Working together towards an ideal infrastructure for language learner corpora

Towards an infrastructure for language learner corpora

Egon W. Stemle (Eurac Research, Italy), Adriane Boyd (University of Tübingen, Germany), Maarten Janssen (University of Coimbra, Portugal), Therese Lindström Tiedemann (University of Helsinki, Finland), Nives Mikelić Preradović (University of Zagreb, Croatia), Alexandr Rosen (Charles University, Czech Republic), Dan Rosén (University of Gothenburg, Sweden), Elena Volodina (University of Gothenburg, Sweden)

Abstract

In this article we provide an overview of first-hand experiences and vantage points for best practices from projects in seven European countries dedicated to learner corpus research (LCR) and the creation of language learner corpora. The corpora and tools involved in LCR are becoming more and more important, as are careful preparation and easy retrieval and reusability of corpora and tools. But the lack of commonly agreed solutions for many aspects of LCR, interoperability between learner corpora and the exchange of data from different learner corpus projects remains a challenge. We show how concepts like metadata, anonymization, error taxonomies and linguistic annotations as well as tools, toolchains and data formats can be individually challenging and how the challenges can be solved.

1. Introduction

Compiling a corpus to conduct learner corpus research (LCR) is an intricate task. First, the design criteria for the learner corpus[1] need to be specified, then a sample of language learners must be defined and data collected, which in turn is often preceded by obtaining approval of a teaching or testing institution, unless the data

[1] The term *learner corpus* covers (written) texts produced by language learners, both studying their mother tongue (L1) and a new language (L2).

is collected directly, for example, in a testing and learning environment or through crowdsourcing.[2] And although the sample should be representative, it is often inevitable to compromise on this aspect. Additionally, data collection might be preceded by a mandatory approval of an ethics committee. If the data is not already born-digital it needs to be transcribed, which in turn requires to identify a transcription tool and to search for and instruct transcribers. Once the data exists in a digital form as a *corpus*, that is, an electronic collection compiled for a specific purpose, this corpus needs to be processed.

Processing often consist of both automatic and manual tasks. Automatic processing steps towards linguistically enriched texts are usually tokenization, lemmatization, and part-of-speech (POS) tagging, sometimes also named entity recognition (NER) or syntactic parsing, which are all more error-prone for second language (L2)[3] corpora (see Section 2.3). Semi-automatic or manual processing steps may include normalization (i.e., correction, emendation or target hypothesis annotation) and error annotation (i.e., identification of errors using an error taxonomy). Once the corpus is ready with its metadata and all annotations, analyses can be computed to answer research questions and the data can be explored using corpus query interfaces.

Suppose that researchers want to investigate the use of adverbs and find an appropriately annotated corpus covering the language and learner group of interest. What kinds of problems might they encounter if they wanted to use this corpus to answer their research questions? Moreover, what could be possible complications if they wanted to carry out a comparative study across languages? How could potential problems be resolved when using existing corpora, and what could have been done differently in the compilation and annotation phases, which would have improved their later reuse?

In this paper we describe an ideal world of learner corpora and LCR, primarily focussing on written corpora. Based on previous research (see, e.g., Rebuschat et al. 2017) and our own experiences, we reflect on problems that can arise while trying to achieve utopia and on how they might be solved or at least minimized by talking to each other across research areas and between projects, learning from each other, exchanging data, tools and experiences. Our arguments in this article will usually be illustrated using examples taken from corpora in Table 1. The corpora are all written corpora, except for COPLE2, which also contains a spoken part, and all

[2] For detailed discussions on the design of learner corpora see, e.g., Gilquin 2015 or Granger 2008.

[3] We use L2 to cover foreign language, second and third language or any other language that is not the L1.

except KoKo contain L2 learner language, yet in many cases our statements apply to all types of learner corpora[4].

Name	Learner Language(s)	Size in Texts (Tokens) *Expected Size*	Anonymized Transcribed Digitally Born	Full Access / License	Query Interface
ASK[5] (2006–2014)	L2: NB	1,700	A T	CLARIN Res (Priv)	Corpuscle
COPLE2[6] (2013–)	L2: PT	1,070 (227k)	A T	tbd	TEITOK
CroLTeC[7] (2016–)	L2: HR	1,042 (200k) (*~1M toks*)	A T D	tbd	TEITOK
CzeSL[8] (2010–)	L2: CZ	8.6k (1.1M)	A T	CC BY-SA 3.0	KonText, NoSkE
Falko[9] (2004–)	L2: DE, L1: DE control	L2: 641 (380k) L1: 152 (92k)	T	CC BY 3.0	ANNIS
KoKo[10] (2010–2015)	L1: DE	1,503 (811k)	T	CLARIN ACA +BY +NC +NORED	ANNIS, NoSkE
MERLIN[11] (2012–2014)	L2: CZ, DE, IT	2,287 (340k)	A T	CC BY-SA 4.0	custom platform, ANNIS
SweLL[12] (2017–2021)	L2: SV	340 (144k) *(~600 texts)*	A T D	CLARIN RES (Priv)	Korp (SV), Strix
Topling[13] (2010–2013)	L2: FI, EN, SV	10,350 (165k)	(A) T	CLARIN RES +NC +DEP 1.0	Korp (FI)

Table 1: Written learner corpora referenced in this article.

[4] For a discussion focusing on interoperability from the perspective of spoken corpora see, for example, Ballier & Martin (2013).

[5] http://clarino.uib.no/ask/

[6] http://alfclul.clul.ul.pt/teitok/learnercorpus/

[7] http://teitok.iltec.pt/croltec/

[8] http://utkl.ff.cuni.cz/learncorp/

[9] https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko

[10] http://www.korpus-suedtirol.it/KoKo/

[11] http://www.merlin-platform.eu

[12] https://spraakbanken.gu.se/eng/swell_infra

[13] https://www.jyu.fi/hytk/fi/laitokset/kivi/tutkimus/hankkeet/paattyneet-tutkimushankkeet/topling/en

In recent decades there has been an enormous upswing in LCR. McEnery (2018: xvii) notes that "[l]earner corpora are barely mentioned in the 1994 [TaLC] proceedings",[14] and in the 2014 TaLC proceedings (Brezina & Flowerdew 2018) "five chapters use well-established learner-corpora, while the remaining three analyse smaller datasets … which, however, could be considered DIY learner corpora of a kind" (McEnery 2018: xvii). In the introduction Brezina & Flowerdew (2018: 1) also emphasize that "[c]orpora play a *crucial role* in second language (L2) research and pedagogy" (emphasis added).

There have also been recent national and international efforts to coordinate data collection and tools development for different user groups, such as Språkbanken[15] in Sweden, ORTOLANG[16] in France and SADiLaR[17] in South Africa, all connected to the Common Language Resources and Technology Infrastructure (CLARIN)[18] in Europe and similar global entities.

These types of efforts have shown that there is a need for electronic research infrastructure consisting of (Volodina et al. 2016):

- Data in electronic form (ideally freely accessible)
- A user interface that facilitates the use of the material in various ways
- A platform for the collection and processing of new material
- The expertise needed

An ideal infrastructure for LCR would build on the concepts of interoperability developed within the general language resources. However, the specifics of learner corpora (among others, learner metadata and error annotation) need to be specified as an extra dimension of interoperability.

Interoperability has been operationalized from different perspectives by, among others, Chiarcos (2012) and Ide & Pustejovsky (2010), with metadata, search and retrieval, as well as formats and data documentation forming the common basis. Chiarcos (2012: 163) classifies the interoperability of language resources into *conceptual* and *structural*. The former deals with the heterogeneity of linguistic annotations with diverging definitions of identical terms or different underlying schemata altogether; here, terms and schemata are to be linked with a common vocabulary and embedded into an encompassing ontology. The latter deals with the heterogeneity of formalisms and data formats that tools use as input or output; here,

[14] As a comparison, 1994 was also the year that the BNC was first published.
[15] https://spraakbanken.gu.se/eng/research/infrastructure
[16] https://www.ortolang.fr/
[17] http://www.sadilar.org/
[18] https://www.clarin.eu/

tools should be able to exchange and process data seamlessly, which ideally enables simple substitutions of entire tools. We interpret the *conceptual* interoperability from the LCR user perspectives in Section 2 and the *structural* interoperability from the point of view of technical solutions used in the LCR domain in Section 3. In Section 4 we are summing up and looking ahead at the (most urgent) needs we can see if we want to facilitate and improve LCR.

2. Conceptual interoperability – user perspective

Potential users and possible uses of learner corpora are manifold. During the discussions at the CLARIN *Workshop on Interoperability of L2 Resources and Tools* in Gothenburg in December 2017[19], the approximately 25 participants with varied research and commercial interests listed the following target groups: language learners, language teachers, students, SLA researchers, general linguists, computational linguists, textbook developers, language test designers, software developers, and lexicographers, but there are certainly also some other target groups.

Different users turn to corpora with varied background knowledge, agendas, and expectations, which creates the need to meet very different documentation and usability requirements. Users must be aware of the actual content of the corpus and of the selection and meaning of metadata, such as sociolinguistic variables (see Section 2.1), error annotations, and target hypotheses (see Section 2.3.1), and of linguistic annotations, like POS labels (see Section 2.3.3). Here, the specifics of semi-automatic and automatic annotations must be taken into account. Automatic tools are built – if not explicitly then implicitly – according to a linguistic theory, and this theory bleeds into their output, including their semi-automatic output. Finally, two or more interfaces tailored to the specific needs or technical skills of different target groups are usually provided for search and retrieval access to learner corpora (see Section 2.4).

Achieving conceptual interoperability, that is, an agreement on definitions of shared concepts, and thereby offering users uniform interpretations of terms is in itself a major challenge. In addition, updating the version of a tool for (semi-)automatic annotation, updating, revising or extending corpus data or changing implicit default settings in the search interface may all cause changes in the data used for later analyses. This is all the more true for changes in annotation

[19] https://sweclarin.se/eng/workshop-interoperability-l2-resources-and-tools

principles and metadata conceptualizations. Looking at all this from the perspective of comparability of different corpora and different research studies, which should be objective, repeatable, and reproducible, it is striking how difficult it is to compare the results from one corpus with results from another corpus or even with results from the same corpus at another time.

In order to improve conceptual interoperability, it is necessary to provide comprehensive descriptions of corpora, including the type of metadata and annotations, and a description of the different categories used, and how all the data were collected and added. However, the documentation should not only focus on the immediate use case, but should be comprehensive enough to adequately cover unexpected usage scenarios and therefore trace, document and make available changes to *all* involved processing tools and interfaces as well as annotation and metadata principles – ideally incorporated into the corpus creation and maintenance procedure.[20]

Below we take a closer look at some of these issues and what might make it difficult to get the most out of comparable corpora.

2.1. Metadata

There is a wide range of metadata that can be imagined as desirable in learner corpora, and in the best of all worlds, the community would have already agreed on a metadata schema that defines mandatory and optional values. So we would know what to collect and how, and we would have a shared understanding of what the data meant. Granger & Paquot (2017) list 10 pages of categories that second language acquisition (SLA) specialists and other linguists would like to see recorded, some of which are claimed to be obligatory: for example, target language (or even languages) of the corpus, L1(s) and other L2 language(s) in the participants' environment, editorial decisions (in written corpora), proficiency level, nationality of the participants, place where the data were collected and institution. In any case, the metadata that is finally included in the corpus has to be filtered according to legal aspects (see Section 2.2 below), availability, and relevance sought by the initial compilers, as well as technical know-how. Guidelines could be collected as a collaborative open database with metadata recommended for LCR corpora, with examples and best practices for modelling the data, an indication of their importance (at the level of mandatory interest) and

[20] See Glaznieks et al. (2014) and Kermes et al. (2016) for detailed examples and discussions of corpus creation and maintenance procedures.

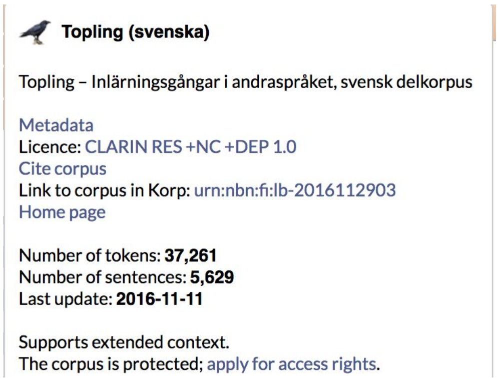
whether they pose data protection problems at a(n) (inter)national or European level.

Burnard (2005: 30) defines metadata as “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”. Today, this *discovery metadata* tends to be communicated through search and discovery services such as CLARIN’s Virtual Language Observatory (VLO)[21], within the corpus data or on the corpus website – but sometimes only in articles. Additionally, corpora usually also have metadata information for individual files or some other finer grained division. This latter metadata is usually encoded in the corpus, either in the individual files or in a separate document or database.

However, a systematic coding of the metadata alone is not sufficient; in order to facilitate search, retrieval and analysis of the corpus data, the metadata must also be usable in the search interface (see Section 2.4), or as Gilquin (2015: 17-18) puts it: “recording all these metadata is of little use if they are not made available to the corpus user, together with the actual data produced by the learners.” In addition, someone who uses the corpus with the metadata needs more detailed information than the annotation schema: for example, values like L1(s), L2(s), native language, foreign languages, preferred languages, etc., some of which are certainly part of each corpus, are strongly dependent on their interpretation within a theory, so that a detailed description seems necessary – a community-wide understanding desirable, but also very challenging. The assignment of a value for the metadata proficiency level can also be unclear: does the Common European Framework of Reference for Languages (CEFR; Council of Europe 2001) B1 level mean that someone has taken and passed an exam for students who were at the B1 level? In this case, the informant might have an even higher level. Or does it mean that someone explicitly rated this text as a B1 level text, or that the exam was part of a B1 level course? Some learner corpora, for example, CzeSL, SweLL, and Topling (Martin 2011) make metadata descriptions and documentation available through links in the concordance tool (see Figure 1 for an example from Topling).

Clearly, information on data collection and the tasks used to collect the data is of paramount importance as it affects the language output in a number of ways, for example, Hinkel (2002: 162-164) shows that the amount of variation between L1 and L2 writers depends on the prompt, Golcher & Reznicek (2011) show that topic affects the text more than L1 does (see also Caines & Buttery 2018). Most projects do include some metadata about the task, but the amount and type varies. In addition, there is the question of the format in which this is accessible.

[21] https://vlo.clarin.eu/



Topling (svenska)

Topling – Inlärningsgångar i andraspråket, svensk delkorpus

Metadata
Licence: CLARIN RES +NC +DEP 1.0
Cite corpus
Link to corpus in Korp: urn:nbn:fi:lb-2016112903
Home page

Number of tokens: **37,261**
Number of sentences: **5,629**
Last update: **2016-11-11**

Supports extended context.
The corpus is protected; apply for access rights.

Figure 1: Display of corpus information with links in the corpus selection menu in KORP (FIN-CLARIN edition; Borin et al. 2012).

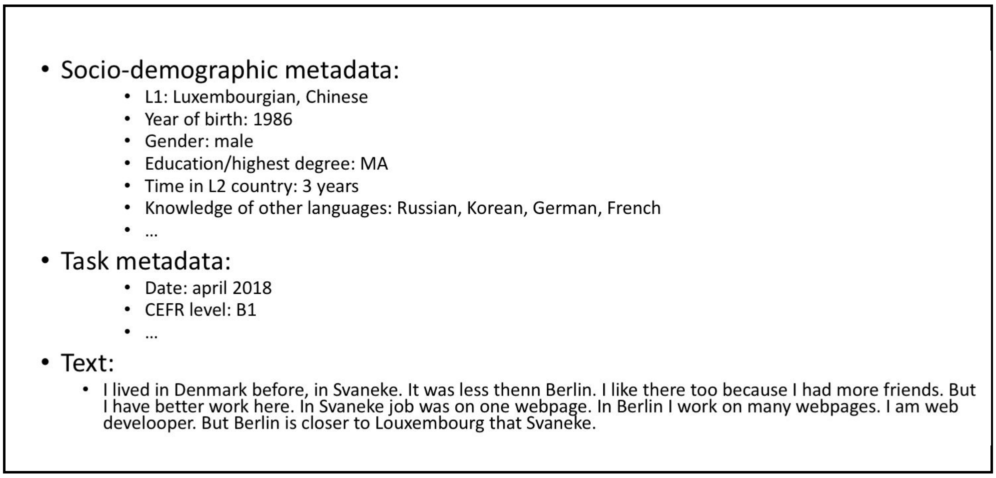
Within SLA it has been common practice to collect data only for use by your own project. This means that when asking for consent the name of the researcher, research project, research institute was probably mentioned and the consent forms only give permission to use the data within that group. This has several implications, some of which will be discussed below in the context of legal and privacy issues (see Section 2.2), but there are also implications for metadata. If data is considered to be something that is only collected for use within a specific project, it is quite possible that information about certain background questions is not collected because it is considered irrelevant, and it may seem more important that completing the background questionnaire does not become too tiresome for informants. However, this may restrict the future usability of the corpus, even if it was concluded that it could be shared with a wider research community. Hence the importance of the above-mentioned initiative undertaken by Granger & Paquot (2017) to standardize learner corpus metadata between projects, something that should be taken into account when setting up a new learner corpus project.

2.2. Legal and privacy issues

Spreading an electronic resource through an infrastructure entails responsibility to the *data subjects*, in our case language learners who have agreed to provide their texts and personal information. The requirement of collecting and storing informed consents, the obligation to remove a learner and erase their data from the registers upon their request as well as national and international laws and ethical regulations regarding personal integrity and discrimination create certain tensions in the opening of data for all types of uses. To justify that the data should be accessible to users outside individual projects, handling of data must be 'bulletproof' at each stage, and several stages have to be considered, namely data acquisition, data storage, data aggregation, data analysis, data usage, data sharing, and data disposal (Accenture 2016). Most of the steps deal with organizational and management decisions/precautions or preparatory steps before uploading data to the infrastructure. In the text below, we focus on those stages relevant to infrastructure usage where learner specific characteristics in the texts and metadata present risks at the *data usage* and *data sharing* stages (Volodina et al. 2018, Volodina & Megyesi 2018).

2.2.1. Personal non-identifiability

Within European countries there exists the requirement to ensure personal non-identifiability when adding essay information with personal metadata. According to the EU General Data Protection Regulation (GDPR), Article 4,[22] "*personal data* means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person". Consider Figure 2, where the addition of information from the two sources – a learner text and metadata – can give away a learner. Even though the name as such is not revealed to the data users, indirect clues can be used to identify a person.



- Socio-demographic metadata:
  - L1: Luxembourgian, Chinese
  - Year of birth: 1986
  - Gender: male
  - Education/highest degree: MA
  - Time in L2 country: 3 years
  - Knowledge of other languages: Russian, Korean, German, French
  - …
- Task metadata:
  - Date: april 2018
  - CEFR level: B1
  - …
- Text:
  - I lived in Denmark before, in Svaneke. It was less thenn Berlin. I like there too because I had more friends. But I have better work here. In Svaneke job was on one webpage. In Berlin I work on many webpages. I am web develooper. But Berlin is closer to Louxembourg that Svaneke.

Figure 2: Example of (selected) metadata and an essay text for an imagined learner.

In view of this, the SweLL project, for example, adopted a rather restrictive approach to metadata (Megyesi et al. 2018). Thus, neither a student's country of origin or nationality (restricting information to L1 only) nor the year of birth are

[22] https://gdpr-info.eu/art-4-gdpr/

given, but rather a 5-year span (e.g., 1970–1974) in order to complicate the possible identification of a learner through aggregated personal information. For the same reason, no information is provided on the educational establishment in which the essays were collected. This is a natural consequence of the national Swedish legislation on open access to public data (Riksdagen 1949, kap 2) on the one hand and the stricter GDPR on the other.

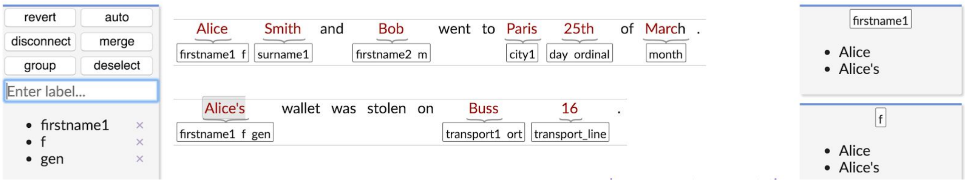
2.2.2. Anonymization and pseudonymization

Ethical Review Boards set further requirements on so-called *sensitive data*, that is, data that can reveal the sexual orientation, religion, political views, or ethnicity of a (potentially identifiable) person, which can lead to discrimination. Unless it can be ensured that the person behind the (meta)data is not disclosed, Ethical Review Boards are entitled to make a request to list all potential scenarios for data usage and also restricts data usage to project-internal use only. This in itself is counter-productive since a research infrastructure aims to provide researchers *outside the project* with electronically *available* data for *potential research questions* that are not always foreseen in advance.

In order to minimize personal identifiability from a text and to make learner data less "sensitive" (as defined by the Ethical Review Boards), learner essays must be *anonymized*, not only in the sense that information in the metadata and in the text itself that could give away the author must be removed, but also in the sense that all personal information about identifiable persons in the text must be suppressed; this is also referred to as *de-identification*. In cases where this is not possible, the only alternative is to omit the texts altogether.

Instead of replacing content by spaces or some form of *XXX*, personal data are often replaced by a code in (learner) corpora. So instead of the place name *Svaneke* in Figure 2, the ASK corpus (Tenfjord et al. 2006) would use @place, CzeSL would use {village}<priv>, and COPLE2 (Mendes et al. 2016) would use PP. This has the advantage of preserving some information about what kind of name (or more precisely, named entity) was used in the original. The codes can be structured to provide more information about the named entity used, for example @firstname_1_female_genitive for the first female name in the text used in genitive case. But codes have the disadvantage that they make the text harder to read, and if the code is not completely transparent, one needs to know the codes beforehand to interpret them. To keep the text readable, several corpus projects opt to replace the names in the text with other real-life names, that is, pseudonymize them. Typically, the most frequent names in the language of the corpus are used, so CzeSL and CroLTeC replace first names with common names in Czech and Croatian respectively, which include *Adam* and *Eva* for Czech, and *Ana, Ivo* and *Vanja* for Croatian. The disadvantage of using pseudonyms, however, is that they make it less clear what the student actually wrote and can both eliminate and introduce mistakes.

In order to combine the advantages of codes and pseudonyms, the data in the SweLL project is anonymized in two steps (Megyesi et al. 2018). The first step is to *mark up* named entities using *placeholders* that consist of a code, like "firstname1 f" in Figure 3. Named entities that are replaced by placeholders are those that involve (1) information that can directly or indirectly reveal the author, and (2) sensitive information about the author. The second step is to automatically *render* the placeholders as real names (pseudonyms). For example, for 'firstname1 f' in Figure 3, a female name is randomly selected from a list of names, in this case *Alice*. This approach has the advantage that it provides more control and a reduced chance to miss information, while being quicker than directly annotating pseudonyms.



Alice Smith and Bob went to Paris 25th of March .
firstname1 f | surname1 | firstname2 m | city1 | day ordinal | month

Alice's wallet was stolen on Buss 16 .
firstname1 f gen | transport1 ort | transport_line

Figure 3: Display of the SweLL anonymization tool.

When using pseudonyms, whether inserted manually or created automatically, there are quite a few things that need to be taken into account, including:

- In the substitution, case needs to be preserved: if we replace *John's* by *Jim* in *John's* car without keeping the genitive, the pseudonymized version would introduce an error the student did not make.

- Especially in the case of automatic rendering, it is important to indicate that a substitution has been made to alert the user that this is partly not the actual text written by the student.

- It is important to substitute different names differently, to avoid sentences like *I have three sisters: Aisha, Karima, and Nura* to become *I have three sisters: Ana, Ana, and Ana*, while still keeping the same pseudonym for a specific name throughout the text.

- The classification of names should be sufficiently fine-grained to ensure selection of the right kind of substitution: in Portuguese, you say *Lisboa*,

> but *o Porto* (with an article), so *Lisboa* should never be substituted by *Porto*.

In the case of automatic rendering, the code should ideally contain all the information needed to render the names correctly, which includes at least a co-indexation number, a fine-grained classification, and an indication of the case. For a morphologically rich language such as Czech, the anonymization placeholder for an occurrence of *Prahy* (Prague.ACC) would need to contain at least the following information in order to render it correctly (here represented in XML using the notation adopted in TEITOK): mark it in the text as the first mention of a city whose name has female gender and singular number and is used in the accusative case:

```
<anon type="city" subtype="fem_sing" case="acc" n="1"/>
```

But even so, there are things that automatic rendering will likely never be able to do, such as keeping spelling or morphological errors in the substituted named entities, as was done manually in CzeSL: to preserve a wrong suffix in the pseudonym, *dám to Petrovy* would be substituted with *dám to Adamovy* instead of the correct *dám to Adamovi* (I'll give it to Adam).

Suggestions to anonymize "structured" or "listed" types of personal information (e.g., personal names, city names, telephone numbers, etc.) can be supported by the use of automated methods. In the medical domain (El Emam & Arbuckle 2013) many such techniques have been developed, but there does not seem to be any LCR applications at this time. In order to obtain the information necessary for the correct rendering of the codes, POS taggers can be used to determine the case and number of the names (see section 2.3.3), and NER tools can be used to detect the names in the text. Of course, both NER and POS tagging must be applied before anonymization to work properly, and processing the non-anonymized text automatically might not always be allowed,[23] so anonymization sometimes has to be done entirely by hand. And although NER and POS can help with anonymization, it is always necessary to check manually because names can be missed by the automated processing and not all names need to be anonymized, but only those that are associated with private information. Automatic techniques for anonymizing sensitive data can only handle simple names and references. So-called "unstructured" types of potentially sensitive information (e.g., *We were happy to participate in a demonstration against Erdogan*) will still need to be marked manually or left out of the corpus completely, as is the case in the CroLTeC corpus.

[23] For example, if the processing pipeline accesses external resources, and thus the data must also be shared with these service providers.

The protection of the integrity of a person is a fundamental good, and in recent years it has received a great deal of attention, particularly through legislation. In any case, LCR research here must continue to be developed, adapted to the national and international laws and brought in line with ethical principles in order to establish and promote best practices. At the moment, it seems that using placeholders with (or without) rendering is the best way forward. What needs to be studied here in particular is the taxonomy of these named entities for pseudonymization and their granularity.

2.3. Annotation

LCR often distinguishes between *linguistic annotation* steps, for example, tokenization, morphosyntactic tagging, lemmatization and parsing, and annotation steps where deviations from the standard version of the target language as well as sometimes concrete correct uses in the target language are marked. The former are typical computer linguistic processing steps (see Section 2.3.3), the latter are characteristic of learner corpora and we will treat them together under the label *error annotation* (see Section 2.3.1).

2.3.1. Error annotation

To approximate a first corrected version of a text, suggestions from spelling or grammar checking tools can be used (Bolt 1992; Granger & Meunier 1994). Based on this – or in most cases alternatively – a manual step is inserted, which we interchangeably call *normalization* or *target hypothesis* (TH; Lüdeling 2005) *annotation*, before using a standard annotation pipeline.[24] Most significantly, the original text is not replaced by the normalized version, but the normalization is added as an annotation layer alongside the original text. On the other hand, in some projects, errors are not annotated at all, only normalized, or only error-tagged (with an implicit TH). In many projects, normalization and error annotation are combined in a single step, but they can also be performed separately; annotators are usually instructed to proceed one way or the other, but often can still make up their own minds.

For an annotation target in the corpus, either a single annotation is assigned within one and only one error annotation phase, or different annotations with often different perspectives are allowed in multiple phases. The annotation schemes then

[24] See Meurers (2015) for an overview of methods for automatic annotation of learner corpora.

have to fulfill different requirements depending on the conceptual decisions regarding error annotation and normalization, and this is reflected in the used format and tools. Finally, even with a single selected annotation concept, various formats or tools may still be available.

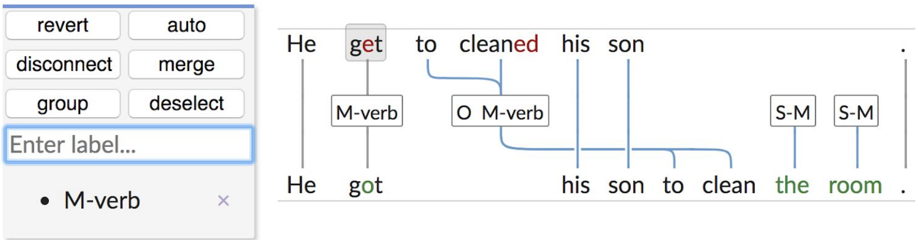


Figure 4: Display of the normalization and error-annotation tool in SweLL.

Errors may also agglomerate on or around a single text segment, for example, a misspelling, an affix that is incompatible with the stem or the case-assigning head, a wrong lexeme, and an unusual word order. Here, the annotation scheme should be able to represent all error types simultaneously. Another observed requirement is to support multiple THs, which in turn can be considered as alternative interpretations or as successive corrections. In such cases, a multi-layer (*parallel*) scheme can be used in which a full-fledged target hypothesis can be annotated in addition to the original version (see, e.g., Golcher & Reznicek 2011, Hana et al. 2010, Boyd et al. 2014, Rosén et al. 2018), possibly accompanied by one or more intermediate layers of analysis. Figure 4 shows the SweLL normalization and error-annotation tool (Rosén et al. 2018): The line above shows the original text, the line below shows the TH, with error labels on the connections between the source and target segments.

Another requirement of the annotation scheme may be to support multiple alternative annotations for both normalization and error annotation. For example, when a learner writes *a friend advice*, it is not clear whether she meant *a friend's advice* or rather *a friendly advice* – and the error type would be different in the two possible TH. In the absence of clear indications as to which of these interpretations is correct, it is sometimes desirable to annotate both alternatives. However, since alternative annotations represent a substantial burden on the format, tools, downstream processing, and user interaction, most annotation schemes force the annotator to make a particular choice.

On the other hand, normalization can also turn its context ungrammatical, so that another requirement for annotation schemata may be to enable and identify subsequent corrections. For example, when a noun head is replaced by a noun of a different gender, the form of an agreeing adjective must be modified to yield a well-formed TH for the noun phrase. In the ASK, SweLL, and CzeSL projects, there is a special error code for this that indicates a subsequent correction.

2.3.2. Error annotation taxonomies

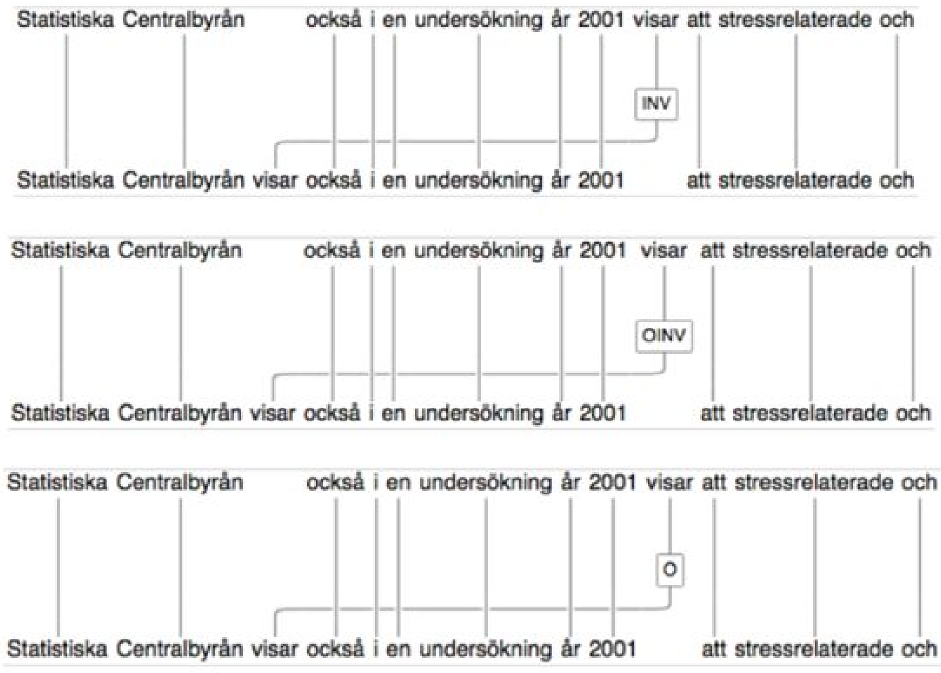
We start with an anonymous quotation: “Taxonomies are like underwear; everyone needs them, but no one wants someone else’s.”[25] With respect to error annotation projects, this is both true and false. Although very few learner corpus projects have managed to reuse each other’s error taxonomies so far, several projects have tried to build on previous work. In the following we will demonstrate the problems of re-using someone else’s taxonomy with an example from the SweLL project. Since the SweLL project is at an early stage, there is a direct incentive to learn from the experiences of other projects to ensure some comparability. To this end, the SweLL project has looked into several *error annotation taxonomies*, i.a. those of ASK and MERLIN .

First of all, the seemingly simple matter of what an *error* is may be complicated, especially when the project is cross-disciplinary. In the SweLL project, SLA researchers had strong objections to the use of the term *error* in relation to the development of learner interlanguage (Selinker 1972: 211-215), a term otherwise widely adopted within natural language processing (NLP) and LCR. Several other terms were proposed and considered, including *norm deviations*, *interlanguage phenomenon* (Díaz-Negrillo et al. 2010), *non-norm adequate form* (Dobric 2015) and *unexpected uses* (Gaillat et al. 2014). Furthermore, a distinction is made between *error correction* (i.e., adding target hypotheses) on the one hand and *correction annotation* (i.e., adding error codes) on the other (Zaghouani et al. 2015). So far, however, *error* is still used with reference to *error* annotation and *error* taxonomies since no other term has been unanimously adopted (Volodina et al. 2018).

The initial SweLL annotation schema was devised as a result of testing the ASK taxonomy and the MERLIN taxonomy on a set of Swedish essays (Volodina et al. 2018). Not surprisingly (see, e.g., Bayerl & Paul 2011; Fort et al. 2012), it turned out that annotating with the 64 tags of the MERLIN taxonomy took more time (twice as much) than annotating with the 23 tags of the ASK taxonomy, and additionally left a lot of inter-annotator disagreement. Since high reliability was more important than a very detailed annotation scheme, the ASK taxonomy was

[25] The related quotation “Standards are like toothbrushes. Everybody wants one but nobody wants to use anybody else’s.” sems to be attributed to Connie Morella: apparently she said this at the ANSI’s World Standards Day awards dinner in 2006. We use a variant from a presentation at the CLARIN workshop on interoperability of L2 resources and tools.

adopted with several modifications and was tested in a pilot study with the involved researchers. Once again, practical usage of the taxonomy led the SweLL researchers to important insights regarding tag names and their coverage. For example in Figure 5, where the Swedish verb [*visar*] (English [*shows*])) was moved to a new place in the corrected version, the three annotators agreed on both the segment in need of correction (the verb [*visar*]) and on the target hypothesis (different but identical placement of the verb *visar* in the corrected sentence), but *not* on the error label. *INV, OINV, O* describe various types of word order errors where *INV* covers "Non-application of subject/verb inversion", *OINV* "Application of subject/verb inversion in inappropriate contexts", and *O* "other word (or phrase) order error". As a result of the pilot study, the SweLL project adopted a new tag, *S-finV*, that covers cases of all word order errors with finite verbs.



Statistiska Centralbyrån också i en undersökning år 2001 visar att stressrelaterade och

Statistiska Centralbyrån visar också i en undersökning år 2001 att stressrelaterade och

Statistiska Centralbyrån också i en undersökning år 2001 visar att stressrelaterade och

Statistiska Centralbyrån visar också i en undersökning år 2001 att stressrelaterade och

Statistiska Centralbyrån också i en undersökning år 2001 visar att stressrelaterade och

Statistiska Centralbyrån visar också i en undersökning år 2001 att stressrelaterade och

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.

Figure 5: Three incompatible error annotations from different annotators in the SweLL error annotation pilot.

Other examples of confusing tag names were

- *PART* : overcompounding, i.e., a word needs to be split into two parts to secure correct spelling, e.g., **varjedag → varje dag* (‘every day’) and
- SPL : oversplitting, i.e., two tokens need to be written together to secure correct spelling, e.g., **person nummer → personnummer* (‘personal number’)

Intuitively, annotators tended to think of a spelling mistake in reference to the *SPL* tag, and of errors with participles in relation to the *PART* tag. The ‘perspective’ of the annotation tags was also a source of disagreement between annotators: does the tag describe the mistake made by the learner (SPL means “has unnecessarily split the word”) or the necessary action that corrects an error (SPL means “needs to be split to get a correct form”)? In this case, the SweLL project joined the two codes into one, *O-COMP*, covering all orthographic errors in compounding.

As a consequence of this SweLL pilot, both the ASK tag names and the number of tags have been reviewed to avoid ambiguity – leaving very little of the original taxonomy as a result. The strongest argument for reviewing the ASK taxonomy was the *possible drop in annotation quality* unless the tagset is reduced or changed, an idea also supported in previous annotation projects (see, e.g., Fort 2016: 24-43).

Even when the error taxonomy and guidelines are in place, various issues are reported from the error annotation process itself. For example, in the CroLTeC project, despite the fact that all annotators received training on the error taxonomy they tended to annotate the cause of the error (mostly the L1 influence), instead of concentrating on the description of the error itself. Low inter-annotator agreement was observed, especially concerning the treatment of multi-word errors. Another observation was that training annotators with essays corrected by teachers of Croatian as a foreign language did not prove to be helpful. The language teachers often corrected only those errors that represent the inadequate adoption of the language material that had been presented during the class at the specific learning level, and these corrections were not helpful for linguistic annotators at a later stage.

While the MERLIN corpus adapted the existing Falko annotation guidelines for target hypothesis annotation (Reznicek et al. 2012), a new annotation scheme was developed given MERLIN’s focus on illustrating the CEFR levels with authentic learner data. The MERLIN annotation scheme (Wisniewski et al. 2014) includes error annotation along with other linguistic characteristics that have been derived from multiple sources: operationalization of the CEFR level descriptions, SLA and language testing research, teacher and expert interviews, and experientially derived indicators (Boyd et al. 2014). Although existing annotation schemes for learner corpora in multiple languages were taken into consideration during the creation of the MERLIN annotation scheme, a new scheme that could be applied to multiple languages was a core contribution of the project.

The manually annotated part of CzeSL is based on a parallel annotation scheme consisting of three layers (Rosen 2015): the original and two layers of annotation with *m*:*n* links between tokens at the neighbouring layers. The links may be labeled by error tags. The two stages of the manual annotation scheme reflect the distinction roughly between errors in orthography and morphemics on the one hand (mostly in non-words) and all other error types on the other. Thus the second layer shows a string consisting of correct Czech forms, even though the sentence may not be correct as a whole. The third layer captures the rest of errors, resulting in a grammatically correct, though stylistically not necessarily optimal target hypothesis. In an automatic post-processing step some implicitly assumed error tags are made explicit and "formal" error tags are assigned, based on the comparison of the non-standard and corrected forms (Jelínek et al. 2012).

Figure 6 shows the sentence *budu se vratit domu*, literally 'AUX.1SG PTCLE return.IMPF.INF home.DIR', normalized as *vrátím se domů* 'I will return home'. The forms *vratit* 'return' and *domu* 'home' are manually corrected due to missing diacritics at the second layer. The annotator used the error tags *incorBase* (for an error in the stem) and *incorInf* (for an error in the inflectional suffix), with *stylColl* as an additional tag (for a colloquial form). The tags *formQuant0* (missing diacritics denoting quantity) are assigned automatically. The fully normalized third layer shows a proper synthetic form for future tense of the perfective verb. The more general *vbx* (verbal complex) error, assigned by the annotator, was made more specific in the post-processing step, resulting in the *cvf* type (compound verb form).

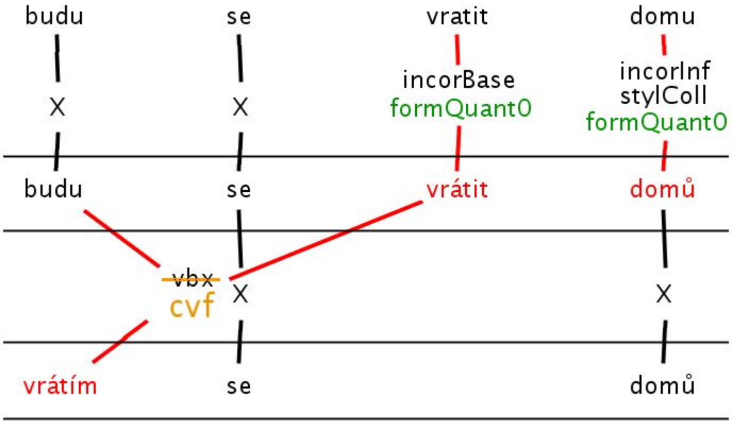


Figure 6: Three layers with links labeled with error types in CzeSL.

This scheme is a compromise between a linear annotation and a scheme with an arbitrary number of layers. Unlike some other schemes annotating multiple tokens, it preserves links between split, joined and reordered tokens across the layers, or between tokens corrected in two stages.

The COPLE2 corpus for Portuguese introduced an implicit error annotation (del Río et al. 2016): Instead of marking the error with a code, the type of error is implicitly indicated by several successive target hypotheses and annotations. There is a purely orthographic correction, a morphosyntactic and a lexical correction, all of which can have a target hypothesis, a morphosyntactic tag, and a lemma. An example from the MERLIN corpus is shown in Figure 7: three different THs concerning a single word (depending on how much correction is done) with their respective linguistic annotations, where the items marked in grey are redundant since they are identical to the preceding level.

<table>
<tr><td colspan="3">… von 20 in Stadt X exestierte Baugenossenschaften …
<i>… from 20 of the building cooperatives [existing] in City X …</i></td></tr>
<tr><td><i>orthographic</i>
TH: existierte
POS: ADJ_ACC.PL
Lemma: existiert</td><td><i>morphosyntactic</i>
TH: existierten
POS: ADJ_DAT.PL
Lemma: existiert</td><td><i>lexical</i>
TH: existenten
POS: ADJ_DAT.PL
Lemma: existent</td></tr>
</table>

Figure 7: Implicit multi-layered error annotation in MERLIN.

From this, much about the type of error can be derived without having to rely on an often theory-dependent error tag: if the word has an orthographically corrected form, it is hence an orthographic error, where a comparison between the two forms tells you exactly what the error(s) are, and the POS tag tells you exactly which kind of word the error was made on. If the word has a morphosyntactically corrected form, it was a morphosyntactic error, which you can subdivide as you wish, for instance gender and number errors on an adjective are agreement errors. This means that you know not only whether it was a number/gender error, but also which number/gender, and in which direction the error went. And if it has a corrected lemma, it is a lexical error. So to find all the lexical errors, you have to look for the words in the corpus that have an (explicit) value for the lexically corrected lemma. If traditional error tags are needed, for instance to facilitate searching, they can be generated automatically from the layered information.

2.3.3. Linguistic annotation

For learner corpora, linguistic annotation, like tokenization, morphosyntactic tagging, lemmatization, and parsing, is problematic in at least two different ways. Firstly, automatic annotation tools do not perform as well on non-standard language. Secondly, it is not always obvious what should be annotated in the first place.

While standard language can be relatively accurately annotated with existing automatic methods, annotating learner language with the same tools is more error-prone due to various (and often overlapping) types of errors, for example, in:

- segmentation: **I have two friends who called S and P they from Afghanistan too I like them becase when i need help they help me and i help them glad*
- spelling: **sommer*, **kulture*
- morphology and agreement: **I has was*
- word order: **We wrote down it.*

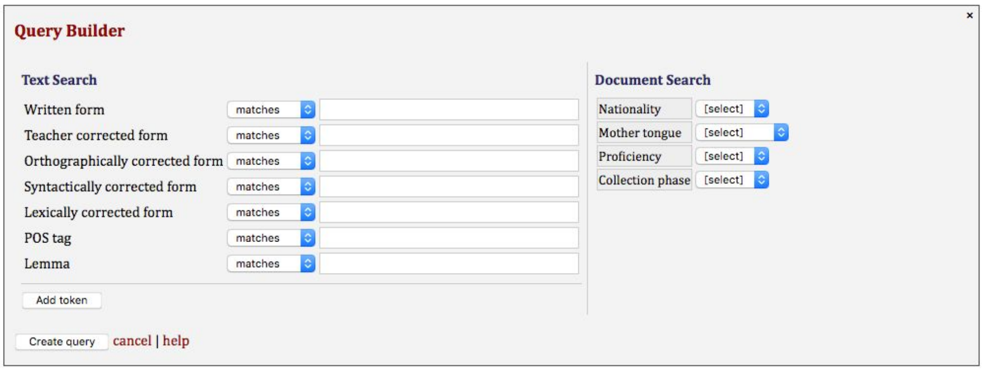
At high proficiency levels, such problems may be less pronounced, but for A1-level corpora, they are very prominent; and when the L1 and L2 are from very different language families, it is often far from clear what the student meant in the first place, which not only makes it difficult to process the original, but even to formulate a TH at all. On the other hand, some factors of the learner language, such as shorter sentences and simpler syntax, may also make a text including non-standard phenomena easier to process. As a starting point for further considerations, which must certainly include the morphological richness of the target language and the respective L1(s) of the language learners, Glaznieks et al. (2014) should be mentioned here; they showed significant improvement of POS tagging results after correcting orthographic errors and uncommon abbreviations in German L1 texts written in secondary school classrooms under standardised conditions one year before the school-leaving examinations. Ultimately, automatic annotation tools producing disappointing results should not be overestimated because due to the often small size of the learner corpora, tagging errors could be corrected by hand.

But all these problems mentioned above are also challenging conceptually: the question what should be annotated needs to be answered (see, e.g., Díaz-Negrillo et al. 2009). When a student writes *walked rapid*, do we want to tag the fact that the student used an adjective, or that the corrected form (*rapidly*) would be an adverb? This is an especially frequent issue for languages with a richer morphology such as Portuguese or Czech.

To overcome the problem of inaccurate tagging results more quickly than manually processing all texts, annotation of the TH can be projected to the original tokens. For example, in the manually annotated part of CzeSL, the TH was tagged with lemmas and morphosyntactic description (MSD) tags by tools trained on regular Czech, and these annotations were projected to an intermediate representation of the orthographically normalized original, annotated with non-disambiguated MSDs and lemmas (Jelínek et al. 2012). Then, the projection results depend on the distance between the TH and the intermediate representation. For identical forms, both lemma and MSD are projected; for forms which are different but share a lemma, the TH lemma is projected, possibly disambiguating the choice of multiple lemmas. Otherwise, the non-disambiguated analysis is retained. In general, spell and grammar checkers can be used for an automatic approximation of normalization and error annotations, as an initial basis for further improvements. For example, the CzeSL corpus has this kind of fully automated MSD and lemma annotations for the original and for a TH (Rosen 2017).

2.4 Search and retrieval

Search and retrieval access via corpus query interfaces to a corpus is often the main goal of the corpus compilation process. It allows for analysing and exploring the corpus, which usually links back to the research questions the corpus was built for in the first place. In order to be usable for a large audience, it is very important to have an accessible interface to the corpus data. Such an interface usually provides a selection of the most relevant linguistic annotations - in order to get an overview - a task-specific selection or a list of all available annotations in the corpus. In addition, a list of searchable metadata that can be used as filters, with a pull-down menu if necessary, is useful, so that users do not need to know beforehand what data the corpus contains, but the interface makes this information available instead. Although too often overlooked in the past, most corpora in Table 1 now have (to varying degrees) a graphical user interface (GUI) for formulating search queries, where the output of the GUI is translated into the query language of the respective corpus tool. Finally, identifiers are often displayed with a gloss and not with the codified internal representation of the corpus. Figure 8 shows TEITOK's Query Builder in the CroLTeC corpus, which lists available search fields and possible metadata values.



Query Builder

Text Search

Written form	matches	
Teacher corrected form	matches	
Orthographically corrected form	matches	
Syntactically corrected form	matches	
Lexically corrected form	matches	
POS tag	matches	
Lemma	matches	

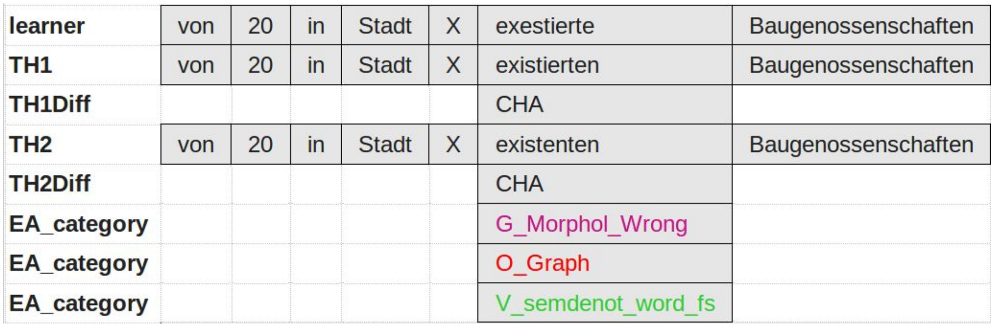
Document Search

Nationality	[select]
Mother tongue	[select]
Proficiency	[select]
Collection phase	[select]

Figure 8: The graphical Query Builder in CroLTeC.

To make access even easier for visitors who are only looking for examples, the interface sometimes also offers an option to search in a manner similar to common search engines. For example in TEITOK, any query that only contains simple characters is interpreted as a search-engine type search, meaning you can type in "house" to search for sentences containing "house" just like you would do for instance in Google. Furthermore, you can type in "hous*" to search for all words starting with "hous", like in the early days of web search engines.[26]

For advanced users, however, it is important to have direct access to the full expressiveness of the query language to be able to find even the most intricate examples in the corpus data. In this way, we were able to find the example from Figure 7, which displays multiple types of errors on the same token, in the MERLIN corpus; the actual data can be seen in Figure 9 below.



learner	von	20	in	Stadt	X	exestierte	Baugenossenschaften
TH1	von	20	in	Stadt	X	existierten	Baugenossenschaften
TH1Diff						CHA	
TH2	von	20	in	Stadt	X	existenten	Baugenossenschaften
TH2Diff						CHA	
EA_category						G_Morphol_Wrong	
EA_category						O_Graph	
EA_category						V_semdenot_word_fs	

Figure 9: ANNIS search result in MERLIN for Figure 7.

In the multi-layer representation of TEITOK, a query to find this example looks as follows, where *form* is the original form and *ort*, *gram*, *lex* are the orthographic, morphosyntactic and lexical TH, and the query checks that all three of them are different:

[26] https://www.livinginternet.com/w/wu_expert_wild.htm

```
[form != ort & ort != gram & gram != lex]
```

while the ANNIS query for this example looks like this, it searches for overlapping error codes (O: orthography, G: grammar, V: vocabulary) instead of searching the TH tokens directly:

```
EA_category=/O_.*/ & EA_category=/G_.*/ & EA_category=/V_.*/ & #1
_=_ #2 & #2 _=_ #3
```

To ensure a smooth learning curve, several corpus tools, including ASK, ANNIS, and TEITOK, provide a GUI that helps to create queries in a user-friendly way and also shows the actual query so that users can gradually become familiar with the query language itself.

The ability to query annotations helps users but for many the possibility to build on and extend the information already contained in the corpus would also be of great value. As Smith et al. (2008: 165) point out, even a richly annotated corpus is often not sufficient for a specific research question and hence additional (manual) annotation is often needed. In many corpora this is currently only possible to do locally for your own use. However, some new online interfaces such as ASK and TEITOK offer the possibility of adding further annotations online and saving them (see, e.g., Meurer 2012). In TEITOK, these added annotations even become searchable like any other annotation, but only for the original user.

2.5. Summary

To summarize, interoperability and comparability between learner corpus projects with regard to the dimensions of error taxonomies, metadata, annotation, and anonymization have proven to be rather challenging. Even when there is an initial aim to ensure comparability with other ongoing or past projects in many of these dimensions, this is hard to achieve. Nevertheless, awareness of the issues across the projects may influence approaches and workflows in newly started projects and help research teams – hopefully – to take an extra step towards a ‘best practice’ in learner corpus infrastructure projects.

3. Structural interoperability – technical perspective

Interoperability has been a focus of recent discussions for electronic linguistic resources ever since the number of available resources started growing (Ide and Pustejovsky 2010). It has also been gaining momentum with research in other fields, like computer mediated communication and social media (Beißwenger et al. 2017), and computational and corpus linguistics, where the special topic of the recent edition of the *Workshop on the Challenges in the Management of Large Corpora*[27] at LREC2018 was "Interoperability of corpus query and analysis systems." Although some state-of-the art approaches deal with both the *conceptual interoperability* (see Section 2) and *structural interoperability* (aiming to build the same formalism for annotations of different origin) of linguistic corpora in general as cited in Chiarcos (2012: 163), there are no standards or other clear guidelines on interoperability in the field of learner corpora. Moreover, there is no clear consensus on what exactly constitutes interoperability in learner corpora. A number of standardization initiatives within the domain of language learning have been created by the IMS Global Learning Consortium[28], such as Caliper Analytics[29] or Question and Test Interoperability (QTI)[30], to name just two. However, none specify the domain of learner corpora.

In an ideal world it should be possible to combine corpus data from different learner corpora, for example, in order to create a corpus of L1 speakers of German learning various L2 languages from a collection of otherwise independent learner corpora. A first requirement for this is that the corpus texts of the different corpora are stored in compatible formats. A second requirement is that the tools and software platforms can work together.

3.1. Document formats

When it comes to the interoperability of the corpus files themselves, which can all be joined in a single file or kept as separate documents, we should distinguish between four different types of data built on top of the actual text written by the student. Firstly, the annotation of the textual elements, mostly for transcribed

[27] http://corpora.ids-mannheim.de/cmlc-2018.html
[28] https://www.imsglobal.org/specifications.html
[29] https://www.imsglobal.org/activity/caliper
[30] https://www.imsglobal.org/question/index.html

hand-written essays: deleted words, added words, etc. Secondly, the linguistic annotation such as MSD, lemma, syntactic function and structure. Thirdly, the error annotation, including the target hypothesis. And finally, the metadata describing background information about the text, such as the L1 and L2 of the learner, the year in which it was written, the way it was collected, the prompts that were used etc.

Textual elements were typically represented in text-based formats in the past, either using a text editor, or symbols to indicate styled elements. Such formats are often hard to process computationally, and are hard to combine with other types of annotation. The more common way to annotate textual elements nowadays is by using a markup language, and the de-facto standard in learner corpora is the Text Encoding Initiative (TEI)[31] XML format. An example of the different formats is given in Figure 10, with POS tags from the universal dependencies tagset. Both the markup and the inline example are in the TEI XML format.

For linguistic annotation there are basically three different techniques: tabular, inline, or tier-based/stand-off. The *tabular format* is most frequently used in NLP applications, such as POS taggers and dependency parsers, or in linguistic search engines such as the IMS Open Corpus Workbench (CWB)[32], Corpuscle[33], or the SketchEngine[34].[35] But the tabular format is not very compatible with textual annotations.

markup	<p><s>A small <hi rend="italics">example</hi>.</s></p>
raw	A small example.
tabular	A DET small ADJ example NOUN . PUNCT
inline	<p><s><w pos="DET">A</w> <w pos="ADJ">small</w> <hi rend="italics"><w pos="NOUN">example</w></hi><c pos="PUNCT">.</c></s></p>
tier	

[31] http://www.tei-c.org

[32] http://cwb.sourceforge.net

[33] http://clarino.uib.no/korpuskel

[34] https://www.sketchengine.eu

[35] Although, these search engines use a more advanced format called vertical file format (vrt), which can encode some annotations beyond the token level.

	DET	ADJ	NOUN	PUNCT
	A	small	example	.
	A small example.			

Figure 10: The different file formats exemplified.

The *inline annotation* is most compatible with the textual annotation, since it can be done directly in markup files such as the TEI XML files (as is done in “inline” in Figure 10). And the inline format is easy to export to a tabular format, making it relatively easy to use in existing NLP tools. But there are two problems with inline annotation. The first is that the XML files with various types of annotations inside become so large that it is virtually impossible to use them without a user interface (which is what, for instance, TEITOK attempts to provide). And the second is that the linguistic annotation cannot intersect with the textual annotation. For token-level linguistic annotation, such as POS and lemma, this is not a problem. But when it comes to error annotation, this becomes a serious problem – error annotations can even intersect with each other. Which is why multi-token error annotation in TEI XML almost unavoidably has to be done in a *stand-off manner*, which is to say, not added to the XML file directly, but kept in a separate file which links to the XML file.

Tier-based annotations and other types of stand-off annotations do not have the problem of overlapping annotations, since there is no restriction on overlapping annotations – the tiers define regions over a timeline, which can either be a real timeline, or a sequence of tokens. But although it is possible to keep the textual annotation in a tier-based format, it has to be explicitly added, and in learner corpora it is frequently discarded. Written (learner) corpora typically use token-based tiers with formats such as EXMARaLDA XML (Schmidt 2010) or PAULA (Zeldes et al. 2013). And in a typical token-based tier setup, it is necessary to keep track of whitespace explicitly, since a sequence of tokens does not indicate by itself whether there are spaces between the tokens or not.

Due to the fundamental differences between the two formats (with TEI and PAULA as typical examples), it is often impossible to convert between them without losing information, even though problems are typically reduced to the treatment of spaces and the error annotation. But conversion between different inline formats or between different tier-based formats is much easier than between inline and tier or vice versa.

The choice of format does not imply anything about the actual tags or annotation scheme used for linguistic or error annotation. But the format does have implications for the way the corpus can be searched: the most commonly used query language for linguistics is the CWB Query Language (CQL), which is not very well-suited for tier-based formats – for using the advantages of the format in the search, a tier architecture needs a more complex search language, which is supported by only a few corpus query languages such as the ANNIS query language[36] and the Prague Markup Language Tree Query (PML-TQ) language[37], whereas there are many different tools for CQL, including CWB, Corpuscle, BlackLab[38], and SketchEngine.

For metadata, the advantage of TEI XML over other formats is that it has a (large) pre-defined set of metadata, specifying in detail which elements can occur in the metadata header and what those elements mean. A tabular format on the other hand typically has to rely on external spreadsheets, and PAULA uses user-defined metadata fields. However, a problem for TEI XML as a standard for learner corpora is that there are various types of metadata that are crucial for learner corpora, but not contemplated in the current TEI standard. To overcome this, COPLE2 introduced two new sections in the teiHeader that have since been adopted by other TEITOK learner corpora (e.g., those in Table 1). A drawback is that decisions about the best solutions for these sections ideally need to be made before documents are added to the corpus. The advantage of several projects using the same system is that they can more easily share their solution, for example, for how to handle additional metadata.

3.2. Tools and platforms

As with annotation schemes, learner corpus developers are faced with decisions about whether to reuse existing annotation and search tools or to develop new ones customized for their research. The reuse of existing tools clearly leads to a degree of interoperability with previous corpora developed with the same tools, but the existing tools may not support all of the desired annotation and search capabilities for a new project, so many learner corpus projects, at least those with time and resources, put significant effort into software development.

While there are many general-purpose linguistic annotation tools and search engines that support most of the desired annotation for learner corpora presented in

[36] http://corpus-tools.org/annis/aql.html
[37] https://ufal.mff.cuni.cz/pmltq/
[38] https://github.com/INL/BlackLab

Section 2.3, target hypotheses and error annotation present a particular challenge. This is evident in the number of learner corpus projects with custom annotation tools, including Falko (Falko Excel add-in[39]), CzeSL (feat[40]), and SweLL (see Section 2.3.1, Figure 4). When developing new annotation tools and formats, it is important to keep in mind that any new annotation format will additionally require support within a search engine for querying and visualization for the annotation to be accessible to most users. The following sections discuss some of the issues in reusing existing tools (MERLIN) versus developing new tools (CzeSL and SweLL) and the final section describes an extensible software platform for both annotation and search (TEITOK, used for COPLE2 and CroLTeC).

3.2.1. Reusing existing tools: MERLIN

The MERLIN project chose to reuse existing corpus annotation and search tools as much as possible, in part due to time and budget constraints. As no single tool supported all of the annotation requirements, a combination of tools was required to support the wide range of manual and automatic annotation that had been designed to illustrate the CEFR scales (Boyd et al. 2014).

First, the transcription process of handwritten documents relied on an XML-based approach developed within the KoKo project: a strictly validating XML editor was used to create documents conforming to a custom schema. Next, the manual annotation in MERLIN (MERLIN 2014), which includes error annotation and linguistic characteristics of the learner language, was performed using the Falko add-in for Microsoft Excel and the MMAX2 multi-level annotation tool (Müller & Strube 2006). The corpus was first annotated with explicit target hypotheses in Excel using annotation guidelines adapted from the Falko project (Reznicek et al. 2012) and then the MERLIN annotation scheme (Wisniewski et al. 2014) was applied in MMAX2. Parallel to the manual annotation, a custom UIMA[41] processing pipeline, which adds additional layers of linguistic annotation from taggers, parsers, etc., was developed for MERLIN's automatic annotation.

Through cooperation with the established Falko project (Reznicek et al. 2012), MERLIN was able to benefit from Falko's existing infrastructure for target hypothesis annotation, corpus data conversion with SaltNPepper (Zipser et al. 2011), and search and visualization with the search engine ANNIS (Krause & Zeldes 2016). In particular, the corpus data conversion tool SaltNPepper was

[39] https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/marc/

[40] https://bitbucket.org/czesl/feat

[41] http://uima.apache.org

crucial to MERLIN’s use of existing annotation tools for each annotation task. It was used to manipulate data in EXMARaLDA, PAULA, MMAX2, and ANNIS format.

However, when reusing existing tools, learner corpus developers need to remain aware of differing needs that have influenced a tool’s features and capabilities, such as the intended corpus users, for example, corpus linguists versus language teachers. For instance, the Falko Excel target hypothesis annotation tool is not able to preserve whitespace and formatting in the learner text (see Section 3.1), while MERLIN needed to preserve the formatting for presentation to users who may not be accustomed to inspecting unformatted texts with whitespace tokenization. As a result, preserving the original text formatting alongside the target hypothesis annotation necessitated the development of additional conversion tools beyond the existing Falko infrastructure.

The contrast between formatted and unformatted texts is demonstrated for a letter from MERLIN in Figure 11, where whitespace-based formatting is shown in comparison to unformatted, tokenized text. For displaying a formatted text to users searching within MERLIN, the whitespace-based formatting is sufficient, but for more flexible visualization options and further linguistic processing, a more structured representation of the document format such as TEI XML would be helpful in order to meaningfully distinguish sections such as a date in a header from other sections of a text.

original formatting	Lieber Michael. 24.6.2012 wie geht´s dir!? Am Montag Fahre ich mit meinem Vater nach Stadt X.
unformatted, tokenized	Lieber Michael . 24.6.2012 wie geht´s dir ! ? Am Montag Fahre ich mit meinem Vater nach Stadt X .

Figure 11: MERLIN Text with and without Formatting.

Since the reuse of existing annotation tools requires conversions between the formats supported by those tools, an ideal format conversion tool should provide detailed documentation about which aspects of the annotation are preserved in a conversion and clear warnings when conversions are not able to preserve information from the input documents.

3.2.2. Building new tools: CzeSL and SweLL

Both CzeSL and SweLL have developed new tools for target hypothesis and error annotation in the multi-layer parallel format shown in Figure 6. The SweLL tool, which stores its annotation in a custom format (Rosén et al. 2018), includes one level of normalization while CzeSL includes two levels (see Section 2.3.2, Hana et al. 2010) stored in the generic Prague Markup Language format (Hana & Štěpánek 2012).

By using custom tools and formats, projects can support an annotation approach that is ideally suited to their research questions and data, but their use may become problematic later if the consequences for the annotation process and the potential uses of the corpus are not foreseen. In particular, visualization of the multi-layer parallel normalizations in search engines is a problem for learner corpora with this type of annotation. The authors are not currently aware of a corpus search engine that can visualize search results in this format. Although *SeLaQ*[42] and ANNIS support the relevant corpus annotation queries, they are not able to display the results in a user-friendly manner. In addition, as a relatively new tool, *SeLaQ* does not include all the features available in more mature corpus search tools. Possible alternatives such as standard concordancers would require substantial modifications of the data and as a consequence only a subset of the annotation could be queried.

3.2.3. Building a platform: TEITOK

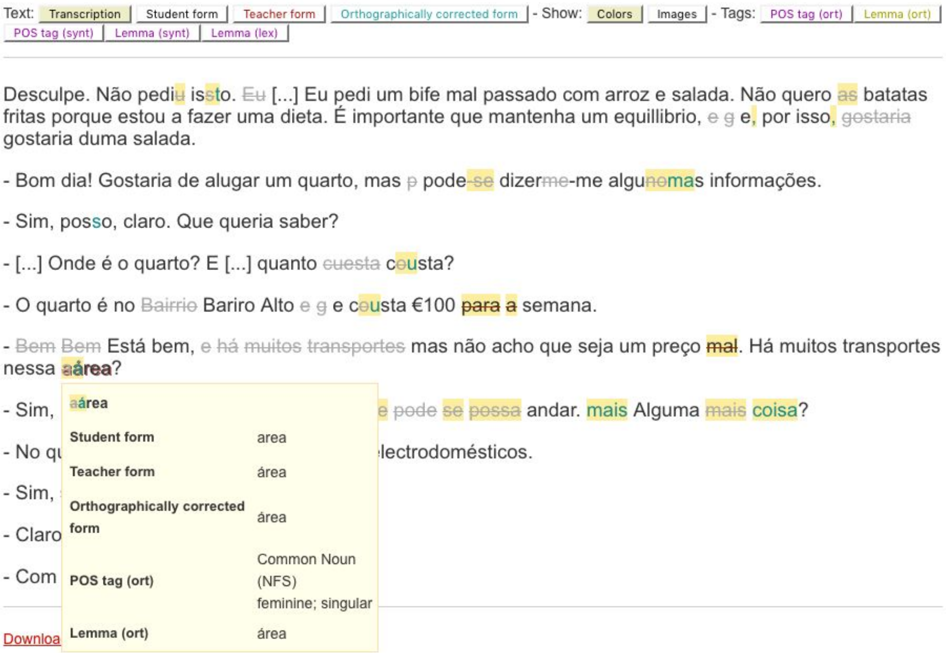
The traditional setup for building corpora is a pipeline: A series of steps is applied to the original texts, with the output of each step being the input for the next. And the steps themselves rely on computational linguists for their application. However, more and more projects are attempting to provide an (online) interface for this type of processing in order to make it easier for linguists to perform the steps autonomously and apply them reproducibly to new texts added to the corpus. Examples of such online interfaces can be found in WebLicht[43], FoLiA[44], and TEITOK[45].

TEITOK (Janssen, 2016) attempts to provide a graphical user interface for a chain of tools (either existing or specifically developed) that runs behind the scenes

[42] The query language developed for the CzeSL parallel format, https://bitbucket.org/czesl/selaq
[43] https://weblicht.sfs.uni-tuebingen.de/
[44] https://proycon.github.io/folia/
[45] http://www.teitok.org/

without the user creating the corpus having to worry about the technical details. This is done in a modular fashion, with all tools interacting directly with TEI XML files. A screenshot of the main text view of a XML document is shown in Figure 12. A platform like TEITOK then ultimately is a complete infrastructure, with the underlying format(s) it is based on, the interface to collect, edit, annotate, distribute, and process data in electronic form on the one hand and an active user community and developer support on the other.

The modular design makes it possible to combine many different options in the same interface: the (administrative) users can correct errors in existing XML files directly from the interface using HTML forms. They can use the interface to directly build the XML files inside the system in a number of ways, most relevantly for learner corpora by either transcribing directly from a sound file using a time-aligned architecture, or transcribing a handwritten text line-by-line from the facsimile images. There is a custom tool to add stand-off error annotation with a target hypothesis; there are scripts to run POS taggers and dependency parsers that can be called from within the system using buttons, and their output is integrated into the XML files. Furthermore, a searchable CWB corpus is built directly from the XML files, which can be searched via the interface as well. And it can combine different sub-corpora with different characteristics and data setup in a single corpus while still being searchable and editable in much the same way. This is particularly relevant for learner corpora, which can have both a written and a spoken part.



Desculpe. Não pedi~~u~~ is~~s~~to. ~~Eu~~ [...] Eu pedi um bife mal passado com arroz e salada. Não quero ~~as~~ batatas fritas porque estou a fazer uma dieta. É importante que mantenha um equillibrio, ~~e g~~ e, por isso, ~~gostaria~~ gostaria duma salada.

- Bom dia! Gostaria de alugar um quarto, mas ~~p~~ pode~~-se~~ dizer~~me~~-me algu~~no~~mas informações.

- Sim, posso, claro. Que queria saber?

- [...] Onde é o quarto? E [...] quanto ~~cuesta~~ c~~o~~usta?

- O quarto é no ~~Bairrio~~ Bariro Alto ~~e g~~ e c~~o~~usta €100 ~~para~~ ~~a~~ semana.

- ~~Bem Bem~~ Está bem, ~~e há muitos transportes~~ mas não acho que seja um preço ~~mal~~. Há muitos transportes nessa ~~a~~área?

- Sim, e ~~pode~~ ~~se~~ ~~possa~~ andar. mais Alguma ~~mais~~ coisa?

- No qu lectrodomésticos.

- Sim,

- Claro

- Com

~~a~~área

Student form area

Teacher form área

Orthographically corrected form área

POS tag (ort) Common Noun (NFS) feminine; singular

Lemma (ort) área

Figure 12: The TEITOK view of an XML document from COPLE2. The yellow highlighted items are teacher corrections, the popup shows several layers of annotation.

For any platform to be usable beyond the scope of the project it was initially built for, it is vitally important that it well documented, and that it can be customized, since different corpora often have different requirements. The setup of TEITOK is very customizable in this sense, since from the start, it was built with not only learner corpora in mind, but a variety of other types of corpora as well, including historical corpora. Therefore, TEITOK can be used for basically any learner corpus, as long as it uses the TEI XML file format.

Of course a platform such as TEITOK, even with customizability, is just a tool that does certain things and not others. But an important idea behind a modular system is that it is not necessary to build a new one if the existing one does not do everything that is needed. If elements are missing, the only thing that is necessary for a new project is to provide new modules for the missing parts or replace existing modules (as long as they support the underlying format). For linguists, this has proven to be a very beneficial setup.

3.3. Summary

Our experience shows that a compatible, stable toolchain should be selected as early as possible in the project and that corpus developers should also consider early on how the annotations will be queried and visualized in a user-friendly search interface so that the data is accessible to researchers with a range of backgrounds. It has proven advantageous to store annotations in an established data format in order to be able to use or adapt existing tools, especially for needs that may not be foreseeable at the beginning of a project. An ideal learner corpus framework would also include tools needed by any manual annotation project such as components for annotation task management and the ongoing evaluation of inter-annotator agreement. And in an ideal world, there would be as few projects as possible building frameworks that would be used by all (new) learner corpora to be gradually improved on the basis of user experiences – although, in practice, having a certain amount of competition amongst as many projects as necessary to drive improvement forward is desirable. For that to work, the frameworks would have to count on the support of the community, be continuously maintained to account for (new) bugs and user feedback, and ideally allow users to add functionalities still missing from the framework; the software should adhere to the principles of free/libre and open-source software: "users [should] have the freedom to run, copy, distribute, study, change and improve the software" (Free Software Foundation 2017).

4. Conclusion and outlook

In this article we gave an overview of first-hand experiences and starting points for best practices from projects in seven European countries dedicated to learner corpus research and the creation of language learner corpora. We saw that we are part of a thriving community with research and commercial interests where the empirically-based method has become established, and work with corpora in particular has been widely adopted. Since the underlying corpora are becoming more numerous and extensive, the corpora and the involved tools are of great value to the scientists who collect and create them, but also to the research community as a whole, because as Wilkinson et al. (2016) point out:

> digital assets … should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are

> high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies.

To this end, they propose four basic principles for scientific data management that we should adhere to: *findability*, *accessibility*, *interoperability*, and *reusability* (FAIR). We believe, the advancement of LCR would thrive on the timely sharing and accessibility of data and tools, and we should do our best to further increase our efforts in their interoperability and reusability.

But what are the stumbling blocks on the way to interoperability between learner corpora, and why can we not combine corpus data from different learner corpus projects? We showed that seemingly inconspicuous concepts like metadata, anonymization, error taxonomies, and linguistic annotations already rocked a few projects, and adapting toolchains or data formats late in a project to meet previously unknown or changed demands mean a few more sticks between the legs.

Metadata affect the applicability in LCR because as Burnard (2005) has pointed out "it is no exaggeration to say that without metadata, corpus linguistics would be virtually impossible", and error taxonomies with their error tags are specifically designed "to cater for the anomalous nature of learner language" (Granger 2002: 18). But with a lack of agreed solutions for many aspects of LCR even minor differences in project goals, research interests, national legislation, ethics board standards, technical expertise, or available resources may entail changes to metadata, error taxonomies, or data formats that render corpus data un-interoperable. It also puts a spot onto an issue when partners within one project are struggling – as we reported – to come to an agreement about what an 'error' is.

Learner corpus infrastructure as such is a new concept, and many groups are coming up with ideas and solutions, so we are right now in a testing period, with a multitude of approaches, workflows, and tools – and it probably will remain like this for a while. So, if an approach seems the best one for a group but no corresponding guidelines or tools exist this does not prove the direction to be wrong. It might just be new. We have all experienced bumps along the road in working with learner corpora, and to distinguish new directions from beaten tracks we strongly believe that a lot is to be gained by communicating more with each other between projects and research areas.

In order to tackle all the challenges we have addressed in this paper, and to integrate, extend and harmonise efforts, we see an EU infrastructure project[46] as an outstanding goal for the future of the LCR community. To achieve this goal, a

[46] http://ec.europa.eu/research/infrastructures/index.cfm?pg=about#

COST Action[47] could be very helpful; it is an ideal instrument to bring together stakeholders with common interests and also research areas into a new, committed community. Additionally, a CLARIN K(nowledge)-Centre for LCR could formalize and centrally register already existing expertise.

Acknowledgments

The initial starting point for work on this paper was the *Workshop on Interoperability of Second Language Resources and Tools* at the University of Gothenburg, Sweden in December, 2017. We would like to thank CLARIN, Swe-Clarin, and CLARINO for their financial support. The work has also been partially supported by the Grant Agency of the Czech Republic (grant no. 16-10185S) and SweLL, an infrastructure grant from the Swedish Riksbankens Jubileumsfond (project IN16-0464:1).

In addition, we would like to thank both reviewers for their thoughtful and thorough comments and their shared enthusiasm for the topic.

References

Accenture. (2016). Building digital trust: The role of data ethics in the digital age. https://www.accenture.com/t20160613T024441__w__/us-en/_acnmedia/PDF-22/Accenture-Data-Ethics-POV-WEB.pdf (last accessed on 20 November, 2018).

Ballier, N., & Martin, P. (2013). Developing corpus interoperability for phonetic investigation of learner corpora. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (eds.) *Automatic Treatment and Analysis of Learner Corpus Data (Studies in Corpus Linguistics)* 59. Amsterdam: John Benjamins Publishing Company, 33–64.

Bayerl, P. S., & Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, 37(4), 699–725. https://doi.org/10.1162/COLI_a_00074

Beißwenger, M., Wigham, C. R., Etienne, C., Fišer, D., Suárez, H. G., Herzberg, L., Hinrichs, E., Horsmann, T., Karlova-Bourbonus, N., Lemnitzer, L., Longhi, J., Lüngen, H., Ho-Dac, L.-M., Parisse, C., Poudat, C., Schmidt, T., Stemle, E., Storrer, A. & Zesch, T. (2017). Connecting resources: Which issues have to be solved to integrate CMC corpora from heterogeneous sources and for different languages? In E. W. Stemle & C. Wigham (eds.) *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities*, Bolzano, 3-4 October, 2017, 52–55. https://doi.org/10.5281/zenodo.1041877

[47] https://www.cost.eu/cost-actions/what-are-cost-actions/

Bolt, P. (1992). An evaluation of grammar-checking programs as self-helping learning aids for learners of English as a foreign language. *Computer Assisted Language Learning (CALL)*, 5(1–2), 49–91. https://doi.org/10.1080/0958822920050106

Borin, L., Forsberg, M. & Roxendal, J. (2012). Korp – The corpus infrastructure of Språkbanken. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, 21-27 May, 2012, 474–478.

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B. & Vettori, C. (2014). The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, 26-31 May, 2014, 1281–1288.

Brezina, V., & Flowerdew, L. (2018). *Learner Corpus Research: New Perspectives and Applications*. London: Bloomsbury Academic. https://doi.org/10.5040/9781474272919

Burnard, L. (2005). Metadata for corpus work. In M. Wynne (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, 30–46. http://ota.ox.ac.uk/documents/creating/dlc/ (last accessed on 20 November, 2018).

Caines, A., & Buttery, P. (2018). The effect of task and topic on opportunity of use in learner corpora. In V. Brezina & L. Flowerdew (eds.) *Learner Corpus Research: New Perspectives and Applications*. London: Bloomsbury, 5–27.

CEFR (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Council of Europe, Modern Languages Division. Strasbourg & Cambridge: Cambridge University Press. https://rm.coe.int/1680459f97 (last accessed on 20 November, 2018).

Chiarcos C. (2012). Interoperability of corpora and annotations. In C. Chiarcos, S. Nordhoff S. & S. Hellmann (eds.) *Linked Data in Linguistics*. Berlin, Heidelberg: Springer, 161–179. https://doi.org/10.1007/978-3-642-28249-2_16

Díaz-Negrillo, A., Meurers, D., Valera, S., & Wunsch, H. (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36(1–2), 139–154.

Dobric, N. (2015). Quality measurements of error annotation - Ensuring validity through reliability. *The European English Messenger*, 24(1), 36–42.

El Emam, K. & Arbuckle, L. (2013). *Anonymizing Health Data: Case Studies and Methods to Get You Started*. Newton MA: O'Reilly Media, Inc.

Free Software Foundation (2017). What is free software? The Free Software Definition (free-sw). https://www.gnu.org/philosophy/free-sw.html (last accessed on 20 November, 2018).

Fort, K. (2016). *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. London: ISTE.

Fort, K., Nazarenko, A., & Rosset, S. (2012). Modeling the complexity of manual annotation tasks: a grid of analysis. In *Proceedings of COLING 2012: Technical Papers*, Mumbaï, 8-15 December, 2012, 895–910.

Gaillat, T., Sébillot, P., & Ballier, N. (2014). Automated classification of unexpected uses of this and that in a learner corpus of English. In L. Vandelanotte, K. Davidse, & C. Gentens (eds.) *Recent Advances in Corpus Linguistics: Developing and Exploiting Corpora* 78, 309–324. http://doi.org/10.1163/9789401211130_015

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (eds.) *The Cambridge Handbook of Learner Corpus Research* (Cambridge Handbooks in Language and Linguistics). Cambridge: Cambridge University Press, 9–34. https://doi.org/10.1017/CBO9781139649414.002

Glaznieks, A., Abel, A., Lyding, V., Nicolas, L. & Stemle, E. (2014). Establishing a standardised procedure for building learner corpora. In T. Nikula, S. Takala & S. Ylönen (eds.) *Apples - Journal of Applied Language Studies* 8(3). Jyväskylä: Centre for Applied Language Studies, University of Jyväskylä, 5–20.

Golcher, F., & Reznicek, M. (2011). Stylometry and the interplay of topic and L1 in the different annotation layers in the FALKO corpus. In *Proceedings of Quantitative Investigations in Theoretical Linguistics (QITL-4)*, Berlin, 29-31 March, 2011, 29–34. https://doi.org/10.18452/1370

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins, 3–33.

Granger, S. (2008). Learner corpora in foreign language education. In N. Van Deusen-Scholl & N. H. Hornberger (eds.) *Encyclopedia of Language and Education. Volume 4. Second and Foreign Language Education*. New York: Springer, 337–351.

Granger, S. & Meunier, F. (1994). Towards a grammar checker for learners of English. In U. Fries & G. Tottie (eds.) *Creating and Using English Language Corpora*. Amsterdam and Atlanta: Rodopi, 79–89. http://hdl.handle.net/2078/75631

Granger, S. & Paqout, M. (2017). Towards standardization of metadata for L2 corpora. https://sweclarin.se/swe/workshop-interoperability-l2-resources-and-tools (last accessed on 20 November, 2018).

Hana, J., Rosen, A., Škodová, S. & Štindlová, B. (2010). Error-tagged learner corpus of Czech. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV)*. Uppsala, 15-16 July, 2010, 11–19.

Hana, J. & Štěpánek, J. (2012). Prague markup language framework. In *Proceedings of the Sixth Linguistic Annotation Workshop (LAW VI)*, Jeju, 12-13 July, 2012, 12–21.

Hinkel, E. (2002). *Second Language Writer's Text: Linguistic and Rhetorical Features*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Ide, N., & Pustejovsky, J. (2010). What does interoperability mean, anyway? Toward an operational definition of interoperability for language technology. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*, Hong Kong, 18-20 January, 2010, 1–8.

Janssen, M. (2016). TEITOK: Text faithful corpora. In N. Calzolari, et al. (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, 23-28 May, 2016, 4037–4043.

Jelínek, T., Štindlová, B., Rosen, A., & Hana, J. (2012). Combining manual and automatic annotation of a learner corpus. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (eds.) *Text, Speech and Dialogue. TSD 2012 (Lecture Notes in Computer Science)* 7499. Berlin, Heidelberg: Springer Berlin Heidelberg, 127–134.

Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J., & Teich, E. (2016). The Royal Society corpus: from uncharted data to corpus. In N. Calzolari, et al. (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, 23-28 May, 2016, 1928–1931.

Krause, T., & Zeldes, A. (2016). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities* 31(1), 118–139. https://doi.org/10.1093/llc/fqu057

Lüdeling, A., Walter, M., Kroymann, E., & Adolphs, P. (2005). Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics 2005*, Birmingham, 14-17 July, 2005, 1–11.

Martin, M. (2011). Connecting functional proficiency levels with linguistic development: Projects Cefling and Topling. Presentation at Tallinn University, 2 June, 2011.

McEnery, T. (2018). Preface. In V. Brezina & L. Flowerdew (eds.) *Learner Corpus Research: New Perspectives and Applications*. London: Bloomsbury Academic, xiv–xvii. https://www.bloomsburycollections.com/book/learner-corpus-research-new-perspectives-and-applications/preface-tony-mcenery (last accessed on 14 January, 2019).

Megyesi, B., Johansson, S., Rosén, D., Schenström, C.-J., Sundberg, G., Wirén, M., & Volodina, E. (2018). Learner corpus anonymization in the age of GDPR: Insights from the creation of a learner corpus of Swedish. In *Proceedings of the 7th NLP4CALL* workshop, Stockholm, 7 November, 2018, 47–56.

Mendes, A., Antunes, S., Janssen, M. & Gonçalves, A. (2016). The COPLE2 corpus: A learner corpus for Portuguese. In N. Calzolari, et al. (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, 23-28 May, 2016, 3207–3214.

MERLIN project (2014): Annotation guidelines. http://www.merlin-platform.eu (last accessed on 20 November, 2018).

Meurer, P. (2012). Corpuscle – a new corpus management platform for annotated corpora. In G. Andersen (ed.) *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian (Studies in Corpus Linguistics)* 49. Amsterdam: John Benjamins Publishing Company, 31–49.

Meurers, D. (2015). Learner corpora and natural language processing. In S. Granger, G. Gilquin, & F. Meunier (eds.) *The Cambridge Handbook of Learner Corpus Research (Cambridge Handbooks in Language and Linguistics)*. Cambridge: Cambridge University Press, 537–566. https://doi.org/10.1017/CBO9781139649414.024

Müller, C., & Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In S. Braun, K. Kohn, & J. Mukherjee (eds.) *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods* 3, 197–214.

Rebuschat, P. , Meurers, D. & McEnery, T. (2017). Language learning research at the intersection of experimental, computational, and corpus-based approaches. *Language Learning*, 67(S1), 6–13. https://doi.org/10.1111/lang.12243

Reznicek, M., Lüdeling, A., Krummes, C. & Schwantuschke, F. (2012). Das Falko-Handbuch: Korpusaufbau und Annotationen. Version 2.0. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuchv2.0.pdf (last accessed on 20 November, 2018).

Riksdagen. (1949). Tryckfrihetsförordningen (1949:105). http://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/tryckfrihetsforordning-1949105_sfs-1949-105 (last accessed on 20 November, 2018).

Río, I. del, Antunes, S., Mendes, A., & Janssen, M. (2016). Towards error annotation in a learner corpus of Portuguese. In E. Volodina, et al. (eds.) *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC*, Umeå, 16th November, 2016, 8–17.

Rosen, A. (2015). CzeSL-MAN – a corpus of non-native speakers' Czech with manual annotation. http://utkl.ff.cuni.cz/~rosen/public/2015-czesl-man-en.pdf (last accessed on 14 January, 2019)

Rosen, A. (2017). Introducing a corpus of non-native Czech with automatic annotation. In P. Pęzik & J. T. Waliński (eds.) *Language, Corpora and Cognition*. Bern & Warszawa: Peter Lang, 163–180.

Rosén, D., Wirén, M. & Volodina, E. (2018). Error annotation of second language learner texts based on mostly automatic alignment of parallel corpora. In *Proceedings of the CLARIN Annual Conference 2018*, Pisa, 8-10 October, 2018, 181–184.

Schmidt, T. (2010). EXMARaLDA: un système pour la constitution et l'exploitation de corpus oraux. In H. Boyer (ed.) *Pour une épistémologie de la sociolinguistique. Actes du colloque international de Montpellier 10-12 décembre 2009*, Limoges: Lambert-Lucas, 319–327.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 10(1–4), 209–232. https://doi.org/10.1515/iral.1972.10.1-4.209

Smith, N., Hoffmann, S., & Rayson, P. (2008). Corpus tools and methods, today and tomorrow: Incorporating linguists' manual annotations. *Literary and Linguistic Computing*, 23(2), 163–180. https://doi.org/10.1093/llc/fqn004

Tenfjord, K., Meurer, P., & Hofland, K. (2006). The ASK corpus: A language learner corpus of Norwegian as a second language. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, 24-26 May, 2006, 1821–1824.

Volodina, E., Granstedt, L., Megyesi, B., Prentice, J., Rosén, D., Schenström, C.-J., Sundberg, G. & Wirén, M. (2018). Annotation of learner corpora: first SweLL insights. In *Proceedings of the Seventh Swedish Language Technology Conference (SLTC-2018)*, Stockholm, 7-9 November, 2018.

Volodina, E. & Megyesi, B. (2018). SweLL PUL - data flow handling. Project documentation. https://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/SweLL%20-%20data%20flow%20handling_Aug18_0.pdf (last accessed on 20 November, 2018).

Volodina, E., Megyesi, B., Wirén, M., Granstedt, L., Prentice, J., Reichenberg, M., & Sundberg, G. (2016). A friend in need? Research agenda for electronic second language infrastructure. In *Proceedings of the Sixth Swedish Language Technology Conference (SLTC-2016)*, Umeå, 17-18 November, 2016, 1–4.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016). Comment: The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, Article number: 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Wisniewski, K., Woldt, C., Schöne, K., Abel, A., Blaschitz, V., Štindlová, B., & Vodičková, K. (2014). The MERLIN annotation scheme for the annotation of German, Italian, and Czech learner language. http://www.merlin-platform.eu (last accessed on 20 November, 2018).

Zaghouani, W., Habash, N., Bouamor, H., Rozovskaya, A., Mohit, B., Heider, A., & Oflazer, K. (2015). Correction annotation for non-native Arabic texts: Guidelines and corpus. In *Proceedings of The 9th Linguistic Annotation Workshop (LAW IX)*, Denver CO, 5 June, 2015, 129–139. https://doi.org/10.3115/v1/W15-1614

Zeldes, A., Zipser, F. & Neumann, A. (2013). PAULA XML documentation. Format version 1.1. https://hal.inria.fr/hal-00783716

Zipser, F., Zeldes, A., Ritz, J., Romary, L., & Leser, U. (2011). Pepper: Handling a multiverse of formats. https://doi.org/10.5281/zenodo.15638