IT staff)

• Conceptual Plan
  – Content
    • data structure
    • development of initial type templates for the semagrams
    • sample entries (e.g. bier ‘beer’ and koe ‘cow’)
    • lexicographer’s manual
    • corpus design
Phase of data acquisition

• Primary sources:

**ANW Corpus** (ca. 100,000,000 words and still growing):
  – literary texts
  – newspapers
  – neologisms
  – specific domains (e.g. gardening, sports, web, etc.)

• Secondary sources:
  – **Dictionaries**, e.g. Van Dale, WNT, OED, eLexiko, Wikipedia, Google, Groene Boekje, ...
  – **Linguistics**: Morfologisch Handboek, Algemene Nederlandse Spraakkunst
Phase of computerisation

• Corpus annotation
  – In house tagged and lemmatised
• Installation of software
  – Dictionary Writing System: ANW editor
  – Corpus Query System: Sketch Engine
• Backup and Version Control
Phase of data processing

• Lemma list
• Sketch Grammar
• automatic data acquisition
  – Spelling
  – Regional variation
  – Synonymy
  – Sense distribution
Phase of data analysis

- Lexicographer’s job
- Editing per semantic class
- ....
Data analysis flow

**Preparation:**
- word family
- check of automatically compiled information
- Grammatical information

**Full check**
Check examples
Add multimedia

**Lexicographers**
Completion of full entry:
- Definition
- Combinations, Idioms
- Examples

**Lexicographic Assistants**

**Editor in chief / project manager**

The best way to get a focus on the job you're doing.
Phase of preparation for online release

Editor in chief / project manager

Online release
- Automatic checks
  - Spelling
  - Broken links
- Release on test environment
- Release
Thanks to:

Chief editor: Tanneke Schoonheim
Project manager: Rob Tempelaars
Lexicographers: Lut Colman, Quin van Dijk, Dirk Kinable, Frans Heyvaert, Josefien Sweep, Vivien Waszink
Lexicographic assistants: Arjen Figee, Pieter Hoekstra, Marja Schutte, Jasmijn Splinter, Wies van Velsen
IT-people: Jan Niestadt, Carole Tiberius, ...
<table>
<thead>
<tr>
<th>Vorbereitung</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANW</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elexiko</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Datenbeschaffung</td>
<td>1</td>
<td></td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Computerisierung</td>
<td>1</td>
<td>5</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Datenaufbereitung</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Datenauswertung</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vorbereitung für Onlinerelease</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>