Large Mailing List Corpora: Management, Annotation and Repository

Damir Cavar, Helen Aristar-Dry, Anthony Aristar

ILIT, EMU (The LINGUIST List)

Agenda

- Goals
- Corpus and Data
- Linguistic Annotations
- Indexing and search technology

Goals: Exploration

- Create a frontend for corpus-linguistic analysis of monotonically increasing corpora
- Enable virtualization of balancing and virtualization of corpora
- Create large scale language resources for language models (technologies) and linguistic research with annotations at multiple linguistic levels (morpho-syntactic, syntactic, semantic, functional)
- …

Goals

- Create linguistically annotated corpora from professional mailing lists
  - Mining and extraction of domain specific information/knowledge
- Enable a study of the developments in a specific field along the timeline, in our case: Linguistics

Goal

- Explore the possibilities to create an infrastructure for linguists to:
  - Make a corpus
  - Annotate a corpus
  - Share a corpus
  - Use other peoples corpora
  - Map annotations to e.g. GOLD concepts or another standard

Data

- The LINGUIST List (LL) is hosting
  - Currently 238 mailing lists stored that are related to linguistic topics: LINGUIST List, Corpora-List, Childes etc.
  - Storage:
    - Original Listserv mail format
    - Database tables in a relational DB-server
    - HTML versions of the mails are generated dynamically from the two source formats for online access from the LINGUIST List pages

Data

- Highly active lists since 1990
  - 195,782 postings estimated
  - growing daily
  - Valuable content:
    - Book reviews
    - Dissertation abstracts
    - Journal TOCs
    - Etc.

The LINGUIST List

- LL-mails are moderated and edited
- Special list submission interfaces for some types of mailings
  - LL-mails are structured and some text elements are typed (e.g. Named Entities, abstracts)
  - Exporting these moderated and edited mails gives us a high quality annotation without additional effort and a low error rate
- LL is multi-lingual

Parameters

- Growing corpus
- Some of high quality (e.g. LL), some not edited with poor linguistic quality
- Linguistic components that are qualitatively more or less reliable
- How much can be done wrt. quality and quantity?

Exploring

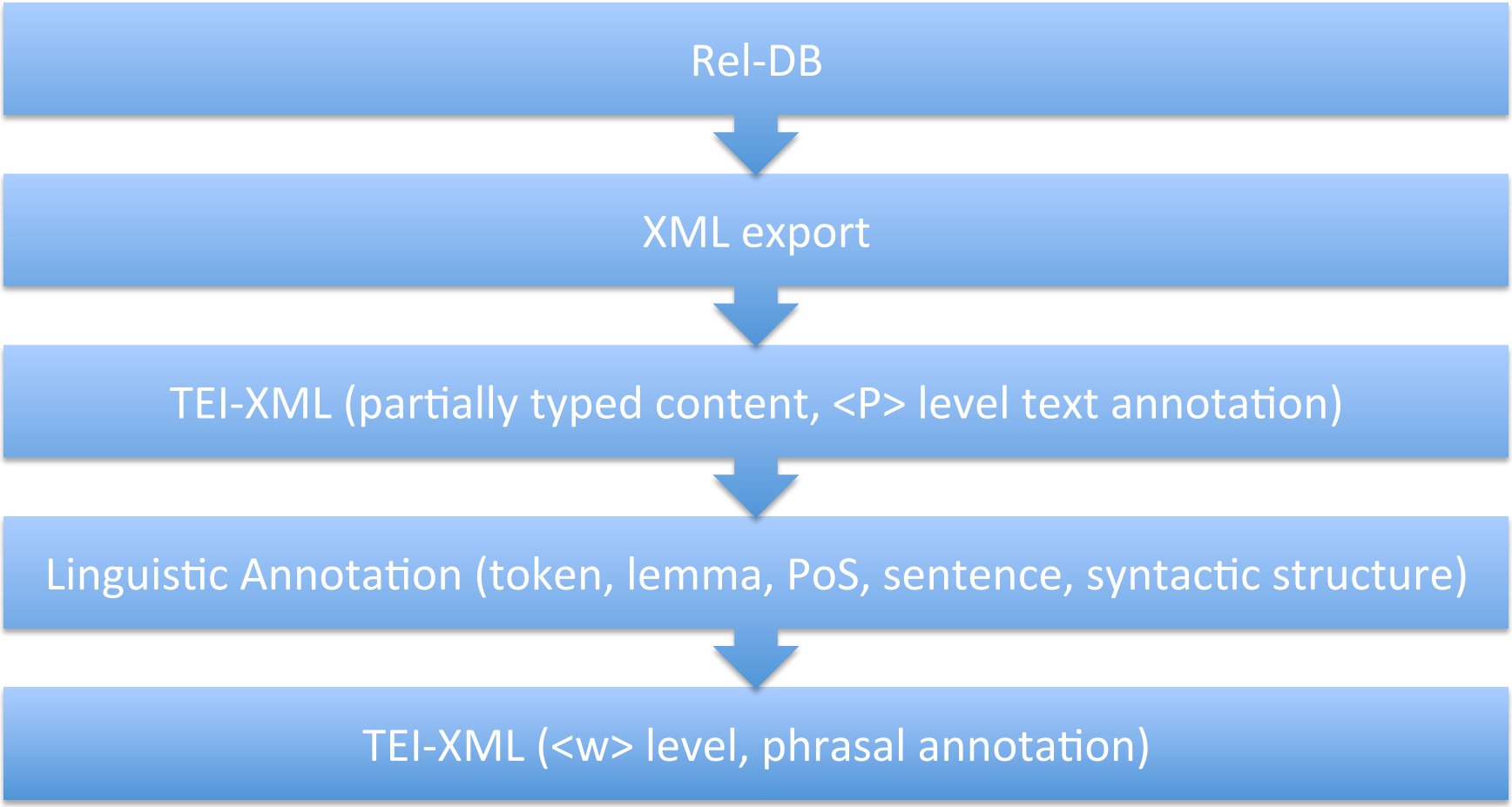
- Redesign a more advanced search functionality
- Infrastructure to advance search and content analysis using
  - Automatic linguistic annotation for the mailing list corpus, statistical methods
  - Bring together new conversion, annotation and storage concepts that facilitate efficient and flexible corpus analyses
- Make specific content available for detailed mining and knowledge extraction

Automatic Conversion

- Generation of one coherent file format from Listserv mails, HTML, DB-tables
- Conversion to some intermediate XML (table-dump to XML)
- Translation to TEI P5 XML documents

```
<dissertation id="2749">
 <disstitle>some title</disstitle>
 <institution_name>University of Edinburgh
 </institution_name>
 <progtitle>School of Informatics
 </progtitle>
 <degreedate>2009</degreedate>
 <dissstatus>In Progress</dissstatus>
 <dissabstract>...</dissabstract>
 <people>
  <person id="111660" role="Author">
   <personfn>David</personfn>
   <personmi></personmi>
   <personln>Smith</personln>
   <institution>Universität Potsdam
   </institution>
  </person>
  <person id="653"
   role="Dissertation Director">
   <personfn>John</personfn>
   <personmi></personmi>
   <personln>Johnston</personln>
   <institution>University of Edinburgh
   </institution>
  </person>
 </people>
```

Conversion of LL-mails



TEI Wrapper

- Meta information about
  - Language
  - Linguistic domain
  - General topic in the lists taxonomy
  - Editor
  - Title and Volume ID
  - Etc.

```
<classDecl>
  <taxonomy xml:id="topic">
    <category>
      <catDesc>Topic</catDesc>
    </category>
  </taxonomy>
  <taxonomy xml:id="lingfield">
    <category>
      <catDesc>Linguistic Field</catDesc>
    </category>
  </taxonomy>
</classDecl>
```

```
<teiHeader>
   <fileDesc>
    <titleStmt>
     <title>All: Editors' Comment; Query on religious language</title>
     <sponsor>The LINGUIST List</sponsor>
     <editor></editor>
    </titleStmt>
    <editionStmt>
     <edition>TEI XML edition 1 <date when="2012-02-27">Mon Feb 27 2012</date></edition>
    </editionStmt>
    <publicationStmt>
     <publisher>The LINGUIST List</publisher>
     <address>
          <addrLine>2000 E Huron River Dr., Ypsilanti, MI </addrLine>
     </address>
     <date when="1990-12-15">1990-12-15</date>
     <idno>1.1</idno>
     <distributor>The LINGUIST List</distributor>
     <availability>
      <p>The LINGUIST List</p>
     </availability>
    </publicationStmt>
    <sourceDesc>
     <bibl>
      <idno>1.1</idno>
      <title>All: Editors' Comment; Query on religious language</title>
      <editor></editor>
      <date when="1990-12-15">1990-12-15</date>
      <publisher>The LINGUIST List</publisher>
      <ptr target="http://linguistlist.org/issues/1/1-1.html"/>
     </bibl>
    </sourceDesc>
   </fileDesc>
```

<text>
 <body>

   <head>Message 1: Syntactic Analysis:
        Sobin</head>

   Date: <date>20-Dec-2011</date><lb/>
     From: Kristen Holding
      <email>kholding@wiley.com</email><lb/>
     Subject: Syntactic Analysis: Sobin
   Title: Syntactic Analysis<lb/>
     Subtitle: The Basics<lb/>
     Published: 2011<lb/>
     Publisher: Wiley-Blackwell<lb/>
     <ref target="...">...</ref>
   Book URL:
     <ref target="...">...</ref>
   Author: Nicholas Sobin<lb/>
     Hardback: ISBN: 9781444338959 ...<lb/>
     Paperback: ISBN: 9781444335071 ...

   <head>Abstract:</head>
   ...

   Linguistic Field(s):
     Applied Linguistics<lb/>
     Syntax
   Written In: English (eng)
   See this book announcement ...<lb/>
   <ref target="...">...</ref>

 </body>
</text>

```
<p>
 <w lemma="title" type="NN">Title</w>
 <pc>:</pc>
 <w lemma="syntactic"
  type="JJ">Syntactic</w>
 <w lemma="analysis"
  type="NN">Analysis</w><lb/>
 <w lemma="subtitle"
  type="NN">Subtitle</w>
 <pc>:</pc>
 <w lemma="the" type="DT">The</w>
 <w lemma="basic"
  type="NN">Basics</w><lb/>
 <w lemma="publish"
  type="VBN">Published</w>
 <pc>:</pc>
 <date>2011</date><lb/>
 <w lemma="publisher"
  type="NN">Publisher</w>
 <pc>:</pc>
 <name type="publisher">
 <w lemma="Wiley-Blackwell"
  type="NNP">Wiley-Blackwell</w>
 </name>
 <lb/>
 <ref target="http://www.wiley.com">
  http://www.wiley.com</ref>
</p>
```

Linguistic Annotation

- Using existing components in
  - GATE, UIMA
  - Stanford CoreNLP or RASP
  - Own components: Finite State Automata (XFST/ Foma-based) for Lithuanian, Croatian…
  - Commercial components: Xelda tools (cover 26 languages)

Linguistic Annotation

- Hardware:
  - Two 2.4GHz Quad-Core Intel Xeon "Westmere" processors, 12GB RAM
- Stanford CoreNLP (Java-based)
  - Linguistic Components running as service called via RPC
  - Approx. 15 tokens per second, or 18 hours for 1 mil. tokens

Linguistic Annotation

- Finite State components
  - Example: Croatian or Lithuanian XFST-based morphology for lemmatization and Part-of-speech tagging (onomastic typing)
    - 3773 tokens per second, or approx. 4 minutes for 1 mil. tokens
    - Self-generated finite state transducers: Croatian, German analyze 30.000 to 50.000 tokens a second, less then 30 seconds for 1 mil. tokens

Indexing and Interface

- Using Philologic 3.x
  - http://ltl.emich.edu/llc/
  - http://ltl.emich.edu/philologic/
- Features:
  - Pre-generated N-gram models, KWIC, meta-based limitations, text mining extensions
  - Processes TEI XML files
    - Meta information
    - Tokenization, Lemmatization, …

Indexing and Interface

- Philologic 3.x:
  - Uses MySQL for meta data
  - File storage on the file-system
  - Mobile-computing enabled (HTML & JavaScript)
- Issues:
  - Tricky to set up
  - Re-indexing is not incremental and time consuming, i.e. acceptable for static corpora
    - 1 mil. tokens in 1 minute, exponential time increase with number of tokens

Philologic

- Philologic 4: Currently being entirely ported to Python 3
  - More flexibility for extension and integration with other technologies
- Possibility:
  - Improve indexing and search over dynamic corpora with alternative storage technologies

Indexing and Storage

- NoSQL component: Redis
  - Key-value storage of a specific type: data-structure storage
    - Key associated with value and both are standard simple datatypes
    - Key associated with complex data-structures like lists or sets with advanced manipulation operations
      - Atomic push and pop and shifts on the lists on heads or tails
      - Range and slicing operators like in Python or Ruby
      - Sets (unique sets of keys): apply set intersection on the Redis server, e.g. tag a set or sub-set of keys and extract the intersection (e.g. all nouns, all Nominative, all with a specific frequency etc.)

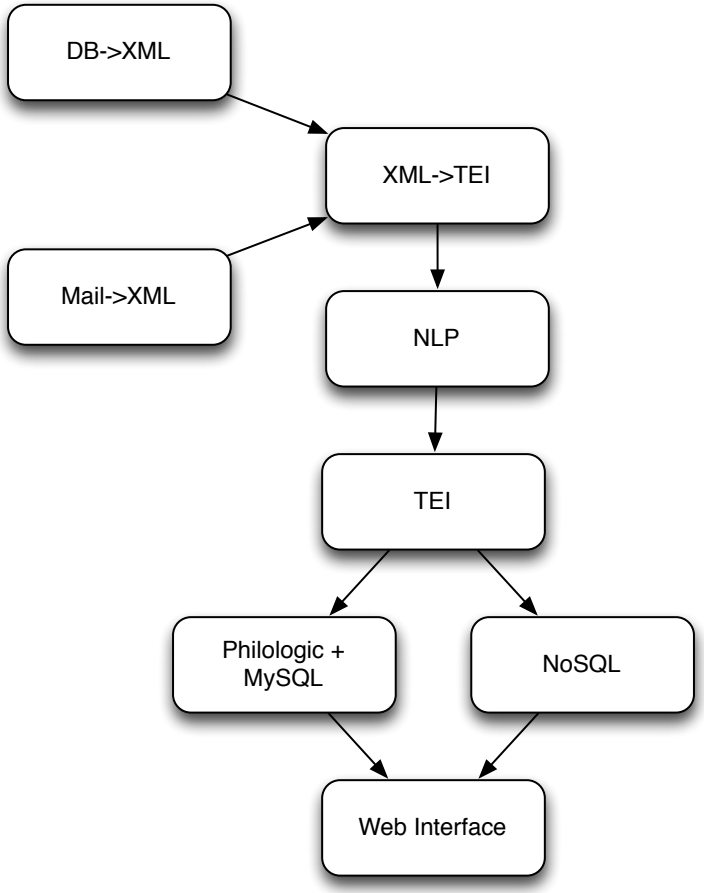
Indexing and Storage

- NoSQL component: Redis
  - Fast, very fast: in-memory storage and operations with asynchronous persistency forks
    - Speed factors: Evaluation reports
      - SET and GET of 256-byte strings
      - 110.000 SET or GET operations in a second

Indexing and Storage

- NoSQL component: Redis
  - Persistent (asynchronously)
    - Possibly loose keys if there is any type of enforced downtime
  - Limitations:
    - 2^{32} keys, 2^{32} elements in value should be enough to cover a large corpus (4.3 bil. tokens)

Overview


DB->XML
XML->TEI
Mail->XML
NLP
TEI
Philologic +
MySQL
NoSQL
Web Interface

Conclusion

- Integration of alternative storages in the Philologic code-base
- Creation of a new interface with possibilities to work with trees, extend the annotations etc.
- Manual check of the automatic annotations, or cross-validation with alternative components
- Visualization