METADATA ANALYSIS FOR HISTORICAL CORPUS LINGUISTICS

Using the Language Change Database

Terttu Nevalainen, Tanja Säily, Turo Vartiainen & Joonas Kesäniemi
WHO ARE WE?

• “Reassessing Language Change: The Challenge of Real Time”
  • Funded by the Academy of Finland, 2014–2018
  • Linguists: Terttu Nevalainen, Tanja Säily, Turo Vartiainen
  • Database architect: Joonas Kesäniemi
  • Assistant: Agata Dominowska
  • Collaborating scholars: Peter Trudgill (sociolinguist), Jefrey Lijffijt, Jukka Suomela (computer scientists), Aatu Liimatta (PhD student)
• **Challenge**: Lack of access to data & previous corpus-based research → research on real-time change largely not cumulative

• **Our solution**: Language Change Database (LCD)
  • Comparative, real-time baseline data for modelling change in progress
  • **Summarises** the results of corpus-based research articles on variation and change in English
  • Can be used to carry out meta-analyses of a large number of linguistic changes
  • Open-access linked data web application, easy to edit & query
Each article is represented by one entry in the database

- Each entry is annotated for linguistic and extra-linguistic features
- **Numerical data** in a computer-readable form
- Initially developed and updated in Helsinki, later each researcher can insert the information of their own work into the LCD
- Currently in beta stage, c. 300 entries
- Search interface now available!

[www.helsinki.fi/lcd](http://www.helsinki.fi/lcd)
LANGUAGE CHANGE DATABASE (LCD)

Study details
- Genre
- Variety
- Grammar
- Dialectology
- Language contact
- Sociolinguistics
- Pragmatics
- Discourse analysis
- Statistical methods

Corpus

Study
- Summary of results
- Time period
- Topics

Publication
- Bibliographic data

Corpus composition file
Annotated data file
Data file
Publication file
“Without except(ing) unless…”: on the grammaticalisation of expressions indicating exception in English
Rissanen, Matti
2002

ENOUGH and ENOW in Middle English
Peitsara, Kirsti
1997

Epicene HE and THEY and the development of English indefinite expression during the period 1500-1800
Laitinen, Mikko
2007

HERE compounds in English: mere satellites of THERE compounds?
Österman, Aune
2007
TILL, a loan connective (subordinator or preposition), borrowed from Old Norse, occurs a few times in Old English texts written in the Northumbrian dialect area, as an equivalent of TO.

- It replaces OP in the temporal and local senses in Early Middle English.
- The earliest instances recorded in the Helsinki Corpus occur in East Midland texts where the Scandinavian influence was the strongest.
- The related compound form UNTIL was borrowed in Middle English; the earliest occurrence occurs in the Órmulum.
- It becomes common in the fourteenth century.
- It is, at least to some extent, dialectally restricted throughout the Middle English period (Northern dialect; Northumbrian, East Midland dialect).
- In Modern English, UNTIL supersedes TILL in frequency numbers. It is the more common form particularly in formal registers. In letters and drama, TILL remains common even in twentieth-century corpus samples. UNTIL is clearly favoured by written language, while TILL is proportionally more common in spoken language. Particularly the preposition TILL seems a usage typical of speech: its relative frequency is even higher than that of UNTIL.

- Register:
  - TILL (speech)
  - UNTIL (writing)
### Table 1. Until and till in the Early Modern English sub-sections of the Helsinki Corpus. Figures per 100,000 words in brackets. (2005)

<table>
<thead>
<tr>
<th></th>
<th>until</th>
<th>till</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>subord.</td>
<td>prep.</td>
</tr>
<tr>
<td><strong>EModE1 (1500-1570)</strong></td>
<td>24</td>
<td>11</td>
</tr>
<tr>
<td><strong>EModE2 (1570-1640)</strong></td>
<td>43</td>
<td>16</td>
</tr>
<tr>
<td><strong>EModE3 (1640-1710)</strong></td>
<td>64</td>
<td>2</td>
</tr>
</tbody>
</table>
CASE STUDY

• At ISLE4, we presented a manual meta-analysis of **connectives** in the history of English
  • Data from Matti Rissanen’s published research
  • Testing the notion of **punctuated equilibrium** using the LCD
  • The Civil War Effect? (Raumolin-Brunberg 1998, Trudgill 2010)
• Example: **till** and **until**
HISTORY OF ENGLISH AS PUNCTUATED EQUILIBRIA

(Bergs 2005: 54, based on Dixon 1997)
until and till in the long diachrony
FROM MANUAL TO SEMI-AUTOMATED META-ANALYSIS

- Annotated data tables implemented with **Excel styles**
- Connects spreadsheet cells to linked-data concepts in a machine-readable form

### Table 1. Until and till in the Early Modern English sub-sections of the Helsinki Corpus. Figures per 100,000 words in brackets. (2005)

<table>
<thead>
<tr>
<th>Time period</th>
<th>Absolute values</th>
<th>Corpus annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>EModE1 (1500-1570)</td>
<td>24 11 35 (18.4)</td>
<td>EModE1 HC 61 (32.1)</td>
</tr>
<tr>
<td>EModE2 (1570-1640)</td>
<td>43 16 59 (31.1)</td>
<td>EModE2 HC 83 (43.7)</td>
</tr>
<tr>
<td>EModE3 (1640-1710)</td>
<td>6 3 9 (5.3)</td>
<td>EModE3 HC 146 (85.4)</td>
</tr>
</tbody>
</table>

**Expression**

<table>
<thead>
<tr>
<th>until subord.</th>
<th>until prep.</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Function**

<table>
<thead>
<tr>
<th>till subord.</th>
<th>till prep.</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Meta-analysis for historical corpus linguistics/ Nevalainen, Säily, Vartiainen & Kesäniemi 17/10/2018**
TOOL FOR RE-USING LCD DATA

• LADA = LCD Aggregated Data Analysis workbench
  • Tool for *experimenting* with LCD data
  • Small-scale *meta-analysis* across studies
• Generates RDF Data Cubes of the annotated data tables
• Provides a *workflow* for creating new aggregated datasets

**Diagram:**
- Annotated data tables → Potentially relevant data → Relevant data → Distributable data
  - Select → Filter & group → Review & refine → Normalize
  - RDF data cubes → Normalized data → Visualize

**Corpus compositions:**
- 17/10/2018
- Meta-analysis for historical corpus linguistics/
  Nevalainen, Säily, Vartiainen & Kesäniemi
### Filter and group

#### Corpora
- HC
- ARCHER
- LOB
- F-LOB

#### Expressions
- UNTIL
- TILL

#### Genres
- Any or no value

#### Functions
- Any or no value

#### Time period
- Some timeperiod

### Results (1)

**The development of TILL and UNTIL in English**

*Tables: 3 Values: 74*
Source tables and filtered values

The development of TILL and UNTIL in English
Rissanen, 2005

Table 1
Table 1. Until and till in the Early Modern English sub-sections of the Helsinki Corpus. Figures per 100,000 words in brackets. (2005)

<table>
<thead>
<tr>
<th></th>
<th>until</th>
<th>until</th>
<th>till</th>
<th>till</th>
<th>till</th>
<th>till</th>
<th>HC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>subord.</td>
<td>prep.</td>
<td>total</td>
<td>subord.</td>
<td>prep.</td>
<td>total</td>
<td></td>
</tr>
<tr>
<td>EModE1 (1500-1570)</td>
<td>24</td>
<td>11</td>
<td>35 (18.4)</td>
<td>45</td>
<td>16</td>
<td>61 (32.1)</td>
<td></td>
</tr>
<tr>
<td>EModE2 (1570-1640)</td>
<td>43</td>
<td>16</td>
<td>59 (31.1)</td>
<td>57</td>
<td>26</td>
<td>83 (43.7)</td>
<td></td>
</tr>
<tr>
<td>EModE3 (1640-1710)</td>
<td>6</td>
<td>3</td>
<td>9 (5.3)</td>
<td>87</td>
<td>59</td>
<td>146 (85.4)</td>
<td></td>
</tr>
</tbody>
</table>
Normalize

The development of TIL and UNTIL in English
Rissanen, 2005

Table 1
Table 1. Until and till in the Early Modern English sub-sections of the Helsinki Corpus. Figures per 100,000 words in brackets. (2005)

Table 5
Table 5. Until and till in five genres of the Archer Corpus. Absolute figures. (2005)

Table 6
Table 6. Until and till in Present-day English corpora. (2005)

Frequencies

<table>
<thead>
<tr>
<th>Group</th>
<th>Absolute</th>
<th>Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC</td>
<td>24</td>
<td>18.41</td>
</tr>
<tr>
<td>UNTIL</td>
<td>11</td>
<td>3.08</td>
</tr>
</tbody>
</table>

Frequencies

<table>
<thead>
<tr>
<th>Group</th>
<th>Absolute</th>
<th>Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC</td>
<td>45</td>
<td>16.45</td>
</tr>
<tr>
<td>TILL</td>
<td>16</td>
<td>3.08</td>
</tr>
</tbody>
</table>

Frequencies

<table>
<thead>
<tr>
<th>Group</th>
<th>Absolute</th>
<th>Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC</td>
<td>43</td>
<td>16.09</td>
</tr>
<tr>
<td>UNTIL</td>
<td>16</td>
<td>3.08</td>
</tr>
</tbody>
</table>

Frequencies

<table>
<thead>
<tr>
<th>Group</th>
<th>Absolute</th>
<th>Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC</td>
<td>57</td>
<td>26.73</td>
</tr>
<tr>
<td>TILL</td>
<td>1570-1640</td>
<td>43.73</td>
</tr>
</tbody>
</table>
CONCLUSIONS

• The Language Change Database provides baseline data for meta-analyses, replication studies and systematic reviews.

• It is intended to make the field of English historical corpus linguistics more cumulative by ensuring easy access to the results of earlier research.

• The LADA tool can be used to experiment with the numerical data included in the LCD entries and to carry out small-scale meta-analyses.

• Both the LCD and LADA will be made available in accordance with the best practices of open science.
REFERENCES

http://www.helsinki.fi/lcd/


• HC = *The Helsinki Corpus of English Texts*. 1991. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English). Helsinki: Department of Modern Languages, University of Helsinki. http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/

