
# EURALEX XIX

## Congress of the European Association for Lexicography

**Lexicography for inclusion**

**7-11 September 2021**

Ramada Plaza Thraki
Alexandroupolis, Greece

www.euralex2020.gr

**Proceedings Book**
**Volume 1**

Edited by Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-1-5

Edited by: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris



2020 Edition

# Audio Recordings in a Specialized Dictionary: A Bilingual Translation and Phrase Dictionary of Medical Terms

**Sviķe S.[1], Šķirmante K.[2]**

[1] *Ventspils University of Applied Sciences, Latvia*
[2] *Ventspils University of Applied Sciences, Latvia*

**Abstract**

The present climate of insufficient funding is having an impact on the development of dictionaries such that new projects would benefit from employing as their source already-existing language material, which could then be made available for publication in contemporary electronic forms. This study reviews the design and development of a bilingual translation and phrase dictionary of medical terms in the form of mobile application "English-Latvian-English Phrasebook and Dictionary of Medical Terms" (called MED). This electronic dictionary is the result of a collaborative effort from researchers from two Latvian higher education institutions, namely Riga Stradins University (RSU) and the faculties of Translation Studies (FTS) and Information Technologies (FIT) of Ventspils University of Applied Sciences (VeUAS). The dictionary presents in a systematic manner the Latvian and English language terminology found in the study materials from RSU's specialty study courses. The collected terminology was thoroughly reviewed for relevance and supplemented with additional terms during the development of the dictionary. The need for such a dictionary was verified through a survey carried out before the implementation of the project. The successful development of the dictionary has benefitted considerably from VeUAS researchers' prior experience in the development of electronic dictionaries (in the form of mobile applications) and the expertise of RSU's medical specialists. As well as describing the functionality of the dictionary, this study describes the database model used in its development and provides an insight into the execution of the project. Additionally, it offers a detailed description of the creation and implementation of a particularly salient feature of the mobile application, namely audio recordings of terms and phrases.

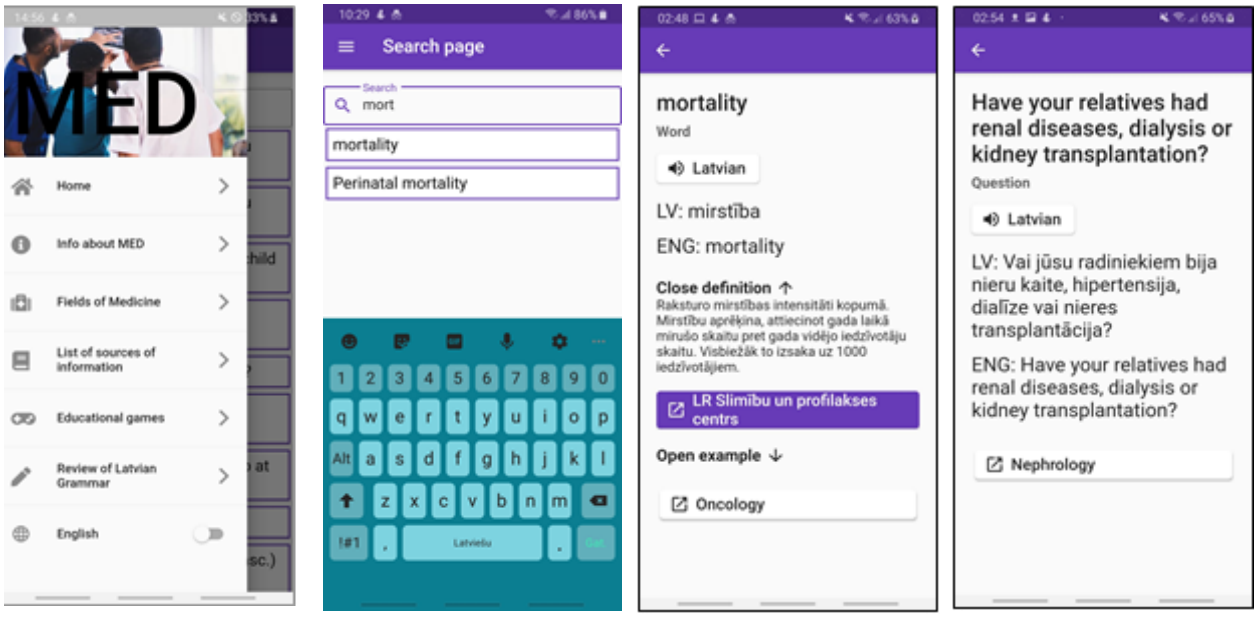
**Keywords**: Audio Recordings, Specialized Dictionary, Mobile Application, Medical Terms.

## 1 Overview of the electronic dictionary's macro and microstructure

In this section, only a brief overview of the structure of the dictionary is presented. The present description is meant to provide a general context for the ensuing discussion of the creation and development of the application's audio recordings and the related technical solutions analysed in this article. For the design of the dictionary's macrostructure, the requests and recommendations obtained from responses to a custom-made survey were taken into consideration. The results of the survey were presented at the international scientific conference "The Word: Aspects of Research", organized by Liepaja University on November 28 and 29, 2019 in Liepaja.

As shown in Figure 1, the macrostructure of the dictionary consists of the following sections, found in the form of a menu in the mobile application:

- Home. Contains a term search page.
- Info about MED. Contains informative texts about the project and the project executors, as well as a description of the mobile application.
- Fields of Medicine. Shows various available subfields of medicine, each containing relevant terms from among those compiled in the dictionary.
- List of sources of Information. Provides detailed information about all the literature sources used.
- Educational games. Provides access to games and interactive exercises for learning medical terms.
- Review of Latvian Grammar. Offers an overview of Latvian grammar with examples.
- Language Change Menu. Allows the user to switch the application's interface language between Latvian and English with the help of a "toggle" button.




(a) (b) (c) (d)

Figure 1: Screenshots of the mobile application. (a) Main menu; (b) Search view; (c) An example of the translation results for the search term "mortality"; (d) An example of the translation results for a question.

The main view (see Figure 1 (b)) is intended for entering medical terms and searching for their equivalents, in the English-to-Latvian or Latvian-to-English combinations. The view has an input field where users can enter a search term using the keypad of their smart device. An additional function incorporated into the application in order to improve user convenience is the option of searching by entering only a part of the word or term. That is, while entering a part of the search term into the input field, all dictionary entries containing said part of the term are shown. After the user chooses from the suggestions displayed, the system searches only for the selected term and the final results appear in the main view. After selecting the corresponding record, the application finds the given term in the database and returns the translation (equivalent), with the option of playing an audio recording with its pronunciation.
The MED dictionary includes three types of entries: English and Latvian terms, phrases, and questions. The need for such a dictionary, containing the features above described, was verified through a survey conducted among medical students and future doctors, and carried out before the commencement of the project. The survey revealed that set phrases and questions are very necessary and useful for communication between doctors and patients. This aspect encouraged developers to include them in the dictionary, together with their corresponding audio recordings. The core section of the dictionary also includes 200 terminological units from the medical field, selected with a focus on terms that might present particular difficulties in translation. The definitions of these terms include hyperlinks to additional information sources, as well as to contextual examples of their use in medical texts.

## 2 Practical work

The electronic dictionary as a mobile application (Android version and iOS version) was created using Google Flutter Framework[1]. The development of the mobile application's Android version was the initial priority. After the testing phase of the application using multiple Android mobile phones and emulators, additional development and configuration tasks were carried out in order to compile the mobile application version for iOS. The Android version of the mobile application was developed on the basis of Android API 19 and using Android Studio. In the testing of this version, various testing emulators (from Android 4.4 to Android 10) and physical smart devices (Xiaomi Mi Note 3 and Samsung Galaxy A7) were used. The iOS version of the mobile application was configured, recompiled and rebuilt using Xcode 11, and iPhone 6 (with iOS version 8.0) to iPhone 11 (with iOS version 13) were used as testing emulators. Development related to mobile application publishing in Apple Store (application version of iOS) and Google Play (application version of Android) is underway. Application development phases were carried out using the Agile Scrum[2] method.
For data storing, Sqlite database technology was used for both versions of the applications. Sqlite database technology provides data storage in a local database, taking into account the specifics of the dictionary and the interest shown by survey respondents to use applications without the need of a Wi-Fi or mobile internet connection.
Google spreadsheets were initially used as the working tables where researchers stored their research results. This working environment had also been used in previous projects on electronic dictionaries implemented by VeUAS

1 Google Flutter Framework, more information here: https://flutter.dev/

2 Agile Scrum development method, more information here: https://www.scrum.org/resources/what-is-scrum

(Rudziša, Sviķe, Štekerhofa 2019: 379–391; Sviķe, Stalažs 2019: 418–429; Sviķe, Šķirmante 2019: 1–17). The terms, compiled from both research and education institutions, were arranged in Google spreadsheets tables. To convert a dictionary from a Google spreadsheets format to a database format, new software was developed using JAVA programming language and the external library Apache POI (Java API for Microsoft Documents)[3], designed to manage Microsoft Word and Excel documents using the JAVA application. The software can autonomously create the database model and its tables, including terms. The database model, shown in Figure 2, is based on the document structure of the electronic dictionary.

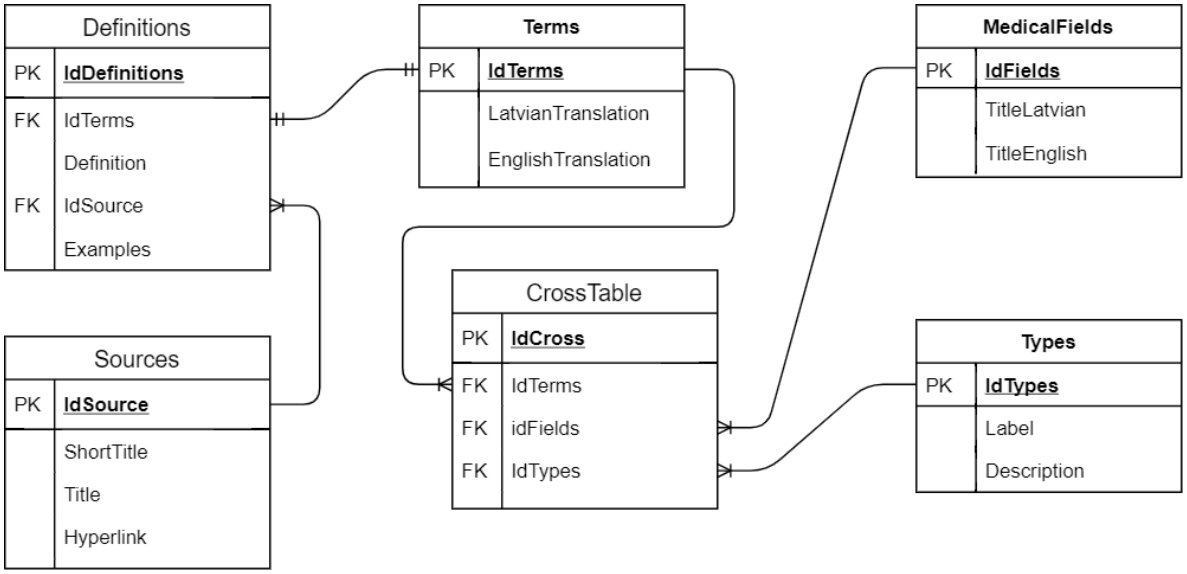



Figure 2: Application database model with relations

Recorded audio files were stored in the mobile application's resources, where each dictionary term or phrase is stored together with its own audio file, with a name corresponding to the identifier of the specific term assigned by the database. For example, the term "blood-vessel" is stored in the database table "Terms" with the IdTerms number 305, and its corresponding audio file title is "305.mp3". The application contains a total of 3785 terms, phrases and questions related to the medical field which are stored with their associated audio files.

## 3 Recording and processing of the dictionary's audio files

The Sanako Lab 100 software was used to record and process the dictionary's audio files. RecordPad Sound Recorder and Audacity were used for re-recording any faulty recordings. For this project, a group of 4th-year students from VeUAS' Faculty of Translation Studies made the audio recordings as part of their compulsory scientific practice. Students were already familiar with the Sanako Lab 100 recording system as the software had been used in their interpretation laboratory during interpretation classes, working in teams or pairs (where one student manages the recordings from the teacher's computer with the Sanako Lab software).

To simplify the recording process of the 3785 audio files, all dictionary terms were grouped in sets of 10, and all the terms of each set were recorded in a single file, with a specific amount of time assigned to each term (5, 10 or 20 seconds). For example, the first term was pronounced at second 0, the second term at second 5, the third term at second 10, and so on, obtaining in this way silent intervals between terms. The recording was done using the established random access method (Behymer 1974) with a fixed-length application. Each term had an individual database identifier (ID) assigned, and the recordings were named according to a numeral scheme indicating the IDs of the first and last terms recorded in the file: e.g. 230_239.mp3. For the automated processing of all audio files, a Python script was developed. The script split each file into 10 sections, each section being then processed separately and stored in a new audio file with the term identifier as the file name. Therefore, from the source file 230_239.mp3 the single-term files named 230.mp3, 231.mp3..., 239.mp3 were obtained. The Python script was then used to process each section of the term file by first removing the silent sections created when using a random access method to extract single terms, and subsequently decreasing the bitrate to 64k to reduce file size. A typical 5-second recording was around 215KB in size before processing, but reduced to around 10KB after processing, without incurring in any loss in the quality of the recording. The final total size for all of the application's audio files is 48.9MB, with an average file size of 13.98KB for each term, phrase or question.

While audio recordings from different speakers are considered an advantage, especially in the case of learning dictionaries (Garrett 2019: 201), the individual voice qualities of each speaker, speed of speech and lagging were the

3 Apache POI Project, more information here: https://poi.apache.org/

cause of some difficulties during the processing the recordings. A significant issue was the adaptation of file processing methods to each speaker to remove silent parts from the recorded audio files. For example, the intensity of sound pronunciation differed among speakers (notably in the case of letters ‘s’, ‘k’ and ‘p’), and therefore it became necessary to process first the sound intensity of each speaker for various sounds, and only then single-term audio files could be automatically obtained.

## 4 Used technologies and techniques

The technologies and techniques used in the development of the application are shown in Figure 3. Application development processes and workflows are described using arrows.

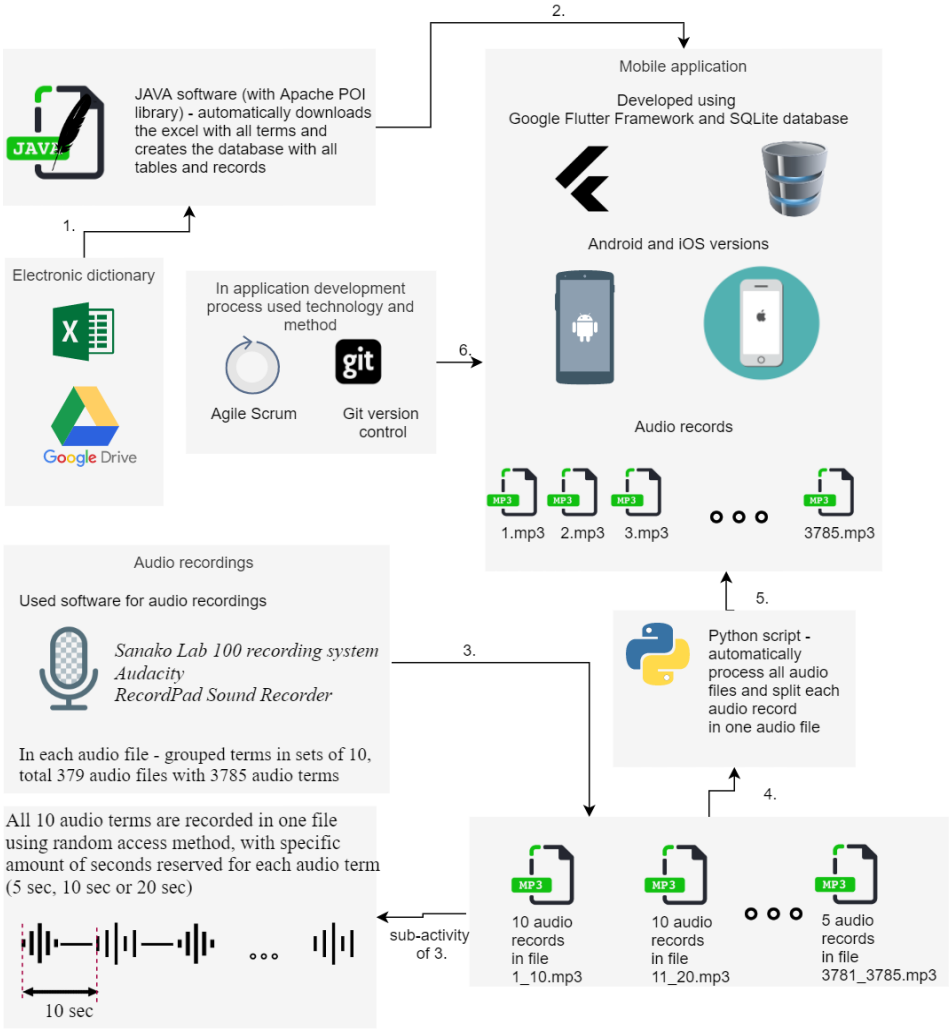



Figure 3: Used technologies and techniques

An automatic download from Google drive was obtained using JAVA software, which was then used to process data from the spreadsheet tables and create an SQLite-type database containing the databases’ tables, relations, Primary Keys (PK), Foreign Keys (FK), as well as data. The Database model is shown in Figure 2. When a SQLite database file is created, it is added into the application project. The same database has been used for both versions of the application (Android and iOS).

Application development and audio recording were tasks carried in parallel. To simplify the recording process for the speakers, ten (10) terms were stored as a single audio file. All recorded files were stored in shared directories (one directory per speaker), as the recording process spanned across several sessions and it was necessary to control the progress of the recordings. Using shared directories made it easy to manage the re-recording faulty recordings. As illustrated in Figure 3, each audio file consisted of an audio wave which included recorded pauses with lower sound intensity so they could later be processed and split automatically. Each speakers’ sound intensities for different pronunciations were analysed and taken into account before the automated processing of audio files.

During the application development process, a custom Python script was written to carry out an automated splitting of the recorded audio files into single-term ones. Each audio file title corresponds to the term’s primary key in the database, thus providing faster and more efficient audio file searches in the application during playback. After file processing, the resulting audio files were stored in the application resource directory and not in the database, since retrieving a term’s audio file from a resource directory is faster than its retrieving and decoding from a database.

The Agile Scrum method was chosen to organize the development process more efficiently. Each development phase consisted of 2-week stages called 'sprints', with specific results obtained at the end of each 'sprint'. To share application source code between programmers and to manage source code versions during different phases of the development process, the version control and GitHub software development platform were used.

## 5 Conclusion and Future Work

This research forms part of a project concerning the conception and development of a bilingual translation and phrase dictionary of medical terms, as well as the creation of a working model of the dictionary in the form of a mobile application. Aiming to maximize the dictionary's usefulness, there exists the possibility of adding new terms, phrases and even sections in the future. The research work here described may be continued in the future in order to expand its theoretical scope, improve the dictionary's current list of entries, and modernize the mobile application according to newfound technological possibilities and users' needs. This study offers a detailed description of the process for the incorporation of audio recordings to the application and characterises the technologies used in the dictionary. The conclusions and suggestions of the present research could be particularly useful in the development of future translation and phrase dictionaries of this type, especially considering the dynamic "shift from p-lexicography to e-lexicography" (Tarp 2012: 107–119).

At the end of the project, it has been concluded that only one voice per gender should be selected to be used in the audio files of the dictionary, as this would provide consistency throughout recordings. For future recordings, each speaker's sound intensities for different letters could be analysed using neural networks. Research in input data corrections algorithms using artificial neural networks in mobile applications is currently underway and soon it might be possible to implement relevant research results in the developed mobile application.

The data material incorporated into the dictionary was divided and compiled according to the areas of responsibility of the working group members of both research institutions. It should be taken into account that when working in large groups it is crucial to monitor completed work, plan regular group meetings, establish future steps, identify problematic issues and implement adjustments. Since the dictionary project was executed in about a year, it was necessary to continuously check whether all steps of the working process were being executed. After the completion of the project, it has been concluded that the work entrusted to the students must be checked thoroughly.

## References

Behymer, J.A., Ogilive, R.A., Merten, A.G. (1974). Analysis of indexed sequential and direct access file organizations. In *SIGFIDET '74: Proceedings of the 1974 ACM SIGFIDET (now SIGMOD) workshop on Data description, access and control,* pp. 389–417.

Garrett, A. (2019). Online Dictionaries for Language Revitalization. In L. Hinton, L. Huss, & G. Roche (eds.) *The Routledge handbook of language revitalization.* Abington-on-Thames: Routledge, pp. 197–206.

Rudziša, V., Sviķe, S., Štekerhofa, S. (2019). Juridisko pamatterminu glosārijs līgumtiesībās Latvijā izdoto nozarvārdnīcu kontekstā. In G. Smiltniece, L. Lauze (eds.) *Vārds un tā pētīšanas aspekti*. 23 (1/2), Liepāja: LiePA, pp. 379–391.

Sviķe, S., Stalažs, A. (2019). "Jaunās botāniskās vārdnīcas" mikrostruktūra: tradicionālais, mainīgais un inovatīvais. In G. Smiltniece, L. Lauze (eds.) *Vārds un tā pētīšanas aspekti*. 23 (1/2), Liepāja: LiePA, pp. 418–429.

Sviķe, S., Šķirmante, K. (2019). Practice of Smart LSP Lexicography: The Case of a New Botanical Dictionary with Latvian as a Basic Language. In I. Kosem, T. Zingano Kuhn, M. Correia, J.P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, C. Tiberius (eds.) *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference.* 1-3 October 2019, Sintra, Portugal, Brno: Lexical Computing CZ, s.r.o. pp. 1–17.

Tarp, S. (2012). Theoretical challenges in the transition from lexicographical p-works to e-tools. In S. Granger, M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press, pp. 107–118.

**Acknowledgements**

This research has been funded by the Latvian Council of Science, project "Smart complex of information systems of specialized biology lexis for the research and preservation of linguistic diversity", No. lzp-2020/1-0179