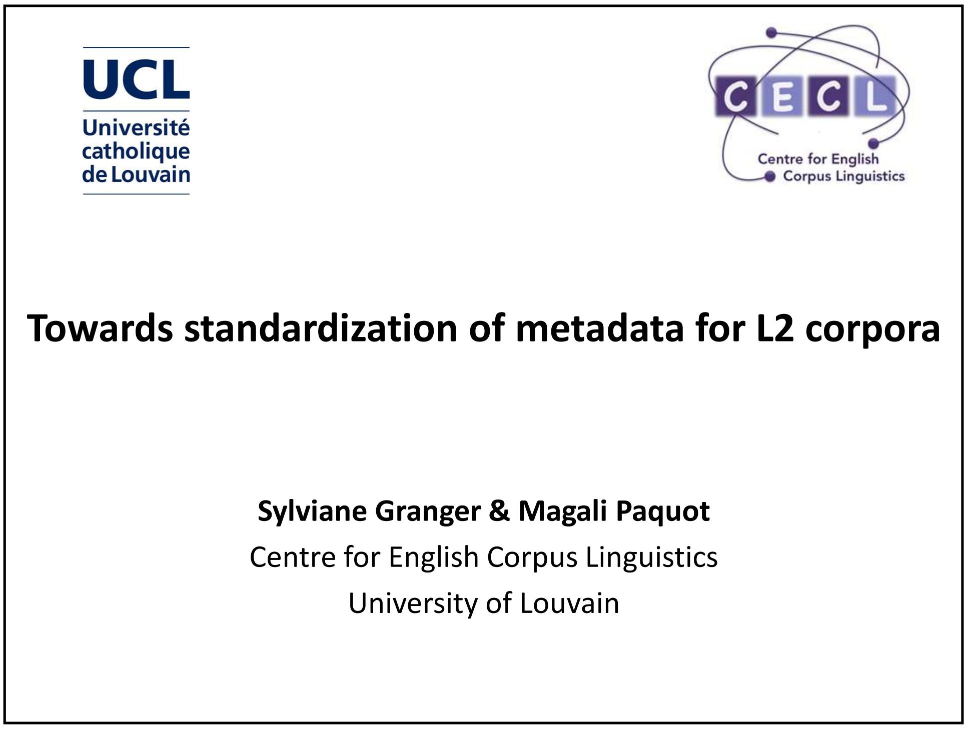
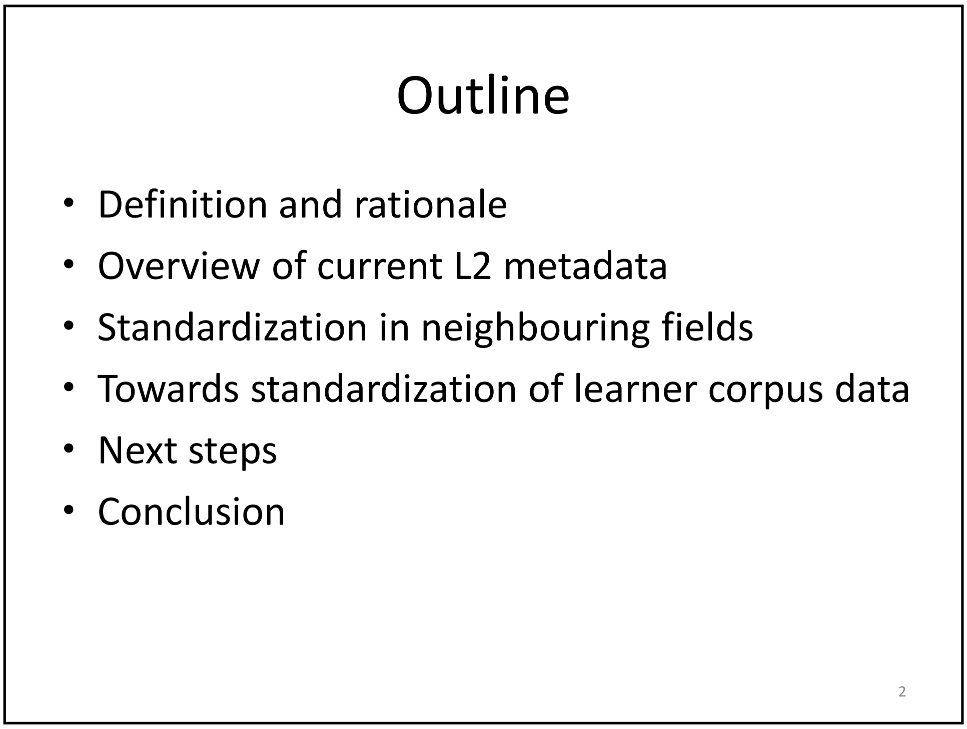


Towards standardization of metadata for L2 corpora

Sylviane Granger & Magali Paquot
Centre for English Corpus Linguistics
University of Louvain



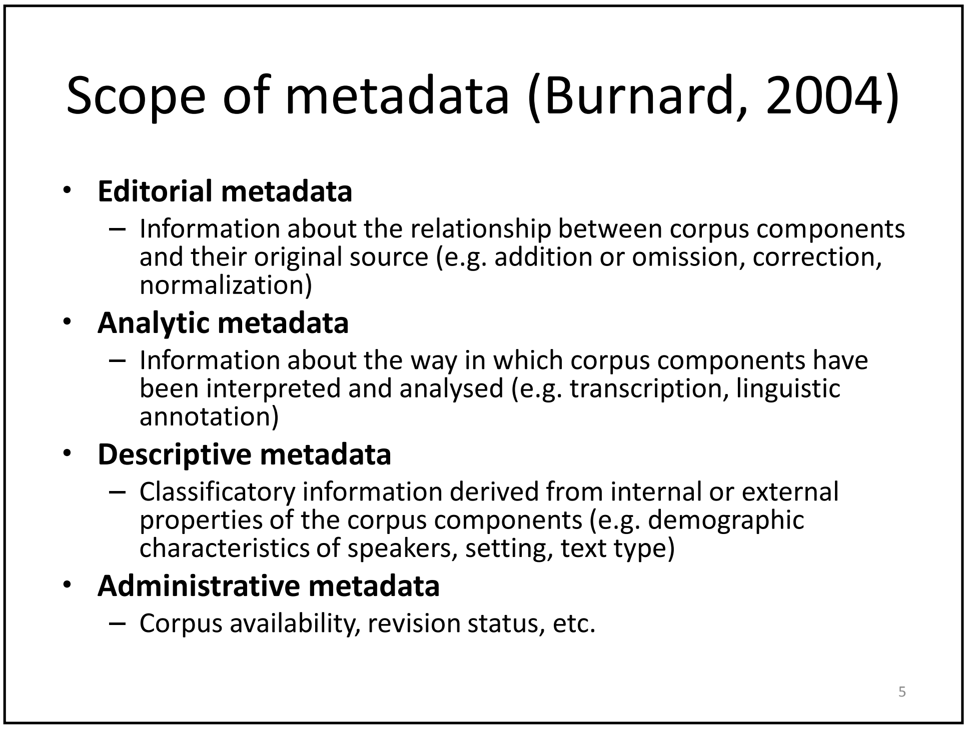
Outline

- Definition and rationale
- Overview of current L2 metadata
- Standardization in neighbouring fields
- Towards standardization of learner corpus data
- Next steps
- Conclusion

Definition and rationale

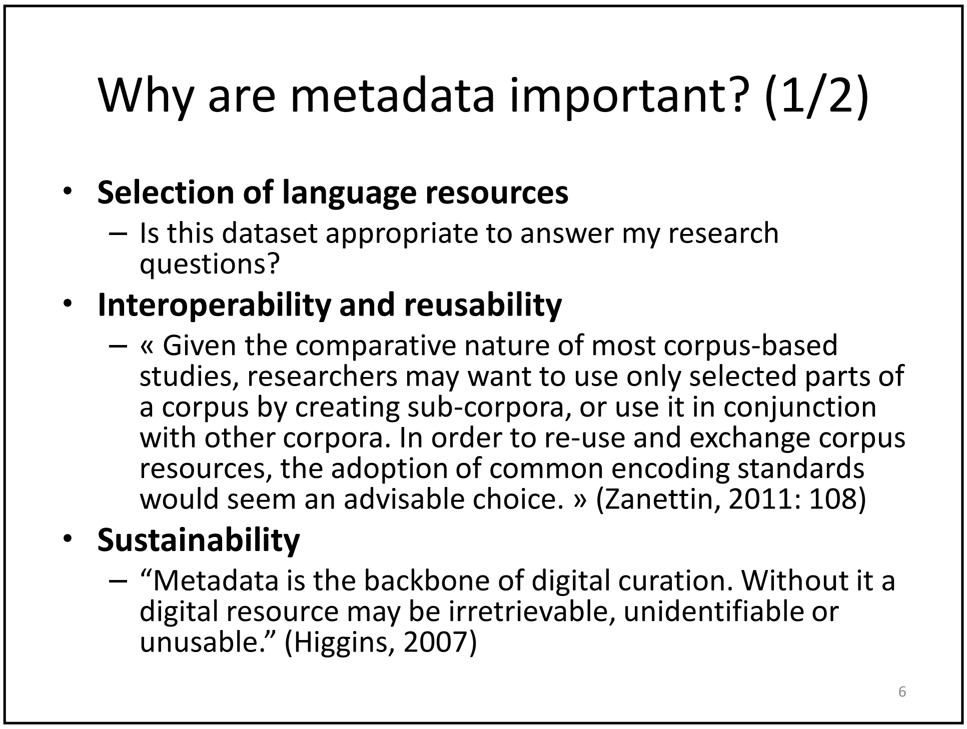
Metadata

- "Data that serve to describe other data" (Heid 2009)
- "The kind of data that is needed to describe a digital resource in sufficient detail and with sufficient accuracy for some agent to determine whether or not that digital resource is of relevance to a particular enquiry" (Burnard 2005)
- "Metadata is descriptive or contextual information which refers to or is associated with another object or resource. This usually takes the form of a structured set of elements which describe the information resource and assists in the identification, location and retrieval of it by users, while facilitating content and access management" (Higgins 2007)
- "Metadata is data about data: information describing properties of linguistic resources, for instance the size of a corpus, the recording date of a speech file, the purpose for which annotations were created." (Frequently Asked Questions - Metadata in CLARIN: basics, https://www.clarin.eu/faq-page/273#t273n2850)



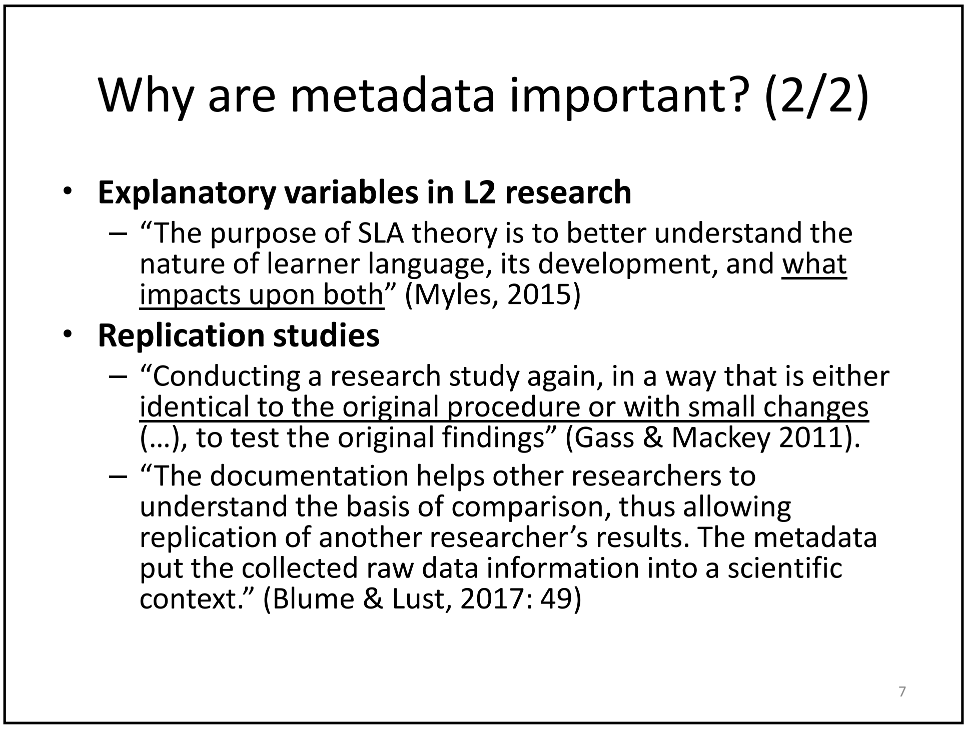
Scope of metadata (Burnard, 2004)

- **Editorial metadata**
  - Information about the relationship between corpus components and their original source (e.g. addition or omission, correction, normalization)
- **Analytic metadata**
  - Information about the way in which corpus components have been interpreted and analysed (e.g. transcription, linguistic annotation)
- **Descriptive metadata**
  - Classificatory information derived from internal or external properties of the corpus components (e.g. demographic characteristics of speakers, setting, text type)
- **Administrative metadata**
  - Corpus availability, revision status, etc.



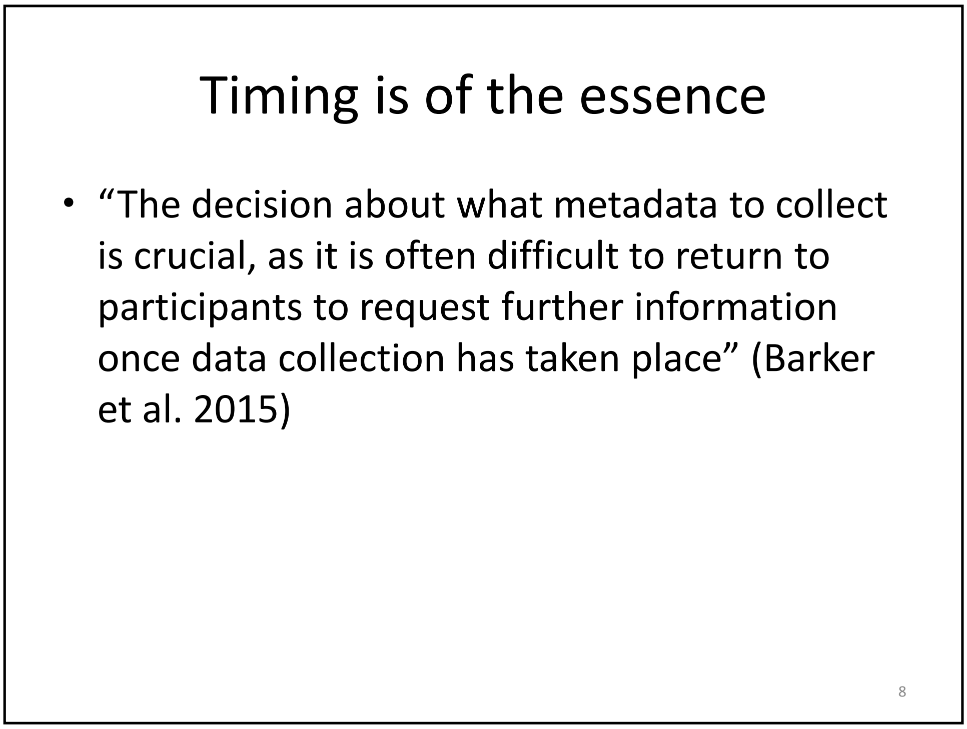
Why are metadata important? (1/2)

- **Selection of language resources**
  - Is this dataset appropriate to answer my research questions?
- **Interoperability and reusability**
  - « Given the comparative nature of most corpus-based studies, researchers may want to use only selected parts of a corpus by creating sub-corpora, or use it in conjunction with other corpora. In order to re-use and exchange corpus resources, the adoption of common encoding standards would seem an advisable choice. » (Zanettin, 2011: 108)
- **Sustainability**
  - “Metadata is the backbone of digital curation. Without it a digital resource may be irretrievable, unidentifiable or unusable.” (Higgins, 2007)



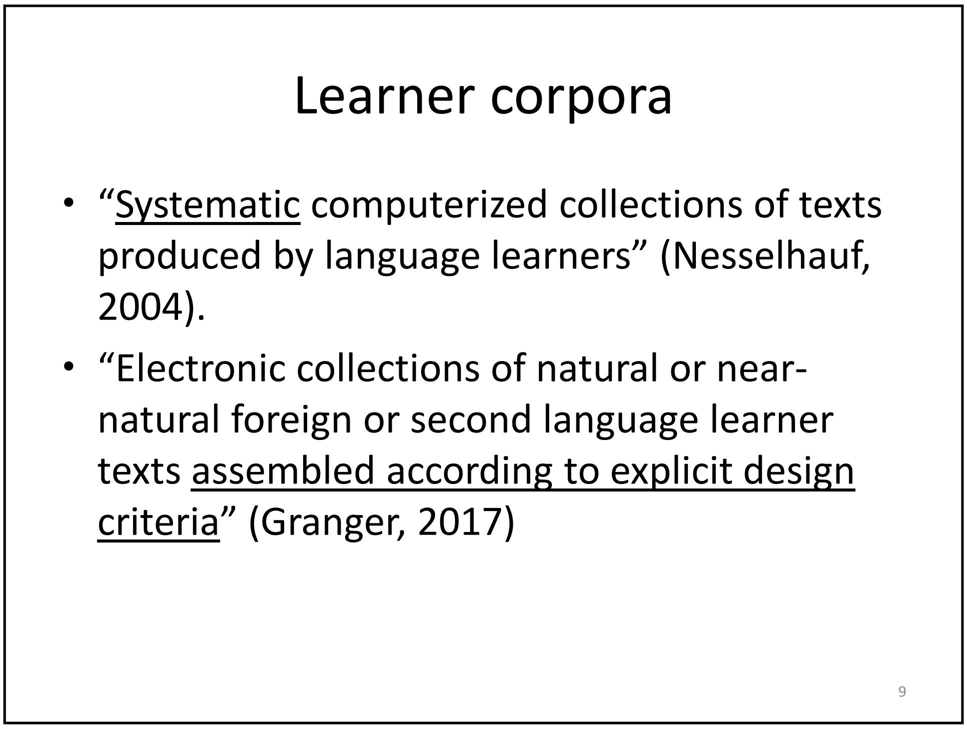
Why are metadata important? (2/2)

- **Explanatory variables in L2 research**
  - “The purpose of SLA theory is to better understand the nature of learner language, its development, and what impacts upon both” (Myles, 2015)
- **Replication studies**
  - “Conducting a research study again, in a way that is either identical to the original procedure or with small changes (…), to test the original findings” (Gass & Mackey 2011).
  - “The documentation helps other researchers to understand the basis of comparison, thus allowing replication of another researcher’s results. The metadata put the collected raw data information into a scientific context.” (Blume & Lust, 2017: 49)



Timing is of the essence

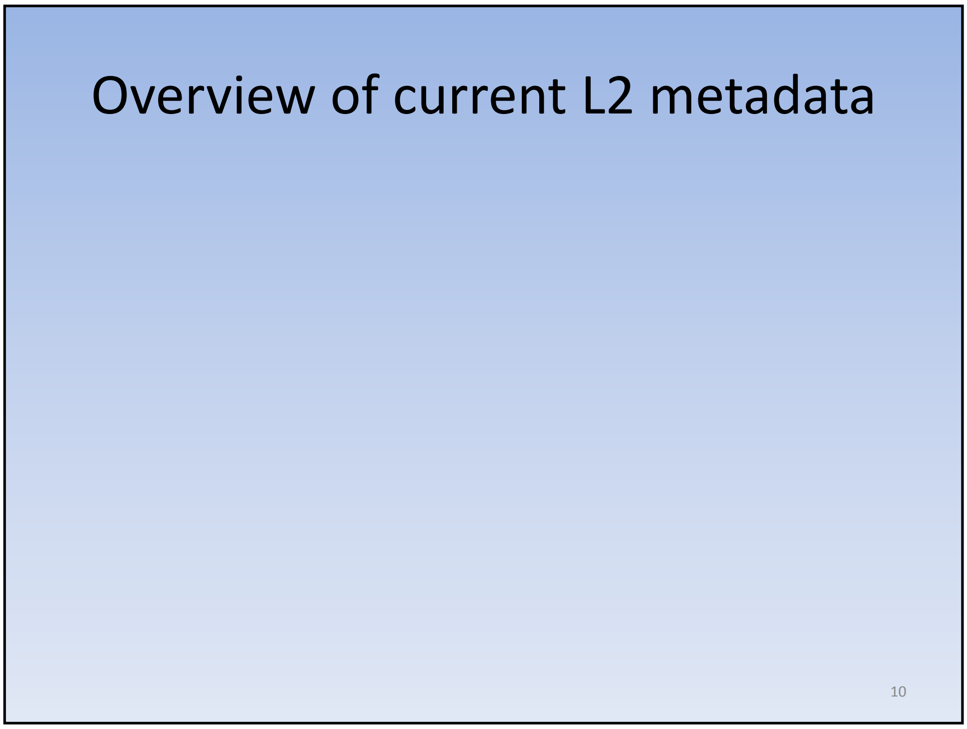
- “The decision about what metadata to collect is crucial, as it is often difficult to return to participants to request further information once data collection has taken place” (Barker et al. 2015)



Learner corpora

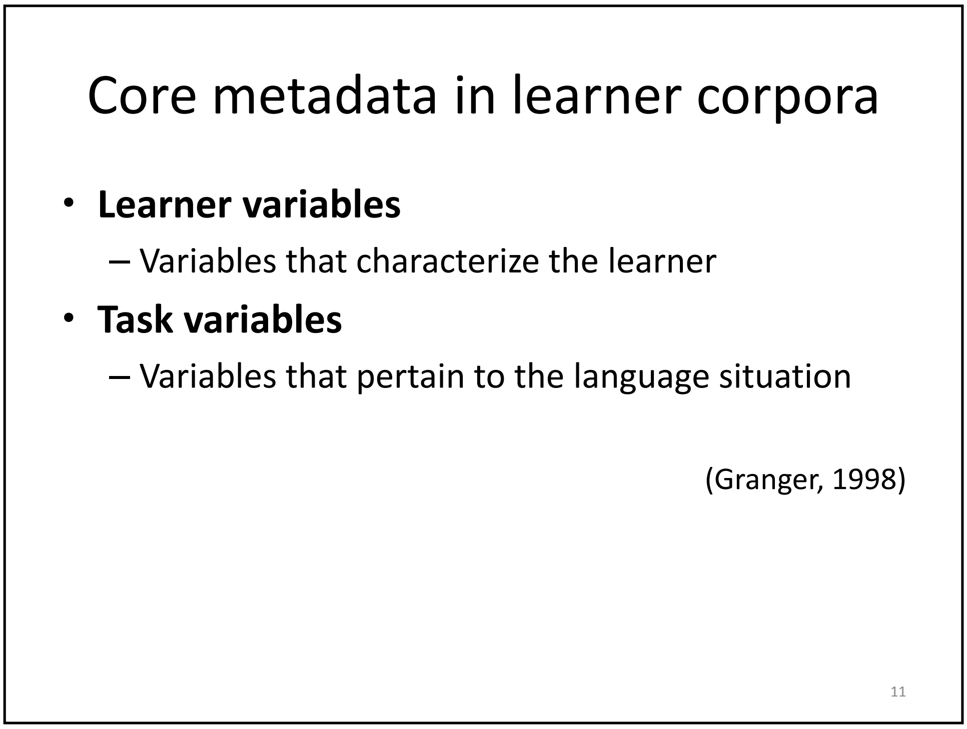
- “Systematic computerized collections of texts produced by language learners” (Nesselhauf, 2004).
- “Electronic collections of natural or near-natural foreign or second language learner texts assembled according to explicit design criteria” (Granger, 2017)

9



Overview of current L2 metadata

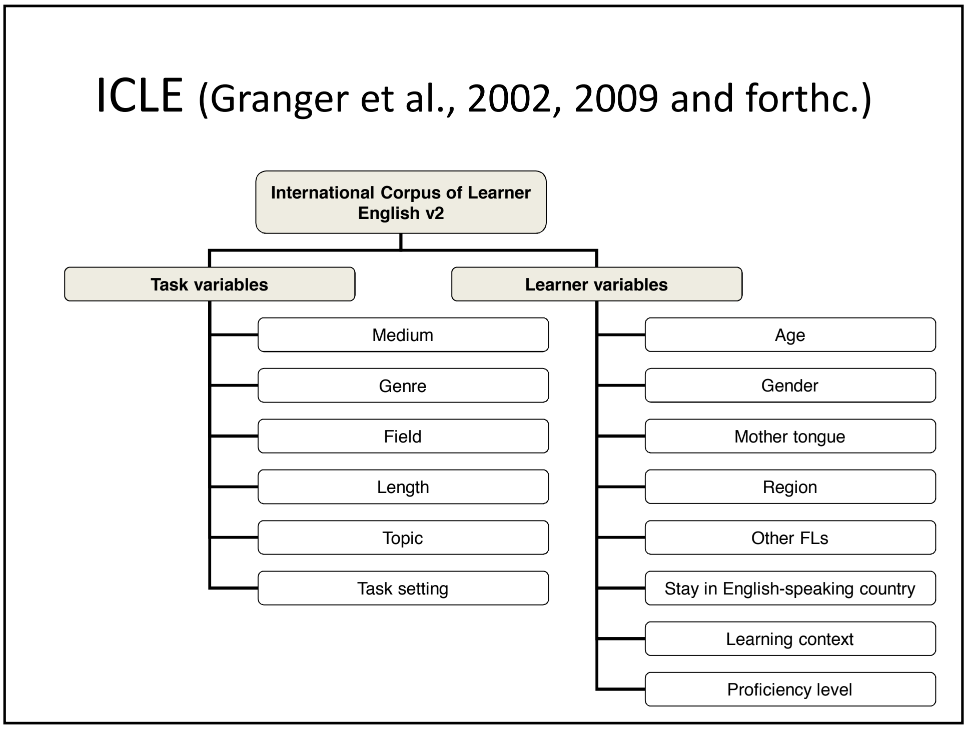
10



Core metadata in learner corpora

- **Learner variables**
  - Variables that characterize the learner
- **Task variables**
  - Variables that pertain to the language situation

(Granger, 1998)



ICLE (Granger et al., 2002, 2009 and forthc.)

International Corpus of Learner English v2

Task variables
- Medium
- Genre
- Field
- Length
- Topic
- Task setting

Learner variables
- Age
- Gender
- Mother tongue
- Region
- Other FLs
- Stay in English-speaking country
- Learning context
- Proficiency level

Metadata in CLC and ICLE (1/2) (Barker 2015)

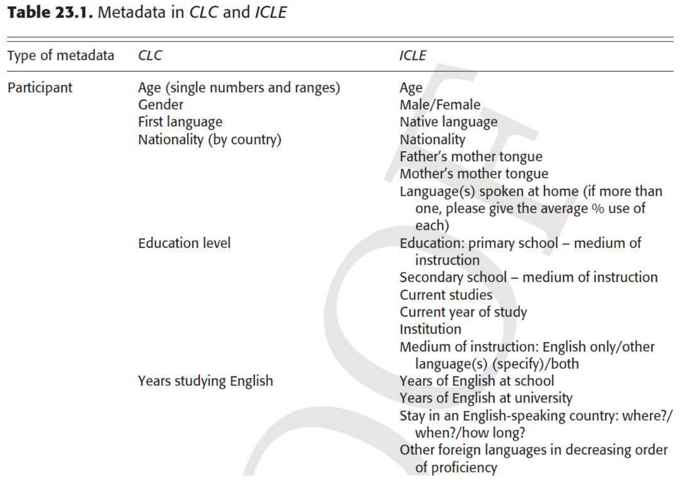
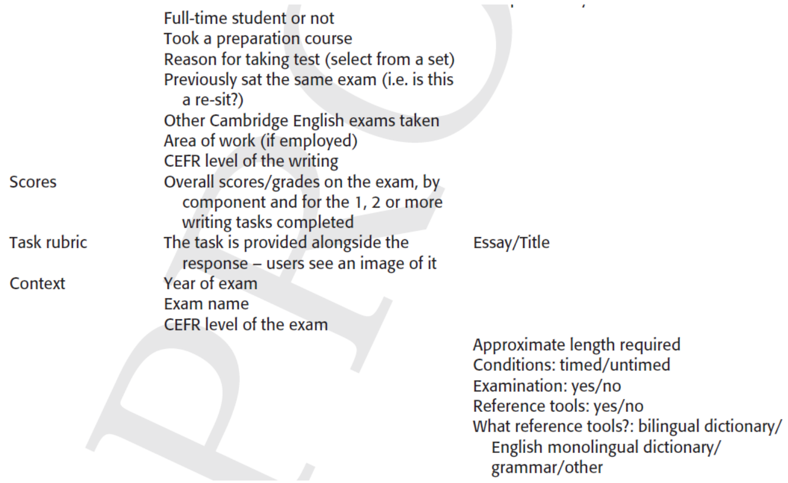


Table 23.1. Metadata in *CLC* and *ICLE*

Type of metadata	*CLC*	*ICLE*
Participant	Age (single numbers and ranges) Gender First language Nationality (by country)	Age Male/Female Native language Nationality Father's mother tongue Mother's mother tongue Language(s) spoken at home (if more than one, please give the average % use of each)
	Education level	Education: primary school – medium of instruction Secondary school – medium of instruction Current studies Current year of study Institution Medium of instruction: English only/other language(s) (specify)/both
	Years studying English	Years of English at school Years of English at university Stay in an English-speaking country: where?/when?/how long? Other foreign languages in decreasing order of proficiency

Metadata in CLC and ICLE (2/2) (Barker 2015)



	Full-time student or not Took a preparation course Reason for taking test (select from a set) Previously sat the same exam (i.e. is this a re-sit?) Other Cambridge English exams taken Area of work (if employed) CEFR level of the writing	
Scores	Overall scores/grades on the exam, by component and for the 1, 2 or more writing tasks completed	
Task rubric	The task is provided alongside the response – users see an image of it	Essay/Title
Context	Year of exam Exam name CEFR level of the exam	Approximate length required Conditions: timed/untimed Examination: yes/no Reference tools: yes/no What reference tools?: bilingual dictionary/English monolingual dictionary/grammar/other

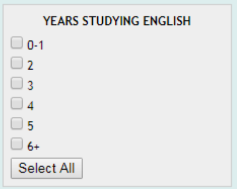
Variables: different labels

- CLC - ICLE
  - First language vs. native language
  - Nationality vs. country
  - Years studying English vs. Years of English at school + Years of English at university

15

Variables: different categories

- Open CLC



YEARS STUDYING ENGLISH

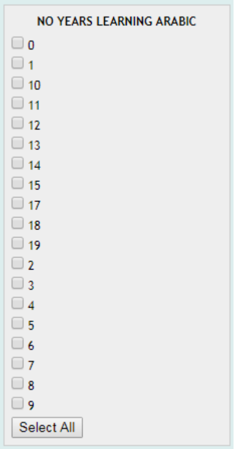
□ 0-1
□ 2
□ 3
□ 4
□ 5
□ 6+

Select All

- Guangwai - Lancaster Chinese Learner Corpus

  No such variable

- Arabic Learner Corpus



NO YEARS LEARNING ARABIC

□ 0
□ 1
□ 10
□ 11
□ 12
□ 13
□ 14
□ 15
□ 17
□ 18
□ 19
□ 2
□ 3
□ 4
□ 5
□ 6
□ 7
□ 8
□ 9

Select All

16

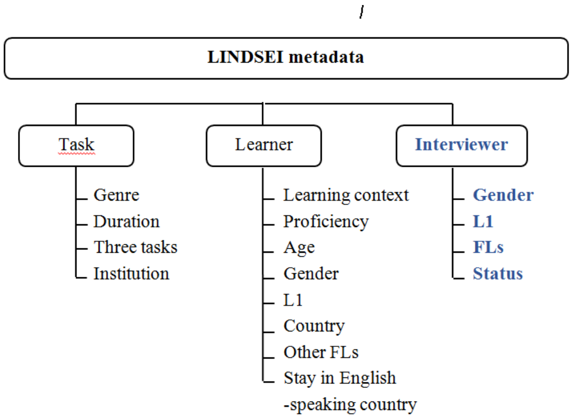
Proficiency

- **Four major ways** of assessing proficiency (Thomas 1994):
  1) impressionistic judgement
  2) use of institutional status as a proxy for proficiency level
  3) use of research-internal or in-house measures of proficiency
  4) standardized test scores
- In current learner corpora: considerable **variation**
  - No information at all
  - Institutional status
  - (more rarely) research-internal or standardized test scores
- Even when scores are available, it is often **unclear**
  - whether it is the learner or the text that has been assessed
  - whether the score involves all skills or only some subskills
- **Mapping** scores from different tests is problematic (Tono 2013)

Heterogeneity of learner corpus types

- Many learner corpora require **additional sets** of metadata
- **Oral data**
  - Editorial & analytical metadata: normalization & transcription guidelines
  - [For interview data]: Descriptive metadata for the interviewer

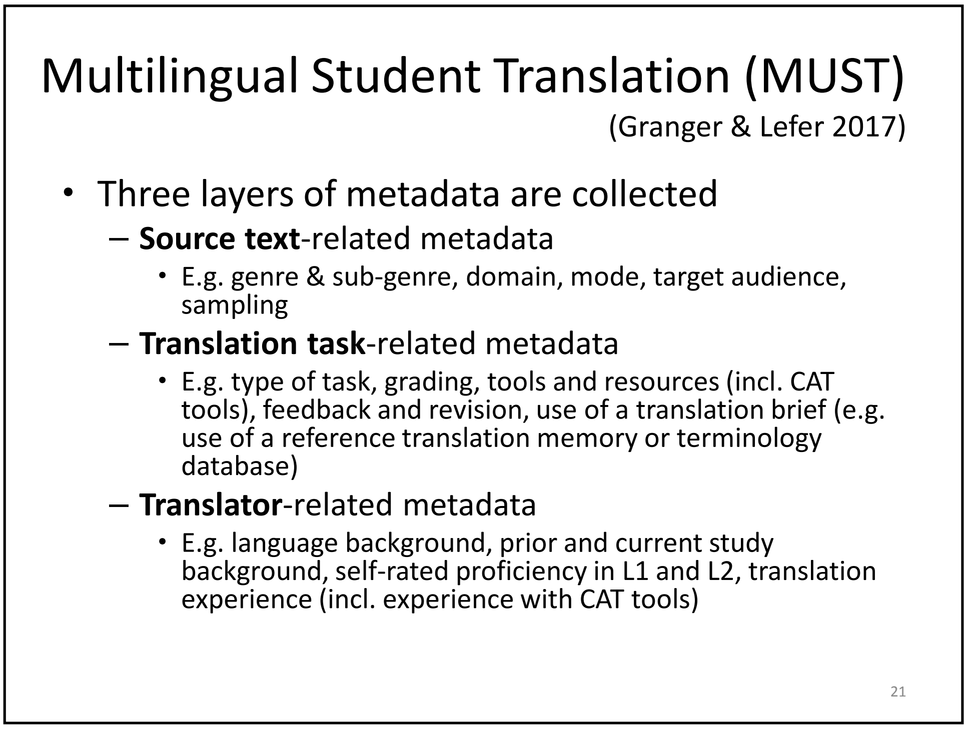
LINDSEI (Gilquin et al., 2010)



Peripheral learner corpora

- Often need more info about the use of external resources

- More controlled tasks
  – Picture used for picture description
- Learner translation corpora
  – Source text

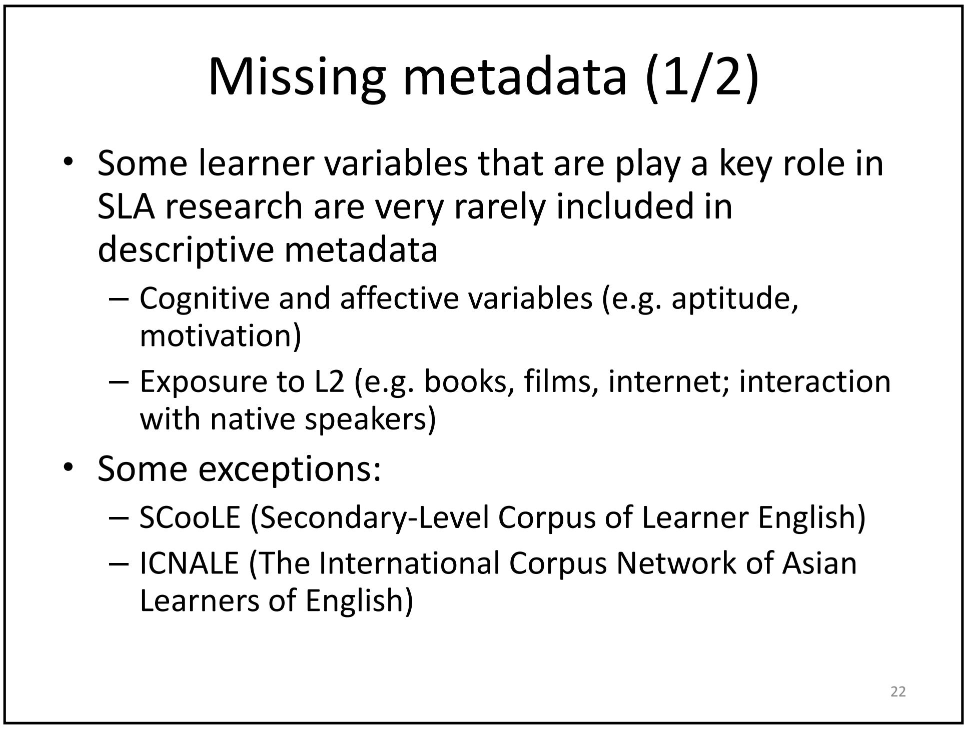


Multilingual Student Translation (MUST)

(Granger & Lefer 2017)

- Three layers of metadata are collected
  - **Source text**-related metadata
    - E.g. genre & sub-genre, domain, mode, target audience, sampling
  - **Translation task**-related metadata
    - E.g. type of task, grading, tools and resources (incl. CAT tools), feedback and revision, use of a translation brief (e.g. use of a reference translation memory or terminology database)
  - **Translator**-related metadata
    - E.g. language background, prior and current study background, self-rated proficiency in L1 and L2, translation experience (incl. experience with CAT tools)

21



Missing metadata (1/2)

- Some learner variables that are play a key role in SLA research are very rarely included in descriptive metadata
  - Cognitive and affective variables (e.g. aptitude, motivation)
  - Exposure to L2 (e.g. books, films, internet; interaction with native speakers)
- Some exceptions:
  - SCooLE (Secondary-Level Corpus of Learner English)
  - ICNALE (The International Corpus Network of Asian Learners of English)

22

SCooLE (Möller 2017)

- **Exposure**
  - Frequency with which English is spoken outside school
  - Frequency with which English books and magazines are read outside school
  - Frequency with which English films and TV programmes are watched outside school
  - Frequency with which English websites are used outside school
- **Cognitive and affective variables**
  - Intelligence (several subscores: verbal, reasoning, concentration, etc.)
  - Motivation (several subscores: perseverance and effort, orientation towards performance and success, etc.)

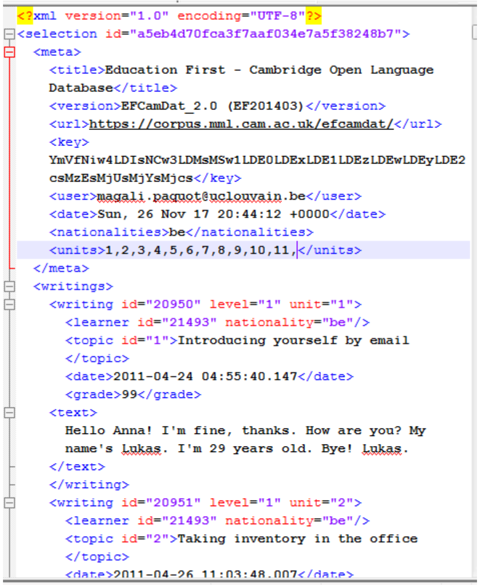
Missing metadata (2/2)

- **General factual information** on the corpus is often absent
  - Editorial metadata
  - Analytic metadata
  - Administrative metadata

Where and in what form?

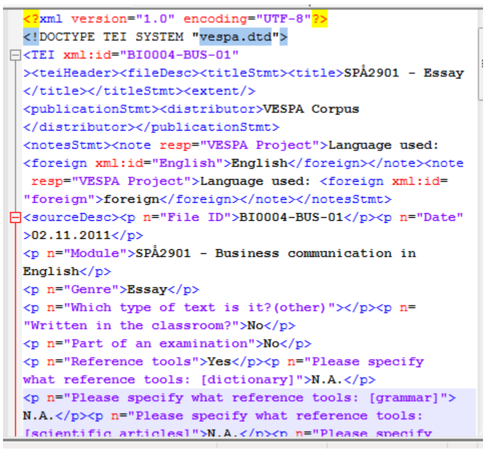
- **Heterogeneity of practices**
  - Descriptive metadata
    - Stand-alone metadata file: ICLE, LINDSEI, ICNALE, ETS Corpus of Non-Native Written English
    - File header: EFCAMDAT, MERLIN, ASK
    - Both representations: VESPA
  - Editorial / administrative / analytical metadata
    - File header: rarely
    - Readme file / corpus manual: ICLE, ICNALE, LINDSEI, VESPA

EFCAMDAT file header



```
<?xml version="1.0" encoding="UTF-8"?>
<selection id="a5eb4d70fca3f7aaf034e7a5f38248b7">
  <meta>
    <title>Education First - Cambridge Open Language
    Database</title>
    <version>EFCamDat_2.0 (EF201403)</version>
    <url>https://corpus.mml.cam.ac.uk/efcamdat/</url>
    <key>
    YmVfNiw4LDIsNCw3LDMsMSw1LDE0LDExLDE1LDEzLDEwLDEyLDE2
    csMzEsMjUsMjYsMjcs</key>
    <user>magali.paquot@uclouvain.be</user>
    <date>Sun, 26 Nov 17 20:44:12 +0000</date>
    <nationalities>be</nationalities>
    <units>1,2,3,4,5,6,7,8,9,10,11,</units>
  </meta>
  <writings>
    <writing id="20950" level="1" unit="1">
      <learner id="21493" nationality="be"/>
      <topic id="1">Introducing yourself by email
      </topic>
      <date>2011-04-24 04:55:40.147</date>
      <grade>99</grade>
    <text>
      Hello Anna! I'm fine, thanks. How are you? My
      name's Lukas. I'm 29 years old. Bye! Lukas.
    </text>
    </writing>
    <writing id="20951" level="1" unit="2">
      <learner id="21493" nationality="be"/>
      <topic id="2">Taking inventory in the office
      </topic>
      <date>2011-04-26 11:03:48.007</date>
```

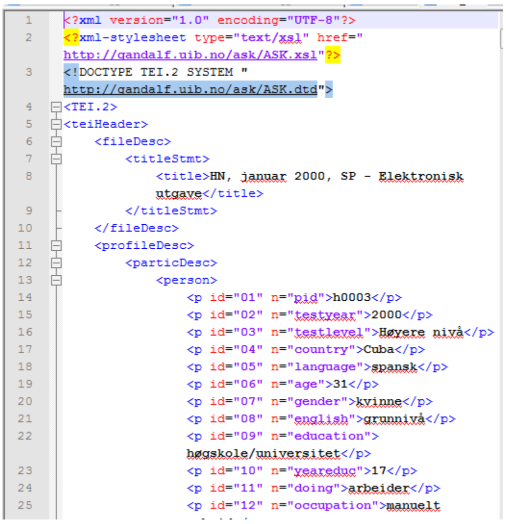
VESPA file header



```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE TEI SYSTEM "vespa.dtd">
<TEI xml:id="BI0004-BUS-01"
><teiHeader><fileDesc><titleStmt><title>SPÅ2901 - Essay
</title></titleStmt><extent/>
<publicationStmt><distributor>VESPA Corpus
</distributor></publicationStmt>
<notesStmt><note resp="VESPA Project">Language used:
<foreign xml:id="English">English</foreign></note><note
 resp="VESPA Project">Language used: <foreign xml:id=
"foreign">foreign</foreign></note></notesStmt>
<sourceDesc><p n="File ID">BI0004-BUS-01</p><p n="Date"
>02.11.2011</p>
<p n="Module">SPÅ2901 - Business communication in
English</p>
<p n="Genre">Essay</p>
<p n="Which type of text is it?(other)"></p><p n=
"Written in the classroom?">No</p>
<p n="Part of an examination">No</p>
<p n="Reference tools">Yes</p><p n="Please specify
what reference tools: [dictionary]">N.A.</p>
<p n="Please specify what reference tools: [grammar]">
N.A.</p><p n="Please specify what reference tools:
[scientific articles]">N.A.</p><p n="Please specify
```

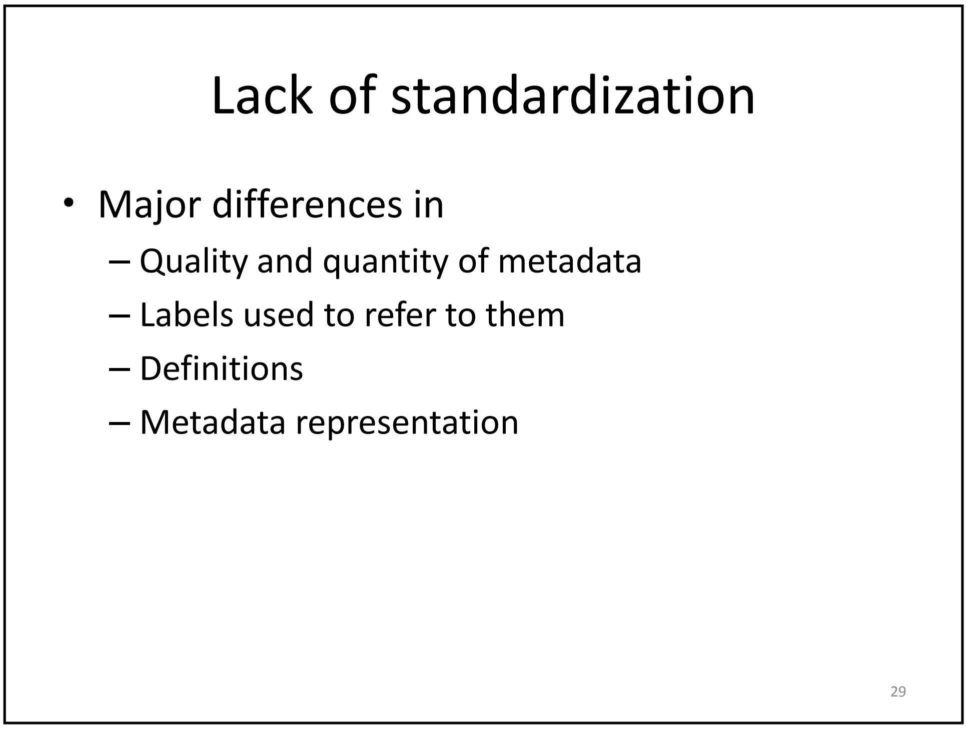
27

ASK corpus



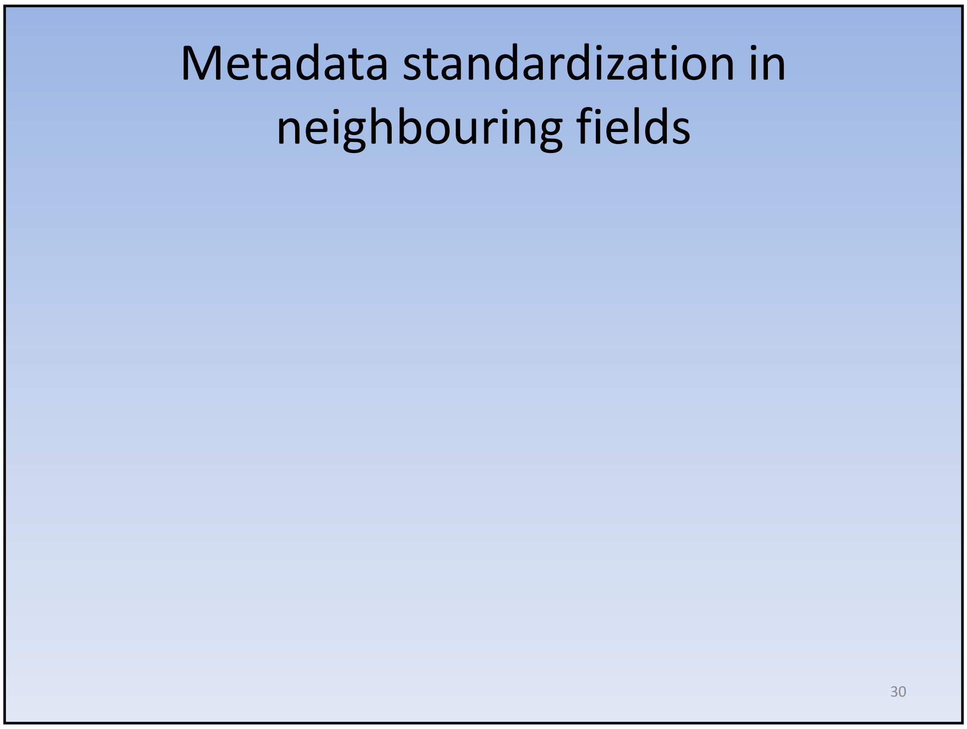
```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="
http://gandalf.uib.no/ask/ASK.xsl"?>
<!DOCTYPE TEI.2 SYSTEM "
http://gandalf.uib.no/ask/ASK.dtd">
<TEI.2>
<teiHeader>
    <fileDesc>
        <titleStmt>
            <title>HN, januar 2000, SP - Elektronisk
            utgave</title>
        </titleStmt>
    </fileDesc>
    <profileDesc>
        <particDesc>
            <person>
                <p id="01" n="pid">h0003</p>
                <p id="02" n="testyear">2000</p>
                <p id="03" n="testlevel">Høyere nivå</p>
                <p id="04" n="country">Cuba</p>
                <p id="05" n="language">spansk</p>
                <p id="06" n="age">31</p>
                <p id="07" n="gender">kvinne</p>
                <p id="08" n="english">grunnivå</p>
                <p id="09" n="education">
                høgskole/universitet</p>
                <p id="10" n="yeareduc">17</p>
                <p id="11" n="doing">arbeider</p>
                <p id="12" n="occupation">manuelt
```

28

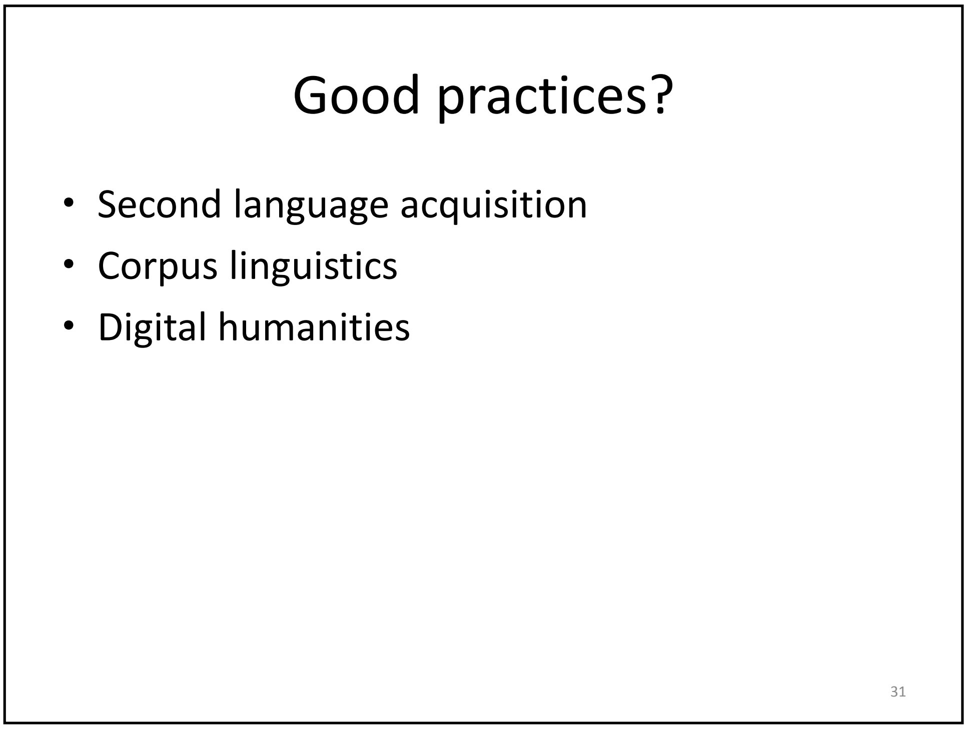


Lack of standardization

- Major differences in
  - Quality and quantity of metadata
  - Labels used to refer to them
  - Definitions
  - Metadata representation

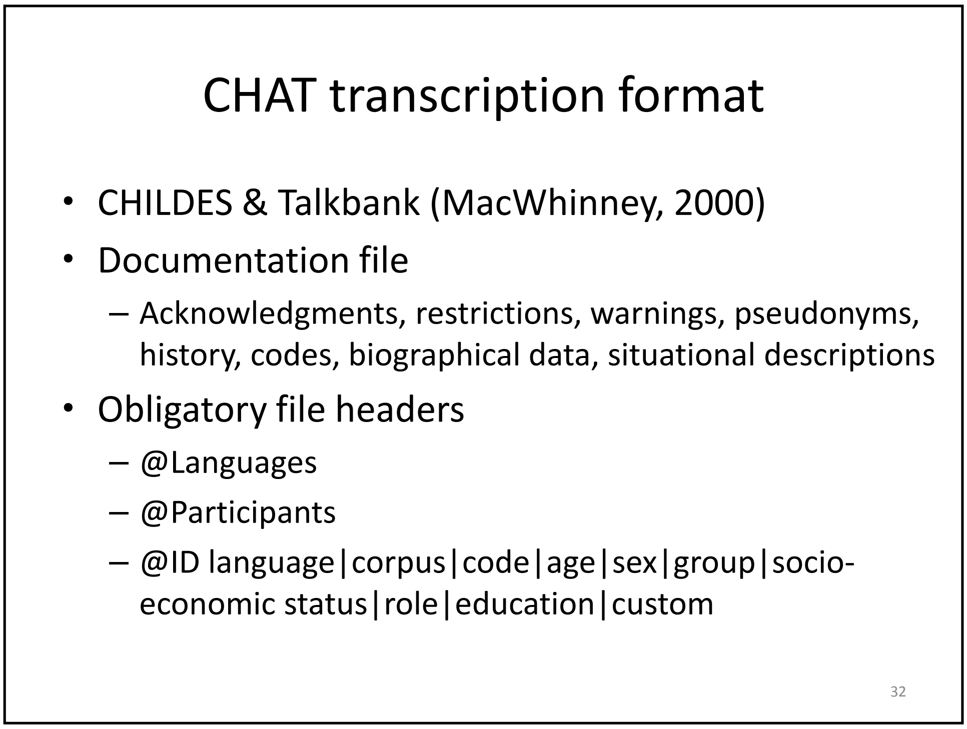


Metadata standardization in neighbouring fields



Good practices?

- Second language acquisition
- Corpus linguistics
- Digital humanities



CHAT transcription format

- CHILDES & Talkbank (MacWhinney, 2000)
- Documentation file
  - Acknowledgments, restrictions, warnings, pseudonyms, history, codes, biographical data, situational descriptions
- Obligatory file headers
  - @Languages
  - @Participants
  - @ID language|corpus|code|age|sex|group|socio-economic status|role|education|custom

CHAT: optional file headers

- @Interaction type
- @Location
- @Number
- @Recording quality
- @Room layout
- @Tape location
- @Time duration
- @Transcriber
- @Transcription
- …

Indexing and registration of materials for language resource archives

- The CHILDES and TalkBank systems provide information that can be incorporated into:
  - OLAC (Online Language Archives Community)
  - CLARIN VLO (Virtual Language Observatory)
    - Component MetaData Infrastructure (CMDI)

IRIS (Mardsen et al., 2016)

- A digital repository of instruments and materials for research into second languages
  - **Ontologies** for categorizing instruments and materials
    - Type of instrument
    - Data type
    - Participant type
      - Adolescent, adult, bilinguals, first language attriters, heritage learners, teacher trainees, young learners
    - Proficiency of learners
    - Domains of use
      - Academic, home, residence abroad, school, work
  - https://www.iris-database.org/iris/app/home/search-help#3

A maze of standards for metadata description …

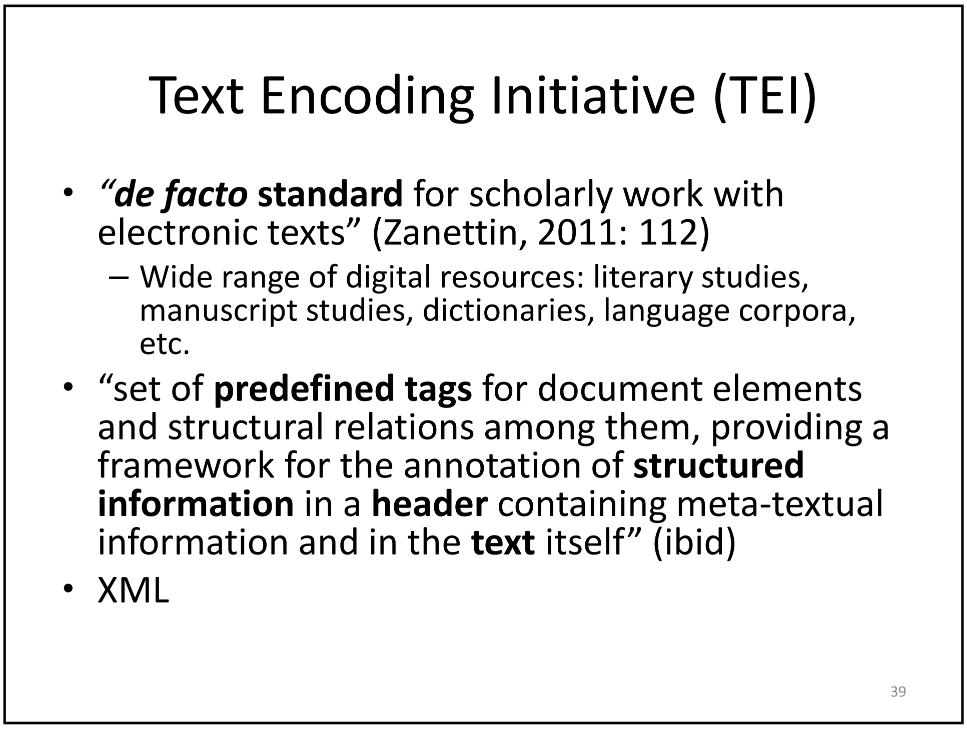
- Text Encoding Initiative (TEI)
  http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CC.html
- Corpus Encoding Standard (CES; XCES)
  https://www.cs.vassar.edu/CES/
- Simple Dublin Core
  http://dublincore.org/documents/dces/
- ISLE Metadata Initiative
  http://www.mpi.nl/ISLE/
- Component Metadata Infrastructure
  https://www.clarin.eu/content/component-metadata

TEI/TEI-Lite in corpus projects

- British Academic Spoken English (BASE)
- British National Corpus
- English-Norwegian Parallel Corpus
- English-Swedish Parallel Corpus
- Michigan Corpus of Academic Spoken English (MICASE)
- Oslo Multilingual Corpus
- PAROLE corpora (CES)
- Polish National Corpus
- Prague Spoken Corpus
- Russian Reference Corpus
- SCIENTEXT

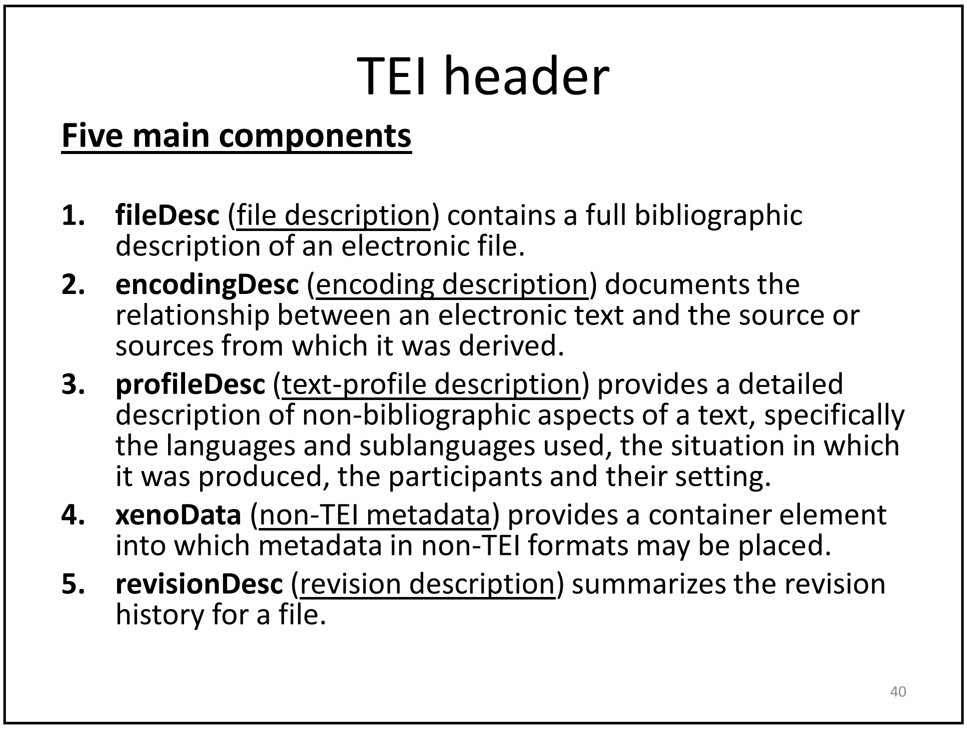
TEI-conformant learner corpora

- ACAW - Aachen Corpus of Academic Writing (ACAW): http://www.anglistik.rwth-aachen.de/cms/Anglistik/Anglistik-Amerikanistik/Anglistische-Sprachwissenschaft/~gdgf/Kerz-Mit/?allou=1
- ASK corpus http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.380.627&rep=rep1&type=pdf
- CityU Corpus of Essay Drafts of English Language Learners: https://corpling.uis.georgetown.edu/amir/pdf/annis_cityu_prepub.pdf
- COPLE2 Corpus http://www.lrec-conf.org/proceedings/lrec2016/pdf/439_Paper.pdf
- NOCE corpus http://www.sfs.uni-tuebingen.de/~dm/handouts/evo-paris-09-05-28.pdf
- Hanken Corpus of Academic Written English for Economics (Mäkinen & Hiltunen, 2016)
- Polish Learner English Corpus (TEI Lite): http://www.lancaster.ac.uk/fass/projects/corpus/cbls/corpora.asp
- VESPA https://uclouvain.be/en/research-institutes/ilc/cecl/vespa.html



Text Encoding Initiative (TEI)

- *“de facto* **standard** for scholarly work with electronic texts” (Zanettin, 2011: 112)
  - Wide range of digital resources: literary studies, manuscript studies, dictionaries, language corpora, etc.
- “set of **predefined tags** for document elements and structural relations among them, providing a framework for the annotation of **structured information** in a **header** containing meta-textual information and in the **text** itself” (ibid)
- XML

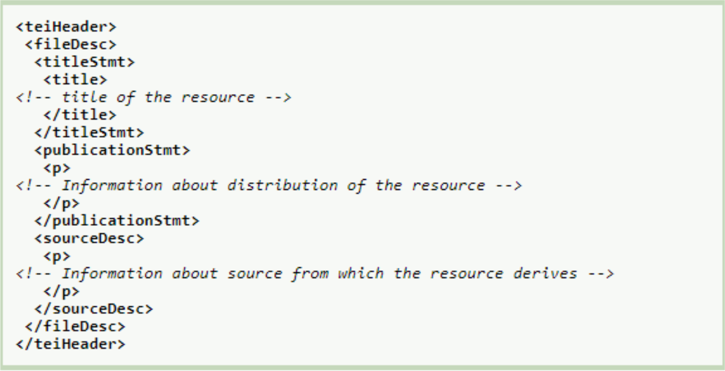


TEI header

Five main components

1. **fileDesc** (file description) contains a full bibliographic description of an electronic file.
2. **encodingDesc** (encoding description) documents the relationship between an electronic text and the source or sources from which it was derived.
3. **profileDesc** (text-profile description) provides a detailed description of non-bibliographic aspects of a text, specifically the languages and sublanguages used, the situation in which it was produced, the participants and their setting.
4. **xenoData** (non-TEI metadata) provides a container element into which metadata in non-TEI formats may be placed.
5. **revisionDesc** (revision description) summarizes the revision history for a file.

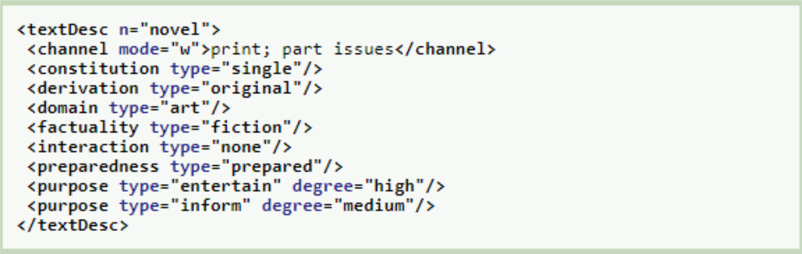
TEI header



```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>
<!-- title of the resource -->
   </title>
  </titleStmt>
  <publicationStmt>
   <p>
<!-- Information about distribution of the resource -->
   </p>
  </publicationStmt>
  <sourceDesc>
   <p>
<!-- Information about source from which the resource derives -->
   </p>
  </sourceDesc>
 </fileDesc>
</teiHeader>
```

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html

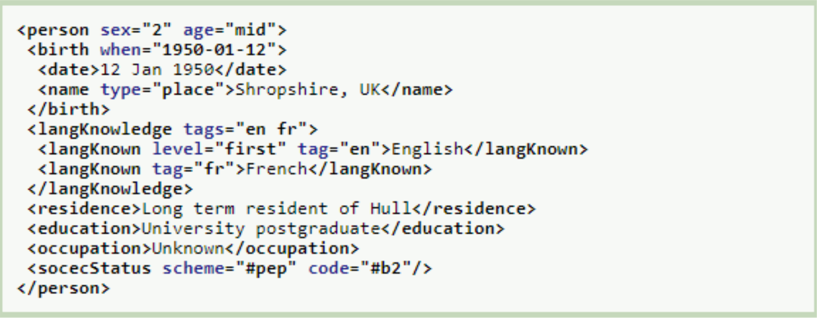
Corpus module: text description



```
<textDesc n="novel">
 <channel mode="w">print; part issues</channel>
 <constitution type="single"/>
 <derivation type="original"/>
 <domain type="art"/>
 <factuality type="fiction"/>
 <interaction type="none"/>
 <preparedness type="prepared"/>
 <purpose type="entertain" degree="high"/>
 <purpose type="inform" degree="medium"/>
</textDesc>
```

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CC.html

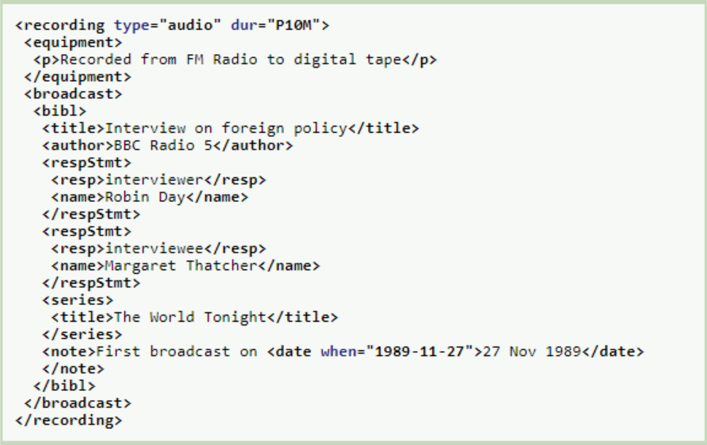
Corpus module: Participant description



```
<person sex="2" age="mid">
 <birth when="1950-01-12">
  <date>12 Jan 1950</date>
  <name type="place">Shropshire, UK</name>
 </birth>
 <langKnowledge tags="en fr">
  <langKnown level="first" tag="en">English</langKnown>
  <langKnown tag="fr">French</langKnown>
 </langKnowledge>
 <residence>Long term resident of Hull</residence>
 <education>University postgraduate</education>
 <occupation>Unknown</occupation>
 <socecStatus scheme="#pep" code="#b2"/>
</person>
```

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CC.html

Transcriptions of speech



```
<recording type="audio" dur="P10M">
 <equipment>
  <p>Recorded from FM Radio to digital tape</p>
 </equipment>
 <broadcast>
  <bibl>
   <title>Interview on foreign policy</title>
   <author>BBC Radio 5</author>
   <respStmt>
    <resp>interviewer</resp>
    <name>Robin Day</name>
   </respStmt>
   <respStmt>
    <resp>interviewee</resp>
    <name>Margaret Thatcher</name>
   </respStmt>
   <series>
    <title>The World Tonight</title>
   </series>
   <note>First broadcast on <date when="1989-11-27">27 Nov 1989</date>
   </note>
  </bibl>
 </broadcast>
</recording>
```

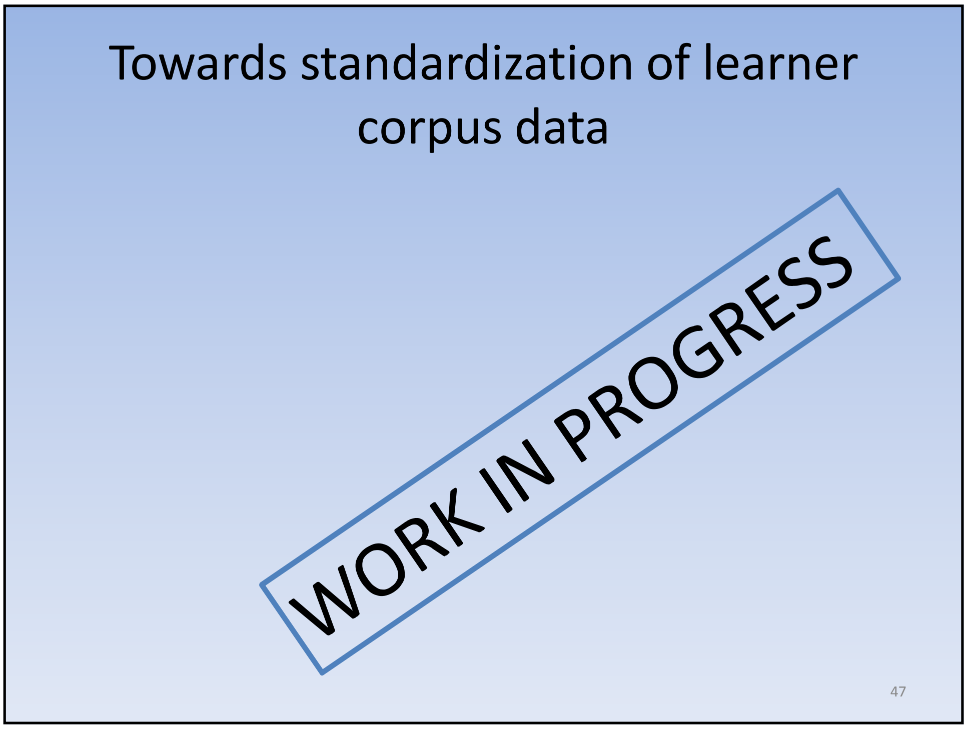
http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html#HD32

Is TEI prescriptive?

- "There has long been a perception that the TEI is a prescriptive model, as indeed in some respects it is: it prescribes a number of very specific constraints for documents claiming to be TEI conformant, for example. However, the prescriptive part of the TEI is concerned only with how the TEI definitions are to be deployed; very few prescriptions are provided as to which of the many hundreds of TEI-defined concepts should be selected in a given context" (Burnard 2017)

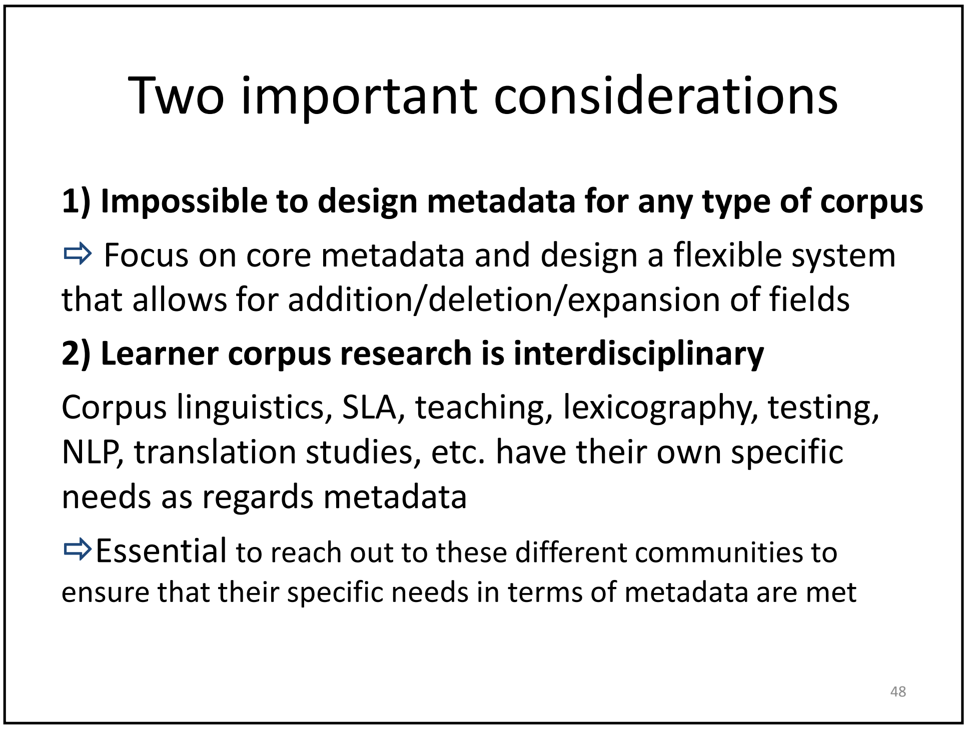
CMC corpora network (Beißwenger et al., 2017)

- Computer Mediated Communication and social media interactions
- TEI special interest group on CMC
  - Extend the TEI framework with additions dedicated to the representation of the structural and linguistic peculiarities of CMC genres
  - Schemas developed following the rules for customization described in the TEI guidelines
  - To be presented to the TEI Technical Council in the form of feature requests, i.e. suggestions for the extension of the 'official' TEI standard.



Towards standardization of learner corpus data

WORK IN PROGRESS



Two important considerations

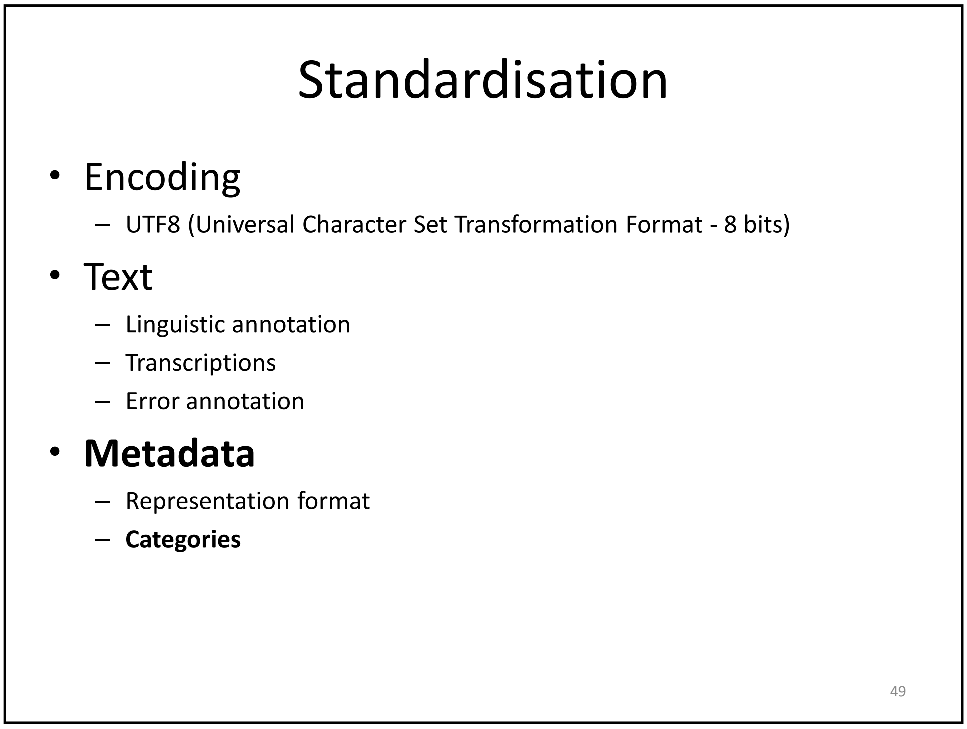
1) Impossible to design metadata for any type of corpus

⇨ Focus on core metadata and design a flexible system that allows for addition/deletion/expansion of fields

2) Learner corpus research is interdisciplinary

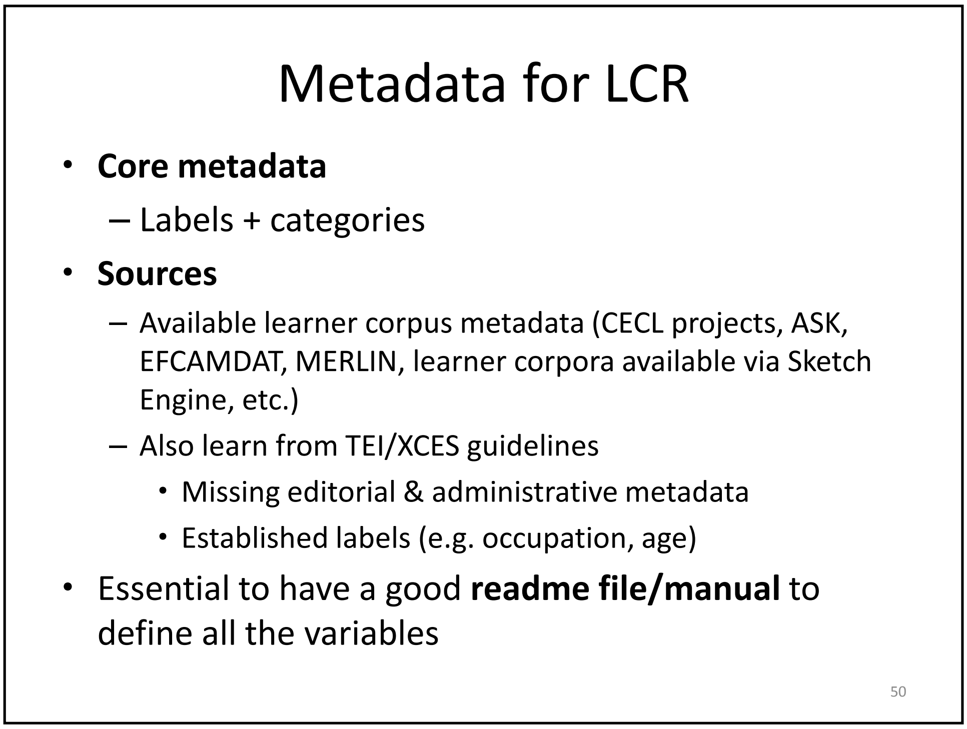
Corpus linguistics, SLA, teaching, lexicography, testing, NLP, translation studies, etc. have their own specific needs as regards metadata

⇨Essential to reach out to these different communities to ensure that their specific needs in terms of metadata are met



Standardisation

- Encoding
  - UTF8 (Universal Character Set Transformation Format - 8 bits)
- Text
  - Linguistic annotation
  - Transcriptions
  - Error annotation
- **Metadata**
  - Representation format
  - **Categories**



Metadata for LCR

- **Core metadata**
  - Labels + categories
- **Sources**
  - Available learner corpus metadata (CECL projects, ASK, EFCAMDAT, MERLIN, learner corpora available via Sketch Engine, etc.)
  - Also learn from TEI/XCES guidelines
    - Missing editorial & administrative metadata
    - Established labels (e.g. occupation, age)
- Essential to have a good **readme file/manual** to define all the variables

Core metadata for learner corpora

- **Our proposal: five main components**

1) Administrative metadata
2) Corpus design metadata
3) Annotation metadata
4) Text metadata
5) Learner metadata

→ see Draft Proposal on handout

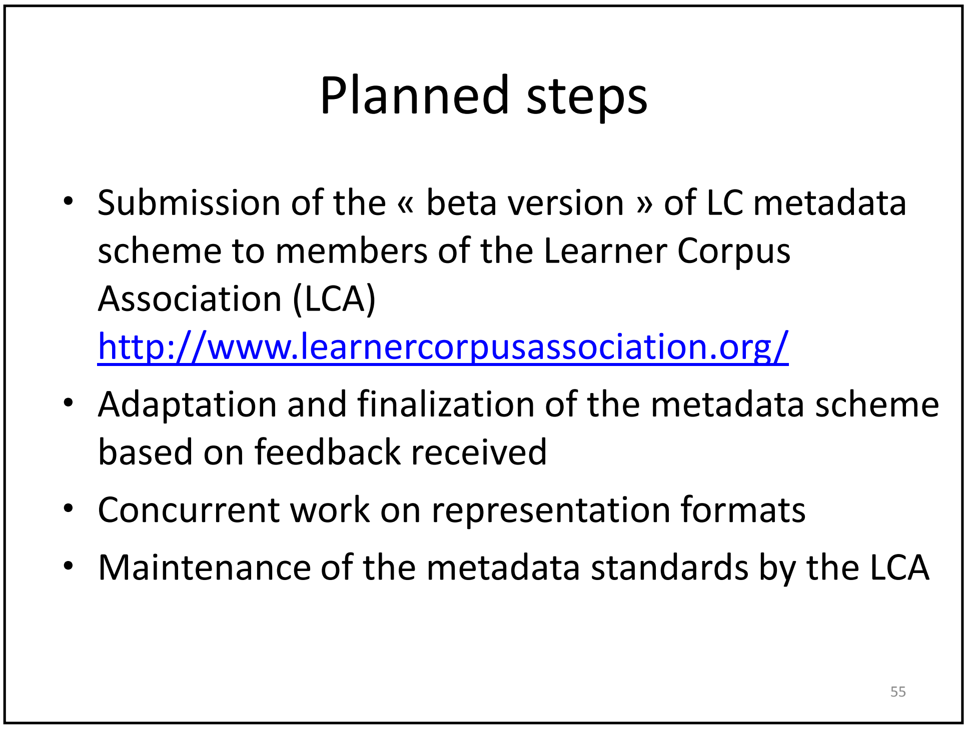
Preliminary validation of core metadata

- Confront draft proposal with metadata from existing learner corpora
  - CECL learner corpora
    - ICLE
    - LINDSEI
    - LONGDALE
    - VESPA
  - Other learner corpora

Additional modules

- Integration of **specialized metadata**
  - Multimodal learner corpora (Freigang & Bergmann, 2013)
    - Info on modality (gestures, eye-gaze, facial expressions)
  - Translation learner corpora
    - Info on source text to be translated, etc.
  - CLIL learner corpora
  - ………

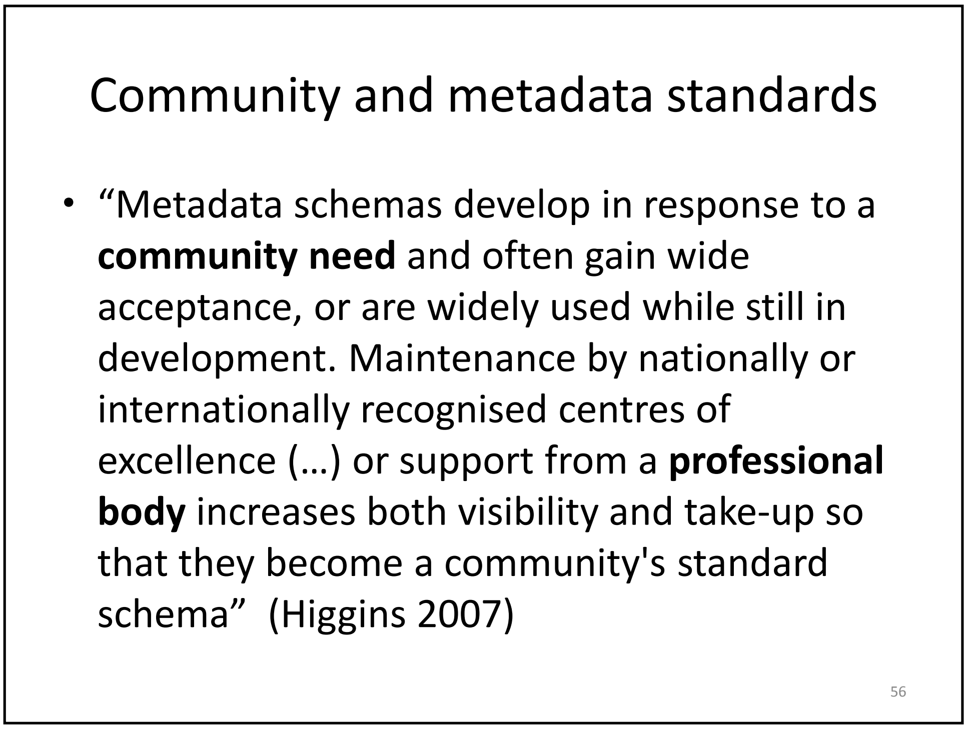
Next steps



Planned steps

- Submission of the « beta version » of LC metadata scheme to members of the Learner Corpus Association (LCA) http://www.learnercorpusassociation.org/
- Adaptation and finalization of the metadata scheme based on feedback received
- Concurrent work on representation formats
- Maintenance of the metadata standards by the LCA

55



Community and metadata standards

- “Metadata schemas develop in response to a **community need** and often gain wide acceptance, or are widely used while still in development. Maintenance by nationally or internationally recognised centres of excellence (…) or support from a **professional body** increases both visibility and take-up so that they become a community's standard schema” (Higgins 2007)

56

Representation format (1/2)

- Metadata location
  - Separate metadata file or file header
- “TEIfy” our LC metadata
  - Administrative metadata:
    - <fileDesc><titleStmt> ; <publicationStmt> <distributor> / <availability> <licence> ; <editionStmt> <edition>
  - Annotation metadata
    - TEI language corpora module
  - Corpus design / text / learner metadata
    - Longitudinal, proficiency levels, etc.

57

Representation format (2/2)

- Help from a TEI expert
- TEI customization? TEI extension?
  - See CMC project (Beißwenger et al., 2017)
- Submit a CLARIN CMI module for LC metadata

58

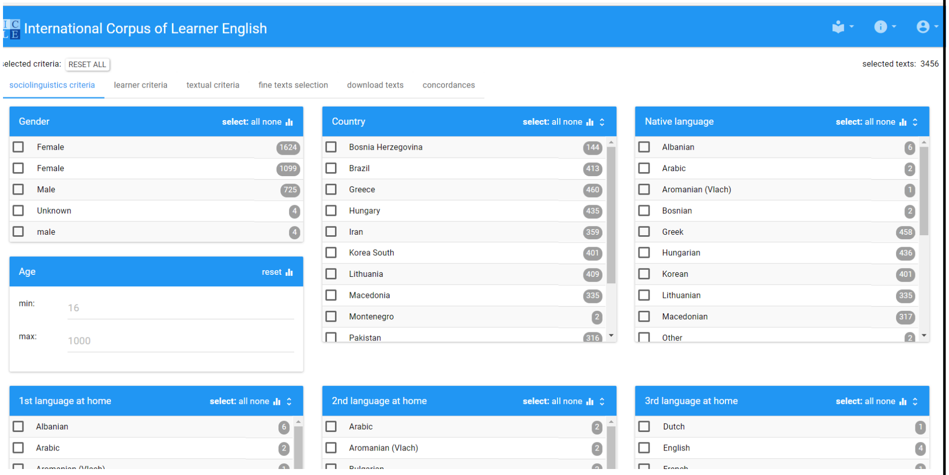
Metadata editor and user interface

- “the extent to which established standards are used is determined by the availability of tools that work with this standardized data” (Lehmberg & Wörner 2008)
- One major desideratum: **user-friendliness**
- Two essential tools
  - Metadata editor (cf. Koeva et al. 2016: “MetaEditor: a tool for manual metadata editing and verification”
  - User interface

Two examples of user interfaces

- **ICLEv3** (Granger, Meunier, Paquot & Dupont forthc. 2018)
  - Check boxes
  - Dynamic size visualization (number of texts and number of words)
- **Hypal4MUST** (Granger, Lefer & Obrusnik in preparation)
  - Pull-down menus
  - Conditional fields

ICLEv3



International Corpus of Learner English

selected criteria: RESET ALL

selected texts: 3456

sociolinguistics criteria | learner criteria | textual criteria | fine texts selection | download texts | concordances

Gender	select: all none
□ Female	1624
□ Female	1099
□ Male	725
□ Unknown	4
□ male	4

Age — reset

min: 16

max: 1000

Country	select: all none
□ Bosnia Herzegovina	144
□ Brazil	413
□ Greece	460
□ Hungary	435
□ Iran	359
□ Korea South	401
□ Lithuania	409
□ Macedonia	335
□ Montenegro	2
□ Pakistan	316

Native language	select: all none
□ Albanian	6
□ Arabic	2
□ Aromanian (Vlach)	1
□ Bosnian	2
□ Greek	458
□ Hungarian	436
□ Korean	401
□ Lithuanian	335
□ Macedonian	317
□ Other	2

1st language at home	select: all none
□ Albanian	6
□ Arabic	2

2nd language at home	select: all none
□ Arabic	2
□ Aromanian (Vlach)	2

3rd language at home	select: all none
□ Dutch	1
□ English	4

Hypal4MUST

- **Conditional input fields** (the choice of a field conditions the number and types of subsequent fields) and **pull-down menus**
  - – E.g. speech or writing. If speech: transcribed or keyed-in, sound files or only transcriptions, etc.
  - – E.g. task duration: untimed or timed. If timed: duration in minutes

Conclusion

Challenge of standardization efforts (Burnard 2017)

- “there is a long-running tension within all standardisation efforts consequent on an opposition between **generality** and **customization**. The more generally applicable a standard, the harder it may be to use productively in a given context; the more tailored it is to a given context, the less useful it is likely to be elsewhere. Yet surely one of the main drivers behind the urge to go digital has always been the ability **not just to have one’s cake and to eat it, but also to have many different kinds of cake from the same messy dough.** For this to work, there is a need for standards which do not limit choice, but rather facilitate an accurate presentation of the choices made”.

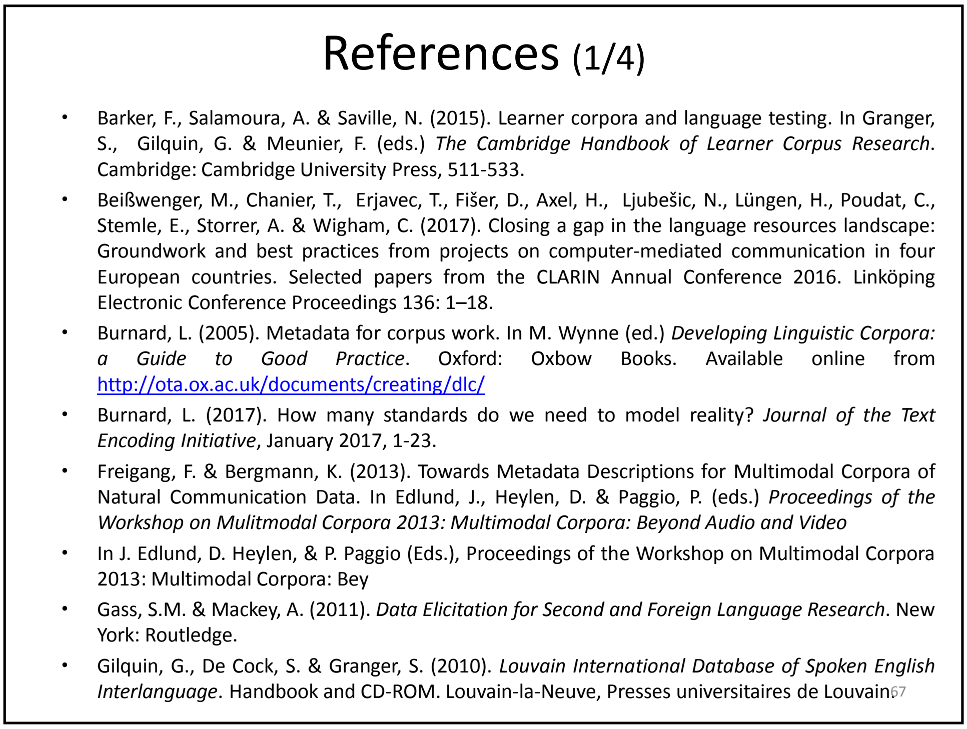
Learner corpus research

- Field in rapid expansion
  - *Learner Corpora around the World* webpage: 163 learner corpora
- Time to work on standardisation
  - Metadata, annotation, transcription
- Improve study quality (cf. also Paquot & Plonsky, 2017)

65

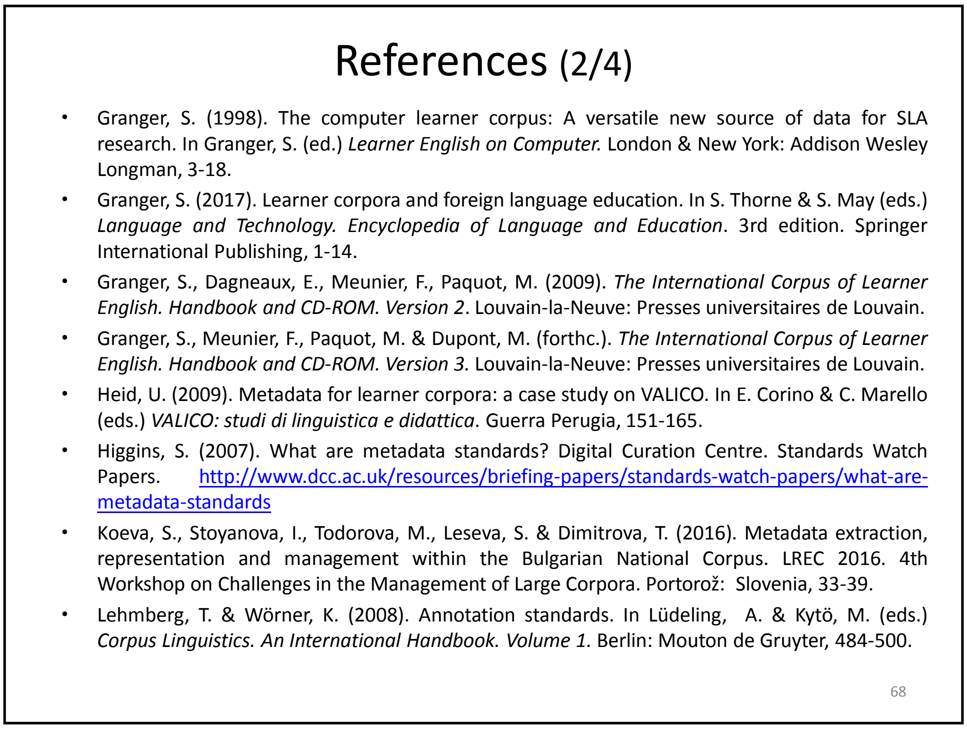
Thanks for your attention and feedback !

66



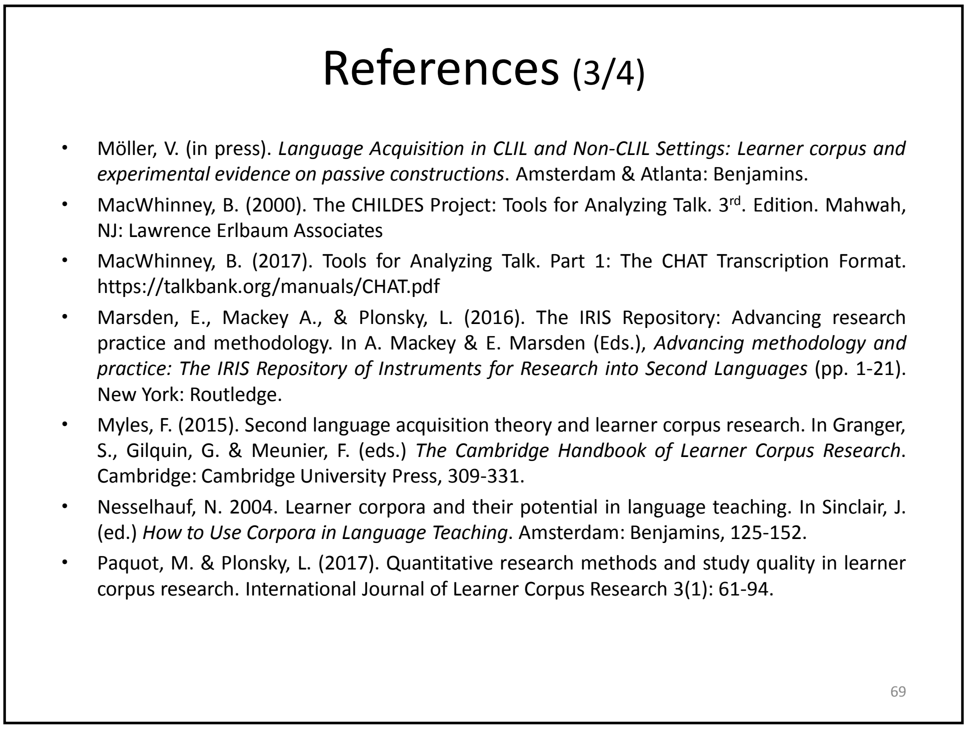
References (1/4)

- Barker, F., Salamoura, A. & Saville, N. (2015). Learner corpora and language testing. In Granger, S., Gilquin, G. & Meunier, F. (eds.) *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 511-533.
- Beißwenger, M., Chanier, T., Erjavec, T., Fišer, D., Axel, H., Ljubešic, N., Lüngen, H., Poudat, C., Stemle, E., Storrer, A. & Wigham, C. (2017). Closing a gap in the language resources landscape: Groundwork and best practices from projects on computer-mediated communication in four European countries. Selected papers from the CLARIN Annual Conference 2016. Linköping Electronic Conference Proceedings 136: 1–18.
- Burnard, L. (2005). Metadata for corpus work. In M. Wynne (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. Available online from http://ota.ox.ac.uk/documents/creating/dlc/
- Burnard, L. (2017). How many standards do we need to model reality? *Journal of the Text Encoding Initiative*, January 2017, 1-23.
- Freigang, F. & Bergmann, K. (2013). Towards Metadata Descriptions for Multimodal Corpora of Natural Communication Data. In Edlund, J., Heylen, D. & Paggio, P. (eds.) *Proceedings of the Workshop on Mulitmodal Corpora 2013: Multimodal Corpora: Beyond Audio and Video*
- In J. Edlund, D. Heylen, & P. Paggio (Eds.), Proceedings of the Workshop on Multimodal Corpora 2013: Multimodal Corpora: Bey
- Gass, S.M. & Mackey, A. (2011). *Data Elicitation for Second and Foreign Language Research*. New York: Routledge.
- Gilquin, G., De Cock, S. & Granger, S. (2010). *Louvain International Database of Spoken English Interlanguage*. Handbook and CD-ROM. Louvain-la-Neuve, Presses universitaires de Louvain



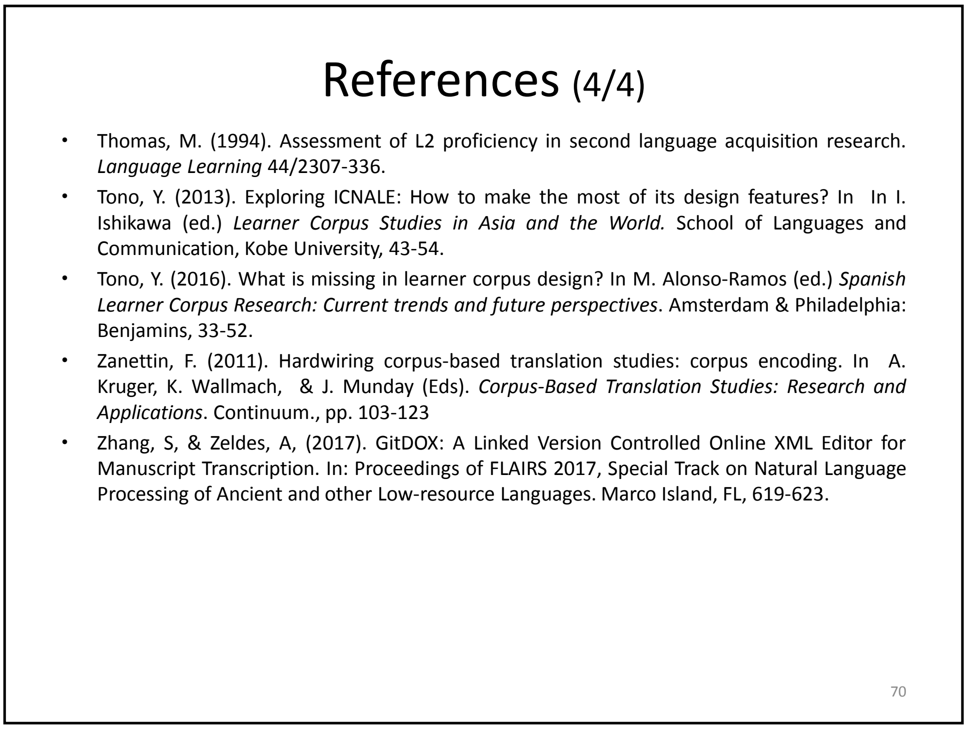
References (2/4)

- Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (ed.) *Learner English on Computer.* London & New York: Addison Wesley Longman, 3-18.
- Granger, S. (2017). Learner corpora and foreign language education. In S. Thorne & S. May (eds.) *Language and Technology. Encyclopedia of Language and Education*. 3rd edition. Springer International Publishing, 1-14.
- Granger, S., Dagneaux, E., Meunier, F., Paquot, M. (2009). *The International Corpus of Learner English. Handbook and CD-ROM. Version 2*. Louvain-la-Neuve: Presses universitaires de Louvain.
- Granger, S., Meunier, F., Paquot, M. & Dupont, M. (forthc.). *The International Corpus of Learner English. Handbook and CD-ROM. Version 3.* Louvain-la-Neuve: Presses universitaires de Louvain.
- Heid, U. (2009). Metadata for learner corpora: a case study on VALICO. In E. Corino & C. Marello (eds.) *VALICO: studi di linguistica e didattica*. Guerra Perugia, 151-165.
- Higgins, S. (2007). What are metadata standards? Digital Curation Centre. Standards Watch Papers. http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards
- Koeva, S., Stoyanova, I., Todorova, M., Leseva, S. & Dimitrova, T. (2016). Metadata extraction, representation and management within the Bulgarian National Corpus. LREC 2016. 4th Workshop on Challenges in the Management of Large Corpora. Portorož: Slovenia, 33-39.
- Lehmberg, T. & Wörner, K. (2008). Annotation standards. In Lüdeling, A. & Kytö, M. (eds.) *Corpus Linguistics. An International Handbook. Volume 1.* Berlin: Mouton de Gruyter, 484-500.



References (3/4)

- Möller, V. (in press). *Language Acquisition in CLIL and Non-CLIL Settings: Learner corpus and experimental evidence on passive constructions*. Amsterdam & Atlanta: Benjamins.
- MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd. Edition. Mahwah, NJ: Lawrence Erlbaum Associates
- MacWhinney, B. (2017). Tools for Analyzing Talk. Part 1: The CHAT Transcription Format. https://talkbank.org/manuals/CHAT.pdf
- Marsden, E., Mackey A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS Repository of Instruments for Research into Second Languages* (pp. 1-21). New York: Routledge.
- Myles, F. (2015). Second language acquisition theory and learner corpus research. In Granger, S., Gilquin, G. & Meunier, F. (eds.) *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 309-331.
- Nesselhauf, N. 2004. Learner corpora and their potential in language teaching. In Sinclair, J. (ed.) *How to Use Corpora in Language Teaching*. Amsterdam: Benjamins, 125-152.
- Paquot, M. & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. International Journal of Learner Corpus Research 3(1): 61-94.



References (4/4)

- Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning* 44/2307-336.
- Tono, Y. (2013). Exploring ICNALE: How to make the most of its design features? In In I. Ishikawa (ed.) *Learner Corpus Studies in Asia and the World.* School of Languages and Communication, Kobe University, 43-54.
- Tono, Y. (2016). What is missing in learner corpus design? In M. Alonso-Ramos (ed.) *Spanish Learner Corpus Research: Current trends and future perspectives*. Amsterdam & Philadelphia: Benjamins, 33-52.
- Zanettin, F. (2011). Hardwiring corpus-based translation studies: corpus encoding. In A. Kruger, K. Wallmach, & J. Munday (Eds). *Corpus-Based Translation Studies: Research and Applications*. Continuum., pp. 103-123
- Zhang, S, & Zeldes, A, (2017). GitDOX: A Linked Version Controlled Online XML Editor for Manuscript Transcription. In: Proceedings of FLAIRS 2017, Special Track on Natural Language Processing of Ancient and other Low-resource Languages. Marco Island, FL, 619-623.