GIVE it a go: TAKE parallel corpora, PUT some statistics, and GET cross-linguistic comparisons!

Natalia Levshina, Thomas Mayer & Michael Cysouw
Philipps University of Marburg

SCL25 Reykjavik, 13 May 2013
Outline

1. GIVE: theoretical background
2. Parallel corpora as a tertium comparationis for cross-linguistic semantic comparisons
3. An onomasiological study of GIVE in 10 European languages
4. Conclusions and perspectives
Semantics of GIVE

• semasiological perspective
  - semantic extensions (polysemy)
  - paths of semantic change
    (Newman’s 1996, von Waldenfels 2012)

• onomasiological perspective
  - how the semantic labour is divided between different expressions of GIVE
  - which conceptual distinctions are cross-linguistically salient
Outline

1. GIVE: theoretical background
2. Parallel corpora as a tertium comparationis for multilingual semantic comparisons
3. An onomasiological study of GIVE in 10 European languages
4. Conclusions and perspectives
Multilingual semantic comparisons

• Elicited data (e.g. Levinson & Meira 2003; Majid et al. 2004; Hartmann et al. In press)
  - problems - decontextualized
  - not natural (different from actual use)
  - only physical stimuli in experiments

• Non-elicited data (parallel corpora)
  - problems - register bias
  - translationese
Corpus-driven semantic maps

• Onomasiological: all instances of items that constitute a semantic field
• Bottom-up: from exemplars to generalizations (relevant dimensions, clusters, prototypes)
• Examples:
  - local phrase markers (Wälchli 2010)
  - motion verbs (Wälchli & Cysouw 2012)
  - letting constructions (Levshina 2012)
Outline

1. GIVE: theoretical background
2. Parallel corpora as a tertium comparationis for cross-linguistic semantic comparisons
3. An onomasiological study of GIVE in 10 European languages
4. Conclusions and perspectives
Aims

1. Theoretical: find common dimensions of variation of GIVE in 10 European languages
   - Germanic: EN, DE, NL, SV
   - Romance: ES, FR, IT, PT
   - Slavic: BG, PL

2. Methodological: will the results vary for different registers?
   - Bible translations
   - film subtitles
Corpora

• Bible translations (New Testament)
  - part of a large-scale DFG project www.paralleltext.info
  - word-aligned with the help of GIZA++

• Subtitles for 10 films (www.opensubtitles.org, www.subscene.com)
  - different original languages and genres, to reduce the translationese effect
  - sentence-aligned with the help of timing information (cf. Tiedemann 2007)
Data set

• Verbs of giving (FrameNet)
  144 different English forms searched for automatically (give, bequeath, contribute, donate, fob, hand, pass, leave, etc.)

• After manual disambiguation and cleaning:

  77 sentences in subtitles (all)
  +
  77 sentences in the Bible (random sample)

154 multilingual contexts
<table>
<thead>
<tr>
<th>ID</th>
<th>Source</th>
<th>EN</th>
<th>BG</th>
<th>DE</th>
<th>ES</th>
<th>FR</th>
<th>IT</th>
<th>NL</th>
<th>PL</th>
<th>PT</th>
<th>SV</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Bible</td>
<td>supply</td>
<td>zapelnjavam</td>
<td>ersetzen</td>
<td>suplir</td>
<td>suppleer</td>
<td>riempiere</td>
<td>vervullen</td>
<td>napelnich</td>
<td>suprir</td>
<td>ge</td>
</tr>
<tr>
<td>3</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
<tr>
<td>4</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
<tr>
<td>5</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>NA</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
<tr>
<td>6</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
<tr>
<td>7</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>zeggen</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
<tr>
<td>8</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
<tr>
<td>9</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
<tr>
<td>10</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
<tr>
<td>11</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
<tr>
<td>12</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
<tr>
<td>13</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
<tr>
<td>14</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
<tr>
<td>15</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
<tr>
<td>16</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>wydach</td>
<td>dar</td>
<td>utge</td>
</tr>
<tr>
<td>17</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>NA</td>
<td>dare</td>
<td>geven</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
<tr>
<td>18</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
<tr>
<td>19</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>wydach</td>
<td>ordenar</td>
<td>ge</td>
</tr>
<tr>
<td>20</td>
<td>Bible</td>
<td>give</td>
<td>davam</td>
<td>geben</td>
<td>dar</td>
<td>donner</td>
<td>dare</td>
<td>geven</td>
<td>dach</td>
<td>dar</td>
<td>ge</td>
</tr>
</tbody>
</table>
Matrix of exemplar dissimilarities

• For each pair of exemplars, compute similarity score $s$: the sum of all similar translations
• Partial similarity = 0.5
• Dissimilarity score $d = 1 - s$
• Multidimensional scaling (smacof package, R)
How many dimensions?
Dutch
Swedish
Português
Spanish
Bulgarian
Dimensions

• Horizontal: distance/contact between the Agent and Recipient
• Vertical: physical vs. non-physical transfer
Clustering the languages

Cluster Dendrogram

Criterion: Does the root contain the canonical *give* in the language?
Is there register variation?
Bible vs. Movies
MDS for Bible only
MDS for movies only

D1

D2

Give

Surrender

Pay

Sell

Treat

Lend

Bequeath

Provide

Give back
Register variation

• The Bible GIVE reveals substantially less variability
• The importance of *yielding* in the Bible

Register variation should be taken into account in future studies based on multilingual parallel corpora
Outline

1. GIVE: theoretical background
2. Parallel corpora as a tertium comparationis for cross-linguistic semantic comparisons
3. An onomasiological study of GIVE in 10 European languages
4. Conclusions and perspectives
Conclusions

• The main dimensions of GIVE in 10 languages involve contact and (non)physical transfer
• Genealogical relatedness determines the way the languages cut the semantic space (basic vocabulary?)
• Register variation is an important factor in onomasiological corpus-driven typology
• More languages are needed to detect areal phenomena and find new dimensions of variation
Thank you!

natalevs@gmail.com