Design and Development of the MERLIN[1] Learner Corpus Platform

Verena Lyding[†], Karin Schöne[*]
[†]Institute for Specialised Communication and Multilingualism, EURAC research, Bozen/Bolzano, Italy;
[*]Multimediales Sprachlernzentrum, Technische Universität Dresden, Germany
E-mail: verena.lyding@eurac.edu, karin.schoene@mailbox.tu-dresden.de

Abstract

In this paper we report on the design and development of an online search platform for the MERLIN[1] corpus of learner texts in Czech, German and Italian. It was created in the context of the MERLIN project, which aims at empirically illustrating features of the Common European Framework of Reference (CEFR) for evaluating language competences based on authentic learner text productions compiled into a learner corpus. Furthermore, the project aims at providing access to the corpus through a search interface adapted to the needs of multifaceted target groups involved with language learning and teaching.
This article starts by providing a brief overview on the project ambition, the data resource and its intended target groups. Subsequently, the main focus of the article is on the design and development process of the platform, which is carried out in a user-centred fashion. The paper presents the user studies carried out to collect requirements, details the resulting decisions concerning the platform design and its implementation, and reports on the evaluation of the platform prototype and final adjustments.

Keywords: learner corpora, corpus search tools, user-centred development

1. Introduction

This article describes the design and development process of a search platform for a trilingual learner corpus with multifaceted target groups. The MERLIN project addresses the need for illustrating the CEFR levels of language proficiency with concrete and authentic examples of language use and tackles two related research questions. On a theoretical level, that is with regard to linguistics and language didactics, it researches on the best ways to pinpoint and describe relevant characteristics of learner language in relation to CEFR evaluation dimensions. On a practical level, that is with regard to system development, it researches about how to encode, make accessible and present this information to a varied set of target groups. This paper describes how the practical research question has been approached. It details the user-centered design and development process and discusses the decisions taken as well as the potential and limits of the developed system.

2. MERLIN Project

The overall objective of the MERLIN[1] project is to address the need for an empirical back-up of the CEFR levels by providing concrete examples of learner language features. Within an interdisciplinary trans-European project team three major tasks were approached within MERLIN:

1. to assemble a trilingual learner corpus,
2. to carefully evaluate relevant parameters for describing learner language and to annotate related features on the learner texts, and
3. to develop a search platform for providing the texts and means for their analysis to a diversified group of users.

2.1 MERLIN Corpus

The MERLIN corpus consists of written productions of foreign language learners of Czech, German and Italian, which are annotated for features of learner language and characteristics of the learners.
The MERLIN corpus is a comprehensive collection of 2,286 authentic foreign language learner texts in Czech, German and Italian, produced in standardized language tests and collected by the established test institutions telc[2] and ÚJOP[3]. The corpus is annotated for learner language features on several linguistic levels, including orthography, grammar, coherence/cohesion etc. (see Abel et al. 2014, Boyd et al. 2014). In addition, personal characteristics of the tested learners, including their L1, age, gender, etc., as well as meta information on the texts, including underlying test task, CEFR level of the test, ratings according to the CEFR[4], are recorded and associated with each text.

2.2 Target Groups

The MERLIN project targets professionals involved with analyzing learner language, which were grouped into four profiles: teachers (incl. material writers), teacher trainers, testers and linguists (including second language acquisition researchers and lexicographers). The target groups differ with regard to how their work relates to the CEFR, what perspectives they take on the data and how familiar they are with corpus interfaces.

3. Requirements Analysis for the MERLIN Platform

In order to inform the design of the MERLIN platform, a study to determine specific requirements of the different target groups was carried out at the start of the project. By means of an online questionnaire we investigated the users'

[1] *Multilingual Platform for the European Reference Levels – Exploring Interlanguage in Context*, www.merlin-platform.eu
[2] http://www.telc.net/
[3] http://ujop.cuni.cz/en/exams/czech-language-for-foreigners
[4] Fair average holistic rating

needs regarding content (i.e. relevant linguistic annotations, metadata, and quantitative text characteristics) as well as interface aspects, including search and display functionalities as well as technical features.
Overall, 55 people from all target groups and covering the three working languages participated in the survey. Most participants indicated to belong to more than one target group[5], and the overall distribution of participants is shown in Figure 1.

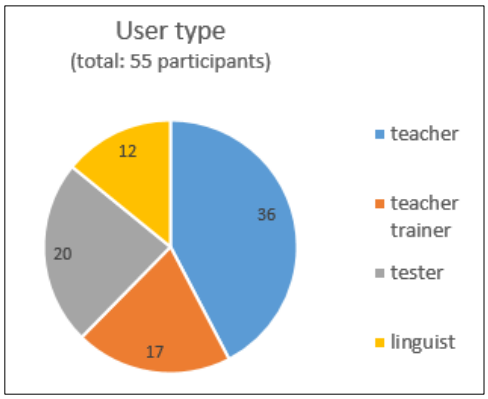

Figure 1: Professions of participants

The requirement analysis revealed their specific information needs and possible usage scenarios. Teachers and teacher trainers expressed the need for illustration of the CEFR descriptors with examples from learner texts. They would appreciate the option to extract sample productions to use them for training purposes and for the preparation of teaching materials, e.g. for self-reflection activities with advanced learner groups. Testers would use a corpus of standardized test samples for training purposes (e.g. have samples from the corpus re-rated and compared with the MERLIN rating) as well as a reference for the assessment of borderline performances. Linguists and SLA researchers would benefit from a thoroughly annotated corpus of learner language to explore L2-competence and to trace errors or features of learner language on different performance levels.
Results indicated that the majority of users considers it important to have search and filtering of learner texts based on different learner language features and metadata. The linguistic features vocabulary (87%) and grammar (76%) were rated most relevant, followed by text characteristics (67%) and sociolinguistic criteria/text type (67%). With respect to metadata, level (87%) and type (80%) of the test tasks as well as quantitative text characteristics like the size of vocabulary (82%) and sentence complexity (76%) were assigned the highest relevance. Compared by user group, metadata are most relevant for linguists (81%) and least relevant for teachers (35%).
Users indicated that groups of texts (78%) are the prior unit for analysis followed by single texts (69%), and that data exports are of high relevance (73%).
Regarding the technical working environments, the survey showed that the Windows operating system (95%) and the browsers Internet Explorer, Firefox or Chrome are part of the most used setups. More than 65% of the respondents need technical assistance for installations, and the majority of respondents was not familiar with using non-office style file types like XML.
In addition to the large-scale survey, semi-guided interviews were carried out with one participant per target language in order to enquire about concrete use cases and related demands. The interviewees expressed a need for looking up prototypical examples of evaluations according to CEFR and to learn about typical learner language features for different groups of learners (e.g. common L1) and different levels. Accordingly a filtering by learner language features as well as by metadata was considered important. The annotated texts were expected to be useful for raising the learner's awareness on his competence level as well as for discussing evaluation criteria and measures with teachers and testers.

4. Corpus Preparation and Storage

Within the project all texts were manually transcribed (using the XML editor XMLmind[6]) and annotated (using the annotation tools MMAX[7]) for learner language features. Furthermore, for each text a minimally error-corrected version ('minimal target hypothesis'), and for the A2 and B2 subset of texts also a fully error-corrected version ('extended target hypothesis') has been created (cf. Reznicek et al., 2012). Furthermore, texts were automatically annotated for lemma and part-of-speech and various statistical measures were computed for German texts (e.g. average sentence complexity, lexical density and diversity, finite verb ratio).
In order to provide the required search functionalities, the MERLIN corpus was transformed and imported into two tools for corpus management and retrieval: the search platform Lucene/SOLR[8] and the search and visualization architecture ANNIS[9]. MERLIN employs Lucene/SOLR for handling string-based searches on the plain texts and target hypotheses, as well as the filtering of texts by metadata and the creation of subcorpora. ANNIS is used to enable targeted searches on learner language annotations (e.g. capitalization error), as well as on words, lemmas and parts-of-speech, as the ANNIS architecture is particularly adapted to querying and displaying multilayer annotated corpora.

5. Design of the MERLIN Platform Structure and Search Interface

5.1 Design Principles

The primary aim of the MERLIN platform is to serve different user groups and usage scenarios. By pursuing a strict target group orientation we thus comply with e-learning standards (Mirbach et al., 2009). Accordingly, the

[5] Multiple selects were possible in the questionnaire.
[6] http://www.xmlmind.com/xmleditor/
[7] https://sourceforge.net/projects/mmax2/
[8] http://lucene.apache.org/solr/
[9] http://corpus-tools.org/annis/

platform design followed two lines: on the basis of the requirement analysis and the expert interviews we modelled target-group specific use cases to determine concrete tasks, data types and display modes of particular relevance to the prospective user groups. For an example see table 1. As for the design of the technical requirements and format characteristics of the corpus data and annotations the results of the technical part of user study as well as general design principles from usability standards (ISO 9241-11) were taken into consideration.

In particular, target group orientation has been implemented in the macro- as well as in the micro-structure of the platform in the following ways:

- Implementation of different search areas, which respond to specific needs of the different target groups regarding search as well as results display
- Modelled usage scenarios for different user groups
- Implementation of different help and support structures

Usability standards are respected by providing for self-descriptiveness, controllability of the interaction and error tolerance. Regarding the users' technical requirements the platform avoids the need to install browser-plugins or additional software, and has been tested for the most frequently used browsers.

Above that, the platform takes into account that teachers and testers are often not familiar with classical corpus interfaces. An end user study conducted by Campillos Llanos among teachers of Spanish as a foreign language who were to evaluate the interface of an oral learner corpus revealed a need for explanation of error descriptors and related terms and the wish for the visual simplification of the search interface (Campillos Llanos 2012, p. 245). As a conclusion Campillos Llanos recommends to present search options in a more dynamic way and to include a comprehensive glossary of terms (ibid p. 246).

As for the MERLIN platform we decided to support the user in several regards: search options are presented in a task-oriented fashion, e.g. "Search for words in the learner texts and display them in context". Example queries present typical searches in a descriptive way and reveal results by just one click. To make sure that the interface does not appear overcrowded, help is contextualized and available in the interface as tooltip. Above that, users that are not familiar with corpus linguistic terms can refer to a glossary. Finally, material related to the learner texts, i.e. test tasks, rating criteria and scales and the annotation rules can be looked up at every point of the search process without interrupting the search.

5.2 Description of the Platform

The overall structure of the platform (see Figure 2) gives access to the search interface (1) and to an area providing background information on the corpus and specific usage scenarios (2) which present search functionalities in non-corpus-linguistics style, but rather in a task-oriented fashion. In addition, the home page offers quick info for getting started (including a video tutorial) (3).

The search interface (3) combines four different areas, which respond to specific needs of the different target groups regarding search as well as results display. In particular, the user study indicated that language teachers, testers and teacher trainers have similar demands that focus on the grouping and retrieval of texts and learner language features, while linguists demand finer-grained linguistic search options.

Initially, the four areas were subdivided as follows (see Figure 3):

- **Simple search** for accessing examples of learner language in their textual context (KWIC – KeyWord In Context) by inputting a search term into a simple text box.
- **Advanced search** providing a variety of search options on original learner texts, target hypotheses, annotations and metadata for accessing examples and annotations within context.
- **Document search** to filter documents by learner characteristics (such as L1, test level, etc.) and learner language features in order to create subcorpora, for reuse in the simple and advanced searches.
- **Search by learner language features** to derive corresponding statistics for individual texts or text groups.

Depending on the search mode, results are displayed as KWIC with or without linguistic annotations, as listings of texts and full text views, or as frequency tables for selected learner language features. Furthermore, metadata can be displayed for all results and texts are provided for download, with the option to include target hypotheses and metadata.

6. Evaluation of the Platform Prototype and Final Adjustments

In a pilot phase the platform prototype was tested and evaluated by targeted future users who were addressed via the project consortium's distribution lists in a direct mailing campaign. The online survey addressed the interface structure, functionality and content of the platform. Regarding the interpretation of results it should be noted that the major part of the 61 respondents were teachers and testers, less teacher trainer and linguists.[10] The overall distribution of participants is shown in Figure 2.[11]

Overall, the aims for using the platform matched the indicated user profiles, e.g. 81% of the teachers managed to use the platform for preparing teaching material. 80% of testers would use the platform for preparing test material, but only 40% as reference for rating. Only a small percentage of teachers and almost no teacher trainer were interested in doing linguistic studies.

[10] By using mailing lists in order to maximize the spreading of the questionnaire, the team was unable to control the exact proportions of participants by target group.

[11] The questionnaire allowed participants to indicate more than one profession.

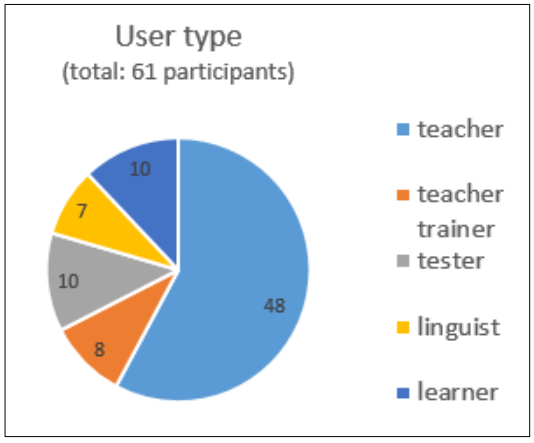

Figure 2: Participants by profession

In general, full learner texts and words in sentence context were considered the most important types of data. This is true mainly for teacher trainers and testers. Linguists were much more interested in learner language features and ratings. Surprisingly, metadata were considered important only by about half of the participants. However, distinguishing the responses of the different user groups, it turned out that mainly teachers were little interested in metadata, while all linguists indicated a strong interest for metadata as well as 75% of the trainers and 80% of the testers. This was taken as indication that metadata might need a better explanation, as teachers might not be familiar enough with the concept of metadata.

More than 80% of the pilot users were satisfied with the subdivision into four search areas, but the single search options were assessed differently, e.g. the document search was well appreciated by teachers and teacher trainers, whereas of lower value to almost half of the polled linguists. The simple search has shown to be well accepted in general, but less valued by linguists. The learner language feature search was most difficult to understand.

79% of the respondents were positive about the provided help on the interface and explanations on the corpus and the annotations. In addition to sample searches, the MERLIN platform offers information on concrete use cases and presents didactically motivated procedures for interacting with the provided learner data. Despite this information given, some users indicated that possible usage scenarios are not clear, which indicates that the given information needs further improvement to be easily found and understood. Furthermore, comments revealed that more sample searches and a clear guidance on the differences and connectivity between the four search areas would be helpful.

In particular, the users had difficulty to understand that the 'document search' serves to create subcorpora for use within simple and advanced search modes. The pilot stage was followed by a comprehensive revision process in which, to name an example the 'document search' was renamed into 'define a subcorpus' and an introductory explanation was added.

7. Conclusion

Results of the pilot study showed that the multiple access modes are suitable to match different target user needs and that it is necessary to reduce complexity when presenting richly annotated data by grouping and faceting search options, offering sample searches and context-sensitive help, and giving clear guidance on what kind of information can be retrieved.

The MERLIN platform aims at bridging the gap between technology development, multi-layer annotations and pedagogical applications by offering four approaches to the data: a simple and an advanced search, a metadata and feature-driven document search allowing for defining subcorpora at the same time, and a separate section for exploring frequency information.

Within MERLIN, it was not feasible to implement functionalities of the collaborative web, but both studies clearly revealed that future corpus users would appreciate support for sharing and commenting search results, subcorpora and best practices.

8. Acknowledgements

The MERLIN project has been funded with support from the European Commission, Lifelong Learning Programme.

9. Bibliographical References

Abel, A.; Wisniewski, K. et al. (2014). A Trilingual Learner Corpus illustrating European Reference Levels. In: Ricognizioni – Rivista di Lingue, Letterature e Culture Moderne 2 (1), 111-126.

Boyd, A.; Hana, J. et al. (2014). The MERLIN corpus: Learner Language and the CEFR. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA)*, Reykjavik, May 26-31, 2014.

Campillos Llanos, L. (2012), Designing a search interface for a Spanish learner oral corpus: The end-user's evaluation. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012). Istanbul, Turkey, pp. 241–8. Available at www.lrec-conf.org/proceedings/lrec2012/pdf/574_Paper.pdf (last accessed on 09.03.2016)

Council of Europe. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press.

ISO 9241-11(1998), Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability.

Mirbach, H.; Sohn, H.M.P.; Pawlowski, J.M.; Reß, L.; Sonnberger, J.; Stracke, C.M. & Strahwald, B. (2009), 'QPL Qualitätsplattform Lernen: Das Instrument zur Qualitätssicherung in der Bildungsbranche', In: Fachausschuss Qualität des D-ELAN Deutsches Netzwerk der E-Learning Akteure e.V., Essen.

Reznicek, M.; Lüdeling, A. et al. (2012): Das Falko-Handbuch. Korpusaufbau und Annotationen. Version 2.01. Berlin.

Figure 3: MERLIN Platform Start Page

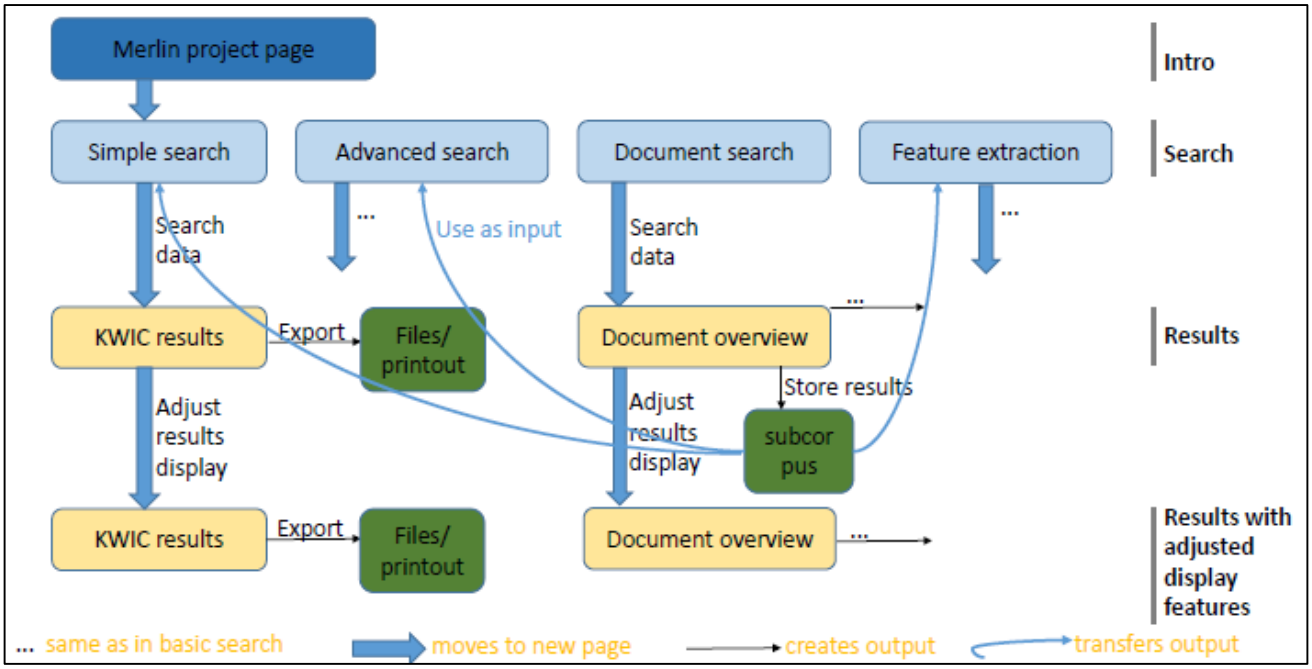

Figure 4: Diagram describing initial search interface design

aim / usage scenario	use case / task	data types	relevant information / features of learner language (LL)	relevant annotation levels	display mode
explore typical errors in context	search for words, adjacent words, POS in learner texts and target hypotheses	annotated learner productions	POS, lemma, features of LL: grammar, vocab., etc.	learner text, TH1	LL feature in context
re-adjust teachers oversensitive to special L1 errors by having them re-rate MERLIN texts without showing the MERLIN ratings and then compare and discuss the differences / results	extract a random sample of written tests on a specific tasks for sample texts by different metadata as e.g. L1, task type, CEFR level (of the test/rated CEFR level)	unannotated learner productions	available metadata, esp. CEFR level of test, fair CEFR level	learner text, (TH1/2)	entire text/ text section, metadata
Underpin the course schedule with lists of learner language features specific for different CEFR levels & identify typical and relevant milestones/errors	extract feature list by CEFR level using different filter criteria as L1 and CEFR level	feature list	available metadata, esp. CEFR level of the test and ratings linguistic annotations (statistical information – features per level)	learner text, meta-data, EA1, EA2	feature list / statistics

Table 1: Example Use Cases and related Tasks and Data Types