Interoperability of Second Language Resources and Tools

Elena Volodina\textsuperscript{1}, Maarten Janssen\textsuperscript{2}, Therese Lindström Tiedemann\textsuperscript{3}, Nives Mikelić Preradović\textsuperscript{4}, Silje Ragnhildstveit\textsuperscript{5}, Kari Tenfjord\textsuperscript{6}, Koenraad de Smedt\textsuperscript{6}

\textsuperscript{1} University of Gothenburg, \textbf{Sweden}; \textsuperscript{2} University of Coimbra, \textbf{Portugal}; \textsuperscript{3} University of Helsinki, \textbf{Finland}; \textsuperscript{4} University of Zagreb, \textbf{Croatia}; \textsuperscript{5} Western Norway University of Applied Sciences, \textbf{Norway}; \textsuperscript{6} University of Bergen, \textbf{Norway}
What is **Interoperability (of Language Resources)**?

*Ideal picture*
What is Interoperability (of Language Resources)?

Real-life picture

- type A
- type B
- type C
- type D
- type E
- type F
- type G
- type H
- type I
- type J
- type K
- type L
What is **Interoperability of Language Resources**?

*by Chiarcos, 2012*

- **Structural**
  - **Annotations** of different origin are represented using the same formalism (e.g. stand-off XML or RDF databases)

- **Conceptual**
  - **Annotations** of different origin are linked to a common vocabulary (terminological reference repository)

What is Interoperability of Language Resources?

by Foulonneau & Riley, 2014

• **Metadata**
  - Descriptions of the data; resource discovery in search engines, portals and registries. (+filtering?)

• **Technical**
  - Data aggregation

• **Content**
  - Comparable content of the resources – based on metadata

---

What is **Interoperability of Language Resources**?  
*by Ide & Pustejovsky, 2010*

- Metadata
  - characteristics of data expressed through a set of labels (syntactic dimension) and categories (semantic dimension)
- Data categories and their semantics
  - e.g. morpho-syntax, syntax, text typologies, etc.
- Requirements for publication of data and notations
  - common practices for creating, documenting and evaluating language resources, e.g. agreement on formats and access; encoding; copyright; etc.
- Requirements for software sharing
  - software formats, data formats, software integration platforms; possibility to combine different tools; evaluation of software; copyright

Interoperability of Second Language Resources and Tools

- Metadata
- Error taxonomies
- Tools
- User interfaces


Workshop on Interoperability of Second Language Resources and Tools
6-8 Dec. 2017 | University of Gothenburg, Sweden
https://sweclarin.se/swe/workshop-interoperability-l2-resources-and-tools

CONTRIBUTION TO CLARIN GOALS
- Promotion of the interoperability of resources and tools in CLARIN by working towards common guidelines for L2 annotation and metadata
- Promotion of the integration of data, tools and services making L2 corpora compatible with corpus tools in CLARIN
- Creating and expanding the CLARIN network of experts in the area of L2 corpora and tools

INVITED TALKS and main program points
Towards standardization of metadata for L2 corpora:
Sylviane Granger, University of Louvain, Belgium

Program in a nutshell:
- Existing corpora
- Corpora under construction
- L2 infrastructures
- Metadata and ethics
- Error annotation

- Tools and software
- Happy user
- Developments on top of L2 corpora
- Discussion à la World café
- Social program

Participants and organizers
Organizers:
- Elena Volodina, Sweden
- Kari Tenfjord & Silje Ragnhildstveit, Norway
- Therese Lindström Tiedemann, Finland
- Nives Mikelic Preradovic, Croatia
- Maarten Janssen, Portugal

Participants:
15 countries; 27 participants
L2 metadata
by Granger and Paquot, 2017

• Administrative
  - title, license, availability, ...

• Corpus design
  - L1s, L2s, size, mode, levels, guidelines, ...

• Annotation
  - type: POS, syntax, errors; tagsets, guidelines, tools, ...

• Text
  - mode, author, title, statistics, task types & instructions, ...

• Learner
  - age, gender, L1s, L2s, level, education, ...

L2 metadata
by Granger and Paquot, 2018

- Administrative
  - title, license, availability, ...
- Corpus design
  - L1s, L2s, size, mode, levels, guidelines, ...
- Annotation
  - type: POS, syntax, errors; tagsets, guidelines, tools, ...
- Text
  - mode, author, title, statistics, task types & instructions, ...
- Learner
  - age, gender, L1s, L2s, level, school, education, ...


L2 metadata
in present-day LCR projects

- Varied between L2 corpora
  - no track of various aspects, e.g. Tasks, guidelines, etc
- Restricted by laws and agreements
  - e.g. aggregated birth year spans in one corpus versus exact birth year in another
- Incompatible
  - e.g. Bosnian, Serbian and Croatian L2s separate in one corpus versus BSC in another
Examples: Korp, Swedish edition
Error taxonomy

**an ideal**

- Same error classification approach across L2 corpora
  - e.g. based on linguistic description (phonology, orthography, morphology, ...) (Dobrić 2015)
- Same granularity
  - 22 tags versus 65 tags
- Theory-independent approach
  - (Tenfjord et al 2006)
- Piloting
  - test on project members first to avoid unreliable / confusing tags
- Annotation
  - e.g. normalization first, error code afterwards (Volodina et al. 2018)
- Annotation quality
  - documented inter-annotator agreement, etc. (Fort 2016)
**Error taxonomy**

*an ideal*

- Same error classification approach across L2 corpora
  - e.g. based on linguistic description (phonology, orthography, morphology, ...) (Dobrić 2015)
- Same granularity
  - 22 tags versus 65 tags
- Theory-independent approach
  - (Tenfjord et al 2006)
- Piloting
  - test on project members first to avoid unreliable / confusing tags
- Annotation
  - e.g. normalization first, error code afterwards (Volodina et al. 2018)
- Annotation quality
  - documented inter-annotator agreement, etc. (Fort 2016)

---

**Error taxonomy**

*in present-day LCR projects*

Taxonomies are like underwear; everyone needs them, but no one wants someone else’s.

(From a presentation by Egon Stemle at CLARIN workshop on interoperability of L2 resources and tools)
Example: ASK taxonomy in SweLL pilot

Gloss: Central Statistical Agency [...] also in a report from 2001 [shows (finite verb)] that stress-related and...

Error code explanations: INV: Non-application of subject/verb inversion, OINV: Application of subject/verb inversion in inappropriate contexts, O: other word (or phrase) order error.
Tools

- Accessible
- User-friendly
- Well-documented
- Accompanied by user manuals
- Collected in one repository for re-use
- Annotation quality
Pluralism of formats and outputs, often inaccessible, or propriatory.
Some examples:
- Feat (Hana et al 2010)
- TEITOK (Janssen 2016)
- SVALA (Rosén et al 2018)
- Falko-tools (Müller & Strube 2006)

Tools
in present-day LCR projects

- Accessible
- User-friendly
- Well-documented
- Accompanied by user manuals
- Collected in one repository for re-use
- Annotation quality

Tools
an ideal
Some tools

SVALA

He get to cleaned his son.

M-verb O M-verb

He got his son to clean the room.

CLARIN
Some annotation formats

<sic type="F" desc="AGR" corr="stilling">stillingen</sic>

<word lemma="dette" features="pron nøyt ent pers @subj">dette</word>
User interfaces

- Basic and advanced modes
- Selection of error types to be correlated with speaker features
- Metadata re-use & filtering
User interfaces

**an ideal**

- Basic and advanced modes
- Selection of error types to be correlated with speaker features
- Metadata re-use & filtering

---

User interfaces

**in present-day LCR projects**

- Various **formats** exist, but most of the search tools rely on xml format → need to have a TEI-conformant version of all corpora
- **Search builder** – varies between interfaces
  - Not all **metadata** is visualized or is made searchable
  - SLA researchers are often ”scared”
Some user interfaces

Query interface

Concordance

Word List

Text


Jeg tror andre land er også muligheter for mennesker som kan komme til Norge for å jobbe. - Kanskje de har problemer med å finne en jobb i det eget land. De kan ikke enkelt vise av å jobbe i en annen land. Men ganger er det på bekostning av familie og vennene som de må fortale.

Det er Norges kum som inn alle forskjellige land er veldig bra for norske folk. De kan være kjent med andre kulturer, andre religioner. Kanskje de kan finne seg vinner.

Integrasjon er viktig for utlendinge. Hvis de skal trives her i Norge må de ta pass på at de nordiske samfunnet. Mysighetene og arbeidsgiveren kan gjøre mye på dette området. Nesen ganger trenger det å skifte hver 5
Prospects

• Non-trivial work, time-consuming, community depending
  • a lá Universal Tagset / Universal Dependencies

• Pluralism (in tools and formats) is healthy, BUT we need a common conversion mechanism, a lá transformators, between the pluralistic approaches

• Need for insights from several perspectives
  • Second Language Acquisition researchers, (learner corpora) linguists, teachers, language testing specialists
  • NLP researchers, software engineers, Systems developers
References 1


References 2


• **Tenfjord, K., Johansen, H., & Hagen, J. E.** (2006). The" Hows" and the" Whys" of Coding Categories in a Learner Corpus (or" How and Why an Error-Tagged Learner Corpus is not "One Big Comparative Fallacy"). *Rivista di psicolinguistica applicata*, 6(3).

Thank you!

Our aim

Type C  Europlug fits all, more than 130 countries
Follow-up

- **Suggestions for future:** [https://goo.gl/bW24Sq](https://goo.gl/bW24Sq)

- **Joint publication** in LCR 2018 post-conference volume (accepted, publication date 2019):
  - Egon W. Stemle (Italy), Adriane Boyd (Germany), Maarten Janssen (Portugal), Therese Lindström Tiedemann (Finland), Nives Mikelić Preradović (Croatia), Alexandr Rosen (Czech Republic), Dan Rosén (Sweden), Elena Volodina (Sweden). *Working together towards an ideal infrastructure for language learner corpora.*

- **CLARIN survey of L2 corpora:**
  - *L2 learner corpus survey – Towards improved verifiability, reproducibility and inspiration in learner corpus research.* By Therese Lindström Tiedemann, Jakob Lenardič and Darja Fišer. CLARIN 2018

- **COST action:** application is planned

- **Follow-up workshop:** is planned
Some user interfaces

Swedish Korp

Search query & filters

Task description

Hits in KWIC format

Metadata

Statistics
Some user interfaces

**TEITOK**

KWIC/XML

Editable metadata

Search & Filters

Statistics

Document

CLARIN
What is **Interoperability of Language Resources**?  
*by Ide & Pustejovsky, 2010*

...a measure of the degree to which diverse systems, organizations and/or individuals are able to work together to achieve a common goal.

**• For computers**
  - **syntactic** interoperability (data formats, communication protocols, data exchange)
  - **semantic** interoperability (ability to automatically interpret exchanged information via a common information exchange reference model)

**• For language resources**
  - focus is rather on **semantic** interoperability, since syntactic ones are technically mappable via a trivial conversion

---

Some tools

feat
Some user interfaces
Some user interfaces

ANNIS

<table>
<thead>
<tr>
<th>learner</th>
<th>von</th>
<th>20</th>
<th>in</th>
<th>Stadt</th>
<th>X</th>
<th>exestierte</th>
<th>Baugenossenschaften</th>
</tr>
</thead>
<tbody>
<tr>
<td>TH1</td>
<td>von</td>
<td>20</td>
<td>in</td>
<td>Stadt</td>
<td>X</td>
<td>existierten</td>
<td>Baugenossenschaften</td>
</tr>
<tr>
<td>TH1Diff</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CHA</td>
</tr>
<tr>
<td>TH2</td>
<td>von</td>
<td>20</td>
<td>in</td>
<td>Stadt</td>
<td>X</td>
<td>existenten</td>
<td>Baugenossenschaften</td>
</tr>
<tr>
<td>TH2Diff</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CHA</td>
</tr>
<tr>
<td>EA_category</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>G_Morphol_Wrong</td>
<td></td>
</tr>
<tr>
<td>EA_category</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>O_Graph</td>
<td></td>
</tr>
<tr>
<td>EA_category</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>V_semdenot_word_fs</td>
<td></td>
</tr>
</tbody>
</table>
"Standards are like toothbrushes; everyone likes the idea of them, but everyone wants to use their own" #FORCE2017
@Metadata2020