Université de Lorraine

IDMC

ATILF

MSC Natural Language Processing

UE 705 – Supervised Project

Database Creation For The Lex.E.M Project

Academic Year: 2020-2021

Host Organization: ATILF

Students:

Morgan RUIZ-HUIDOBRO

Soklay HENG

Supervisors:

Etienne PETITJEAN

Marie Laurence KNITTEL

Samantha RUVOLETTO

December 18, 2020

Contents

Abstract

A lexical database contains information about words but is considerably different from a dictionary. The main difference is their contents and the usage that each of them offers. A dictionary explains or translates words, while lexical database's main purpose stands in this research. This document is the bibliographic report of our supervised project. The goal of the project is to create a database for the Lex.E.M[1] project. The Lex.E.M project is part of a collaboration with kindergarten from REP[2], and is currently directed by the ATILF[3] laboratory.

In this report, we will talk about the lexicon acquisition of 2 to 5 year-old children, and the reason why the lex.E.M project is important in order to create remediation tools to help children access the lexicon. We will also describe in more details the process and useful concepts of database creation.

[1] Lexique pour les Écoles Maternelles (lexicon for kindergarten)

[2] reseaux d'éducation prioritaire (high priority education network)

[3] Analyse et Traitement Informatique de la Langue Française

Introduction

Communication between people plays a vital role in our daily lives because we need to express words to let our recipient knows what we are trying to imply; therefore, lexical knowledge is required for each individual to make the interaction go smoothly; otherwise, there will be a misunderstanding or the communication will not happen. How can a person build lexical knowledge? A baby is not innate with a language, so they need to communicate and observe to learn the language spoken around them. They learn word by word along with phonology and meaning of words in their native language; hence, they will gradually master syntactic and morphological rules of their native language, resulting in setting up an effective communication system[1]. However, children with lexical deficiency face problems with vocabulary comprehension and production during the developing stage of acquiring their native language.

In response to this problem, we are motivated to work on building a database containing a list of vocabulary. This database will be useful to create remediation tools for those children who need special assistance in helping them to achieve a successful acquisition of their spoken language. Our project centres on creating a database of French word list which will be beneficial for pupils from Réseau d'Éducation Prioritaire(REP)[4]. We will collect French basic vocabulary produced by children in a real context, and with this database, a web interface will be created. This will be an assisting tool in research to find ways for children to learn the lexicon faster and more effectively.

[4] REP is a French priority education policy aiming to promote schools by strengthening the pedagogical and educational action from nursery school to college in areas where the great social and financial disparities exist[2]

1 Lex.E.M project presentation

1.1 Language acquisition

Eve V. Clark in *The lexicon acquisition*[3], describes the lexicon acquisition for children. The book mentions that language acquisition begins with words. Possessing some words allows children to make a generalisation and to organize them in categories. After that, children need to instantiate syntactic categories. Without words, children don't have access to phonological contrasts[5] and complex phonological structure, word structure, and syntax. The question frequently asked in linguistics is how the mental lexicon is organized. The lexical items stored in memory can be not only a word, but also an idiomatic phrase.

The author describes the supposed lexical entry of the mental lexicon. It presumably contains four types of information, such as semantic, syntax, morphology, and phonology.

An adult English speaker disposes of a production vocabulary in between 20 000 and 50 000 word forms, and a comprehension vocabulary is even greater. From those numbers, we can observe the considerable task that children have to face to acquire the lexicon. Children begin their journey at age one, but the lexicon growth is very slow in this first stage. By the age of two, we consider that normally developing children learn 10 new words a day. They are supposed to have 14000 words in their production vocabulary by age six. The growth of the lexicon is faster until their teenage years, after that the growth slows down into adulthood. The first six years of life are therefore really important for the development of the lexicon.

Mabel L. Rice in *Children's Language Acquisition*[4] discusses the issues children can have while acquiring language. It also mentions the research done that in the long term will be useful to help children who have difficulties.

To communicate effectively, children need to master four types of knowledge, such as semantic, syntactic, morphological and phonological knowledge (as mentioned above in the content of the mental lexicon).

In children development, language plays a big part in the social interaction it is allowing. Children who have trouble with language are at risk for educational achievement; it also represents a risk for their future reading skills, limits their verbal skills and in the long term affects their social skills. Helping children who have difficulty requires strategies specially designed to meet the needs of each individual. The earlier children receive help,

[5] refers to a minimal phonetic difference, small difference in between sounds (phoneme)

the better it is, especially in the early stage of development. Accessing corpora on child language allows research on language acquisition. In the long term, it is allowing to develop strategies to help children in their development.

MacWhinney in *The child language data exchange system*[5] discusses the project of a shared database that regroups data on child language. It is useful in research and allows global searches for word combinations across the corpus.

The concept of a shared database is beneficial, especially when you want to research child language where you can have access to a lot more data. Having a big database allows reducing the intra-individual variability that can exist in languages in language acquisition. This variability can come from the input quality[6], the parenting style, the children cognitive development, and eventually from some impairment or pathology.Generalising the data allows to conduct research that will represent phenomena closer to reality. The data you are searching will be available and found immediately in the shared database, which is faster than recording child language again. It is difficult to record data on child language because most of it is produced at home with the family. French corpus on language production of 2 to 5 year-old children is scarce; therefore, creating a shared database would be interesting.

1.2 Current issue

Teachers have found that children from REP (francophone or allophone) present notable lexical deficiency, including basic vocabulary used in school. These deficiencies are concerned with comprehension and production.

To elaborate new remediation tools that support lexicon access for kids of 2-3 years old, it is necessary to know with precision the basic vocabulary used by children from standard areas (not REP) in schools. To this day, there is no available database of linguistics production of kids in school. The only existing database is Manulex [6], composed from the lexicon of primary school textbook (CP to CM2). It's a written lexicon for a different audience, and potentially not representative of the children real production.

So, the current issue regarding the Lex.E.M project is the lack of data. If they had the data on child language of 2 to 5 years old, they could develop tools to help children from this age group.

[6] exposure to language the child have, what will be internalized by him and will be applied later in life

1.3 Project goals

The end goal of the Lex.E.M project is to assist children who have difficulties in acquiring the lexicon. To achieve the objective, the project is composed of multiple steps. The first one is to create a new corpus, that contains the basic vocabulary used in kindergarten. They created the Corpus Jeunes Enfants en Milieu Scolaire (JEMS), which records child language in a school context, and we will talk about it in the next part. The second step is the elaboration of a database containing children real production and adult/child interaction. The third step is to create an interface available on the web. The fourth step is to develop tools to help children with difficulties, and especially for kids who are 2-3 years old.

2 Supervised project presentation

Our project focuses on the second step of the Lex.E.M project, the creation of a database on child language. The creation of the database includes several steps. The first step will be to familiarize ourselves with existing corpora on child language and to homogenize the format between all of them. The second step is to retrieve lexical form and morpho-syntactic labelling. The third step will be to add the missing data derived from the difference between the corpus. Then, we will be able to associate various frequency indices with each lemma. The last step for us will be to create a table containing all the information created by the database.

We are interested in this project because of its close link with linguistic, knowing that our backgrounds lie in linguistic. The idea of creating a database that will be beneficial for future research in children development is appealing to us. For the non-native French speaker member of the group, there is also a challenge since the corpora used are in french.

3 Corpora

In this part, we will talk about the already existing corpora on child language. We have researched the content and the form of each corpus. We will also give information on the format used. In addition to this, we will talk about which information we want to include in the database.

3.1 Existing corpora

There are currently three existing corpora that could be interesting to include in the database because all of them centre on child language.

The first one is the Child Language Data Exchange System (CHILDES)[7, 8, 9]. It is an international database containing some corpora on French language. Corpora that contain information on early child phonology are available on the PhonBank website. This can be useful for us because the majority of these corpora contain transcriptions of child production. There are seven PhonBank corpora (GoadRose, Hunkeler, Kern-French, Lyon, Paris, Standford-French, and Yamaguchi) that have full transcriptions. Those corpora used the Phon program to produce the transcriptions. They are included in the CHILDES database, but all of them are not presented in the same way. For example, the Hunkeler and Lyon corpora seem to analyse the production word by word, and they also contain morphosyntactic information.(see figure[6] in annexes). Some corpora on CHILDES don't have any morphosyntactic information.

The second corpora is the Communication Langagière chez le Jeune Enfant (Colaje)[10, 11, 12]. The ANR Colaje project is in continuation of the ANR young researchers' project LEONARD[13].The goal of the Colaje project was to create a database with voice recording and transcription. At the time, the CHILDES corpora were already existing but were lacking French corpora, so the second goal of the Colage project was to have a francophone database on the monolingual and bilingual adult/child communication. At the end of the project, this database had to be shared and continually improved with new data. The project also wanted to add non-verbal cue in the database, so they had to set up specific annotation to keep track on gesture, eyes, voice quality, and the target of the gesture or the vocal production. Each recording also contains metadata, such as the age of the child, the language used, the interlocutor, context, location, duration, and the time of recording. In addition to this, the recording is also accompanied by a summary, a situation time-frame, and the principal characteristic[7] of the child.

They used the transcription format created for the project CHILDES (CLAN software) and the transcription format CHAT; for the phonological need, they used PHON created by Yvan Rose. (see figure [7][8][9] in annexes).

[7] such as the child motor skill, cognitive skill, relational skill, and his linguistics characteristic

The Colaje database contains three corpora. The one that interests us contain 10 follow-ups of children until age of 3 in interaction with adults. This corpus comes from precedent work, project ANR JCJC Léonard, project Adonis ENFLANG and project CREAGESTES. This corpus contains 300h of French, 24h French/English, 24h French/Italian and 50h of LSF.[8]

The third corpora are the Corpora Jeunes Enfants en Milieu Scolaire (JEMS). They are corpora containing adult/child interaction in a school context. The participating schools have recorded 25 videos; 19 videos were usable (no background noise, and great distinction of each speaker), and 17 videos were transcribed. The Lex.E.M project contains 7 videos; 6 videos are usable and transcribed. PHON was used to realize the transcription process. In this corpora, we can find the orthographic transcription, the phonological form the child should have produced and the actual phonological production produced by the child. We can also find some information about the situation.

3.2 Corpora format

3.2.1 CLAN software

CLAN[9] is a software containing multiple programs. CLAN can be separated into two categories based on the functionalities. The first one is the CLAN editor; it is used to edit file in CHAT or CA(conversion analysis) format. The editor part provides additional function, such as video playback, linkage to audio and video, data validation, adding code to files, and shipping data to other programs. The second part of CLAN consists of a set of data analysis programs.

The CLAN software allows interoperability between programs thanks to the CHAT format. For example, CHAT files can be exported and imported from Praat which is a program for phonological and phonetic analysis. The Phon program is also an example of program fully interoperable with CLAN.

3.2.2 CHAT transcription format

CHAT is a system which enables us to produce computerized transcripts of face-to-face conversational interactions with a standardized format[7].(see figure 9 and 10 in annexes) CHAT stands for "Codes for the Human Analysis of Transcripts" and is a standard system used for transcribing for TalkBank and CHILDES projects. It is important to

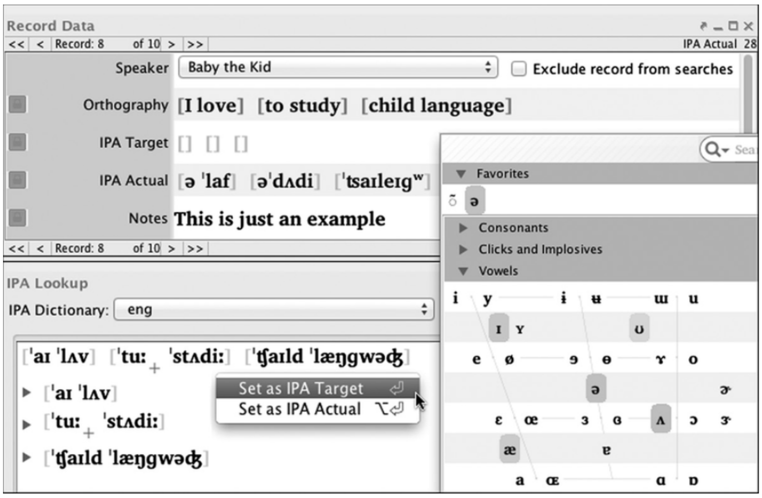
[8] Langue des Signes Française; french sign language

note that all of the transcripts in TalkBank databases are in CHAT format. It is designed to be used with both normal and disordered participants. It also aims to be used with all types of learners, such as children, second-language learners, and adults recovering from aphasic disorders. The CHAT program enables us to track various structures and provides automatic indices computation and morphosyntax analysis. In addition to this, because CHAT can be converted to XML (a language used for text documents on the web), it is also compatible with other programs, such as ELAN, Praat, Phon, and Transcriber.

3.2.3 Phon format

Phon is a software program supporting data corpora of phonology and text[14]. It supports all types of corpus studies concerning with language data produced by child or adult, such as phonological, textual, acoustic, and clinical studies. What is special about Phon is that it provides support for phonological units research, including phones, phonological features, stress, and tones. Special features for clinical speech analysis and acoustic data analysis are also available in Phon.

Queries and analyses results from Phon can be saved in print-ready HTML format, CSV, and Excel workbooks. In addition to this, it will generate links between query/analysis results and Phon data where the transcript data corresponding to these results will be open automatically by clicking on these links.



Record Data

<< < Record: 8 of 10 > >>   IPA Actual 28

Speaker: Baby the Kid   □ Exclude record from searches

Orthography: [I love] [to study] [child language]

IPA Target: [] [] []

IPA Actual: [ə ˈlaf] [əˈdʌdi] [ˈtsaɪleɪgʷ]

Notes: This is just an example

<< < Record: 8 of 10 > >>

IPA Lookup

IPA Dictionary: eng

[ˈaɪ ˈlʌv] [ˈtu: ˈstʌdi:] [ˈʧaɪld ˈlæŋgwəʤ]

▸ [ˈaɪ ˈlʌv]

▸ [ˈtu: ˈstʌdi:]

▸ [ˈʧaɪld ˈlæŋgwəʤ]

Figure 1: Transcription with Phon.

3.3 Content of our database

The common point in all the corpora we have seen is the presence of transcription for child language production. We can expect this sort of information in the database we want to create. What's lacking in some corpora is the morphosyntactic data.

As seen prior the goal of the database is to have information on child language and thus have a representation of the mental lexicon. Hence, we can hope to find in the database information that is supposedly contained in the mental lexicon, such as semantic, syntactic, morphological, and phonological information.

Therefore, the information we can expect in the database are the orthographic form, the morphosyntactic tags, the definitions, and the phonological forms (expected/produced). We will also need to include contextual information, and we can add some non-verbal cue for some corpora.

4 Database creation

4.1 What is database?

The database is an orderly collection of ordered information or records, normally stored electronically in a computer system. Data in the most common types of databases currently in use are usually modelled in rows and columns in a set of tables to make data processing and querying efficient. The database could contain multiple tables, and each of these tables contains various fields that are related to the information stored in the table[15]. (see annex 11 for an example of database)

4.2 Advantage of database

Since the database is a collection of information, it allows you to input criteria, and conduct a search to create a list of leads that meet those criteria. It will save your time. With just a couple of clicks, information is immediately available, and you can look at data in all sorts of ways. Unlike traditional data storage, where you need to find your required information manually one by one; and therefore, consumes so much time. For example, there are thousands of journals stored in a library bookshelf, and you need to find the year of publication of a specific journal, so you would probably spend your whole day to find that specific journal manually just to find the year of publication.

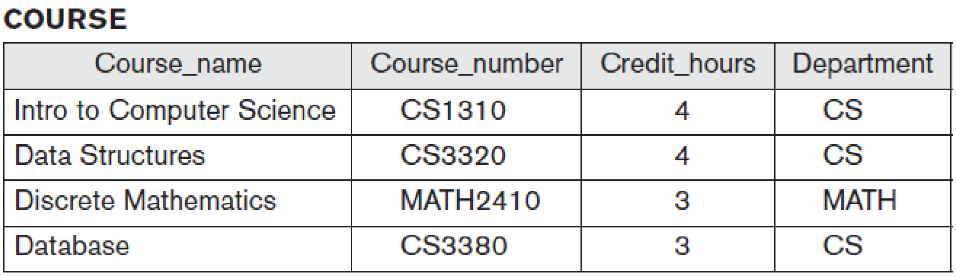
In our project, what can be considered as traditional data storage is dictionaries. However, it often requires a lot of pre-processing before using them in research. A lexical database is preferable to a dictionary when you have a research question. With a lexical database you can study the state of a language or eventually its evolution. For example, with lexical database on child language you can study the process of language acquisition, such as the development of grammatical or syntactical form.

4.3 Types of database

There are two types of database structures. Those are single-file or flat file database and multi-file relational or structured database.

4.3.1 Single-file or Flat file database

A flat database refers to a database which stores data in a plain text file[16]. Each line of the text file contains a single record, with fields separated by delimiters, such as commas or tabs. A flat file database cannot contain multiple tables like a relational database because it uses a simple structure. This simply means that it is a database of just one table.



COURSE

Course_name	Course_number	Credit_hours	Department
Intro to Computer Science	CS1310	4	CS
Data Structures	CS3320	4	CS
Discrete Mathematics	MATH2410	3	MATH
Database	CS3380	3	CS

Figure 2: A Single-file Database.

4.3.2 Multi-file relational or Structured database

A relational database refers to a database storing several data tables of rows and columns. Those rows and columns are connected via special key fields. It uses Structured Query Language (SQL) which is a standard user application that provides an easy programming interface for database interaction[17]. This feature helps you to retrieve an entirely new table from one or more table with a single query[18].

One or more data or record characteristics relate to one or many records to form functional dependencies. These are categorized as follows:

- one to one - where one table record relates to another record in another table
- one to many - where one table record relates to multiple records in another table
- many to one - where more than one table record relates to another table record
- many to many - where multiple records relate to more than one record in another table

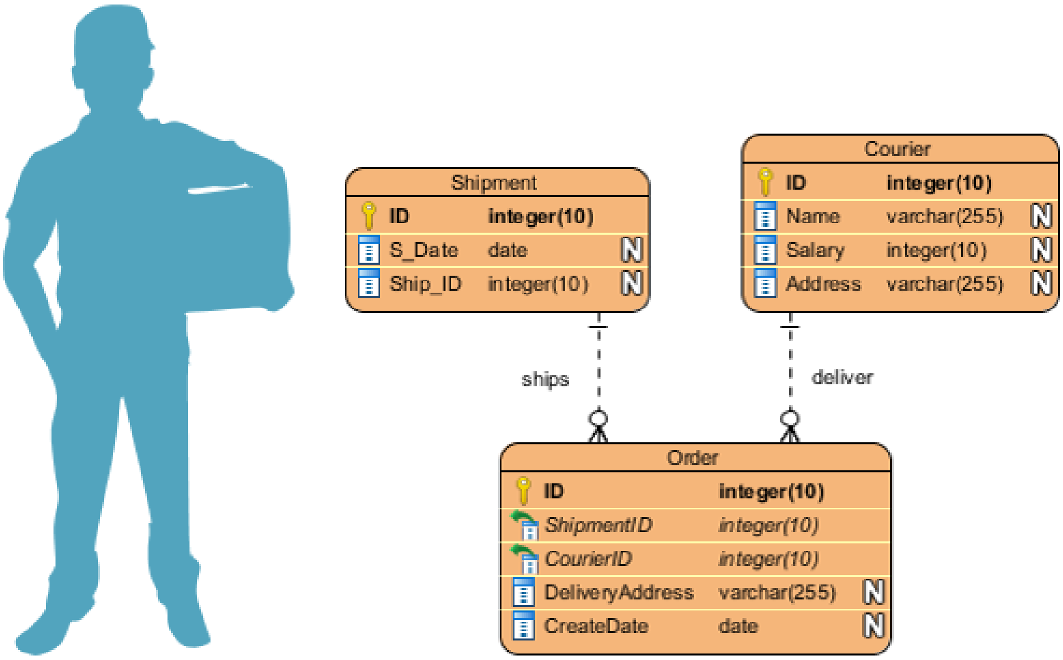


Figure 3: A Relational Database.

4.4 What is Database Management System?

A database management system (DBMS) is a collection of programs that allows users to create and maintain a database[19]. It facilitates the processes of defining, constructing, manipulating, and sharing databases between different users and applications. Defining refers to a process of specifying the structures, data types, and constraints of the data to be stored in the database. Constructing involves storing the data on some storage medium that is controlled by DBMS. Manipulating provides some functions such as querying and updating data, and generating a report from the data. Sharing enables multiple users and programs to access the database simultaneously.

These are some examples of popular DBMS, such as MySQL, Microsoft Access, Oracle, PostgreSQL, dBASE, FoxPro, SQLite, IBM DB2, LibreOffice Base, MariaDB, Microsoft SQL Server.

The database and software together form a database system. An application program accesses the database by sending requests for data to DBMS. Then, a query causes some data to be retrieved.

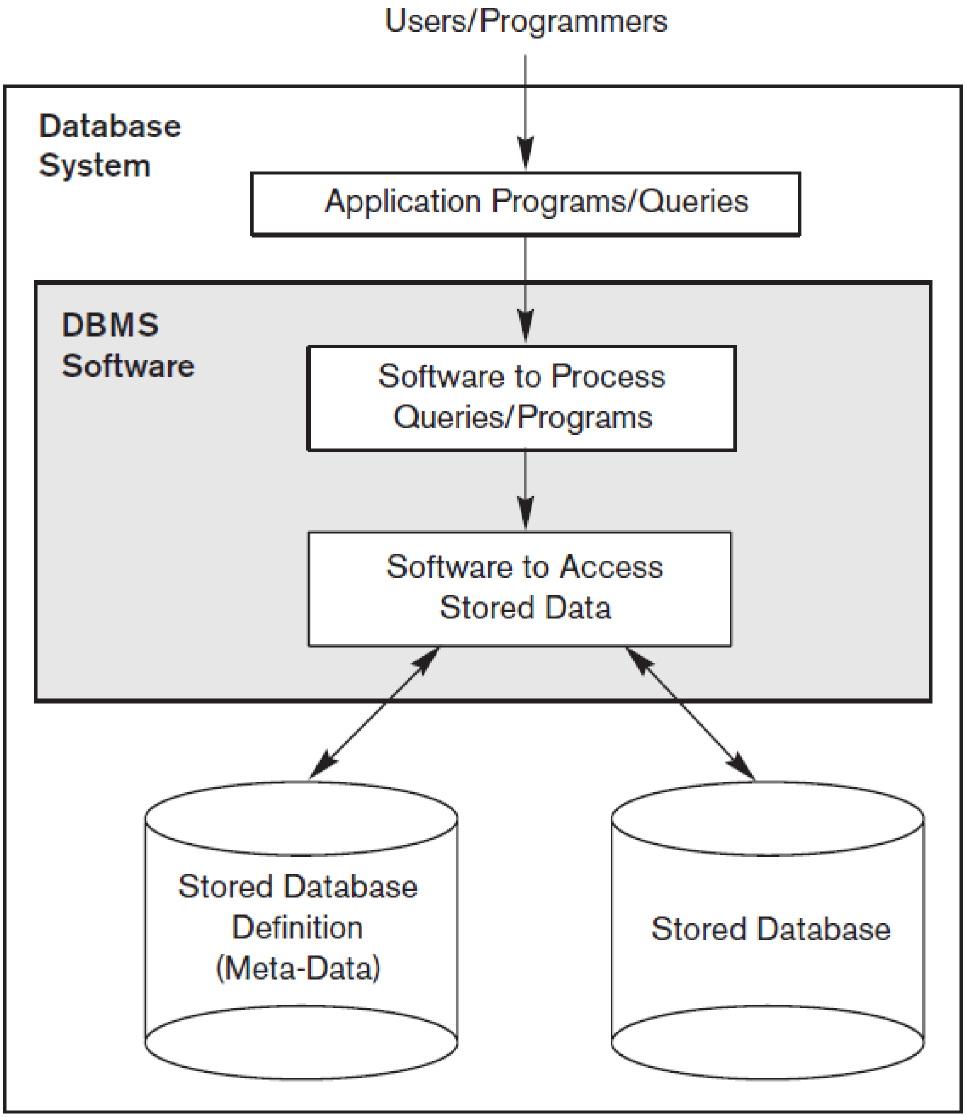

Users/Programmers
Database System
Application Programs/Queries
DBMS Software
Software to Process Queries/Programs
Software to Access Stored Data
Stored Database Definition (Meta-Data)
Stored Database

Figure 4: A Database System Environment.

4.5 Database operation

CRUD is an acronym of create, read, update, and delete, and CRUD is the basic operations in database programming[20]. These operations are useful in creating and managing data. The relational database consists of row and column tables. In a relational database, each row of a table is referred to as a tuple or a record. Each column in the table represents a particular attribute or field. The four CRUD functions can be used by users to execute various types of operations on chosen data in the database. This may be achieved using a code or a graphical user interface.

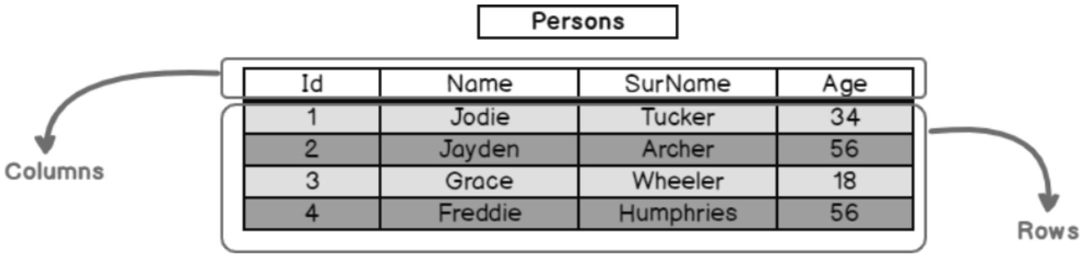
- Create: enables users to create a new record in the database. In the SQL relational database application, the Create function is called INSERT.
- Read: provides a search function. It enables users to search and retrieve specific records in the table and read their values. Users may be able to find desired records using keywords, or by filtering the data based on customized criteria. In the SQL, it is called SELECT.
- Update: is used to make a modification on existing records existing in the database. In SQL, the update function is called UPDATE.
- Delete: enables users to remove no longer needed records from a database. In SQL, the delete function is called DELETE.

5 Table creation

The database table is where all the data in a database are stored. The table is identified by a unique name and is made up of columns and rows. Data is logically organized in a row-and-column format similar to a spreadsheet[21].

A row refers to a smallest unit of data that can be inserted into a database. A row represents a unique record in the table. This simply means that it contains a record or data for the column[22].

A column contains the definition of each field. It represents a field in the record. Each column was given a name, so that it is used to reflect the contents of each cell in that column[22].



Persons

Id	Name	SurName	Age
1	Jodie	Tucker	34
2	Jayden	Archer	56
3	Grace	Wheeler	18
4	Freddie	Humphries	56

Columns
Rows

Figure 5: A Table Containing Rows and Columns.

6 Conclusion

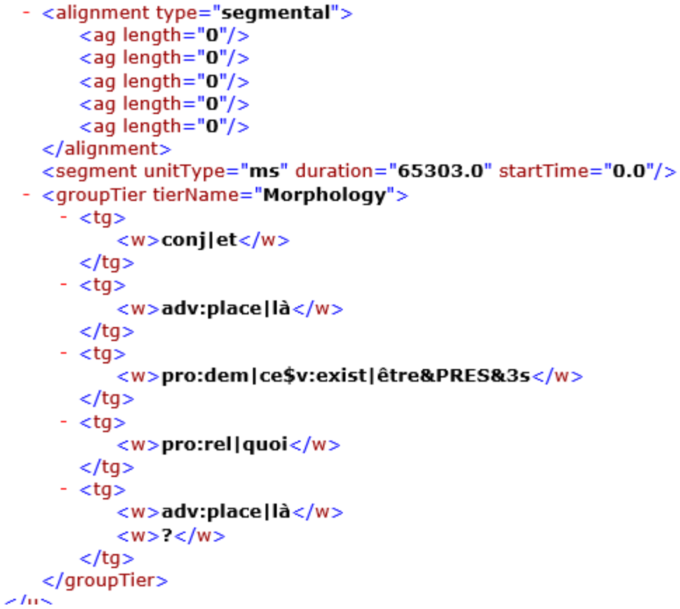
After we have familiarized ourselves with the three existing corpora mentioned above and some useful concepts of how to create a database, we will briefly introduce a pre-plan of what steps will be taken into account for the realization part in this section. The first step will be to see what information is lacking in some corpora and try to obtain them. We already know that some of them lack morphosyntactic information. Consequently, in the next step, we will have to obtain the lemma produced by the adults and the children, which refers to the orthographic form. We will need to annotate some sentences produced by those pupils which have not been annotated to obtain the lemmas. To pre-annotate, we can use TreeTagger, a system which enables us to put a morphosyntactic tag on a lemma.

Therefore, our main objective for the second half of the project will be to to create a new specialized children lexical database beneficial for those children, which will be taking a step in the second semester.

Acknowledgment

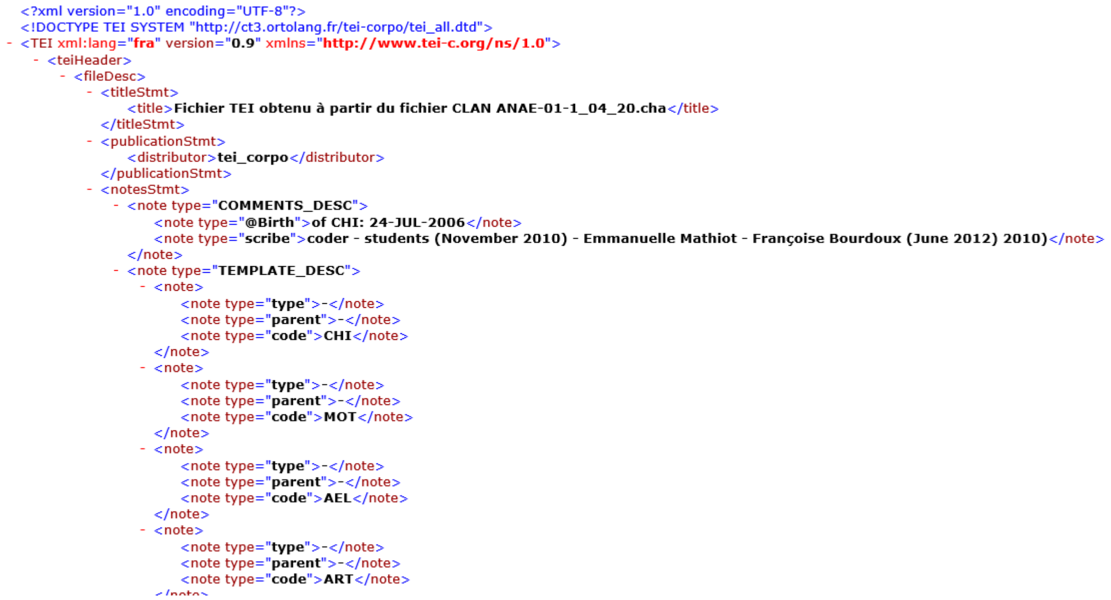
We would like to thank our supervisors Etienne Petitjean, Marie Laurence Knittel and Samantha Ruvoletto for their support and guidance during this project.

Annexes



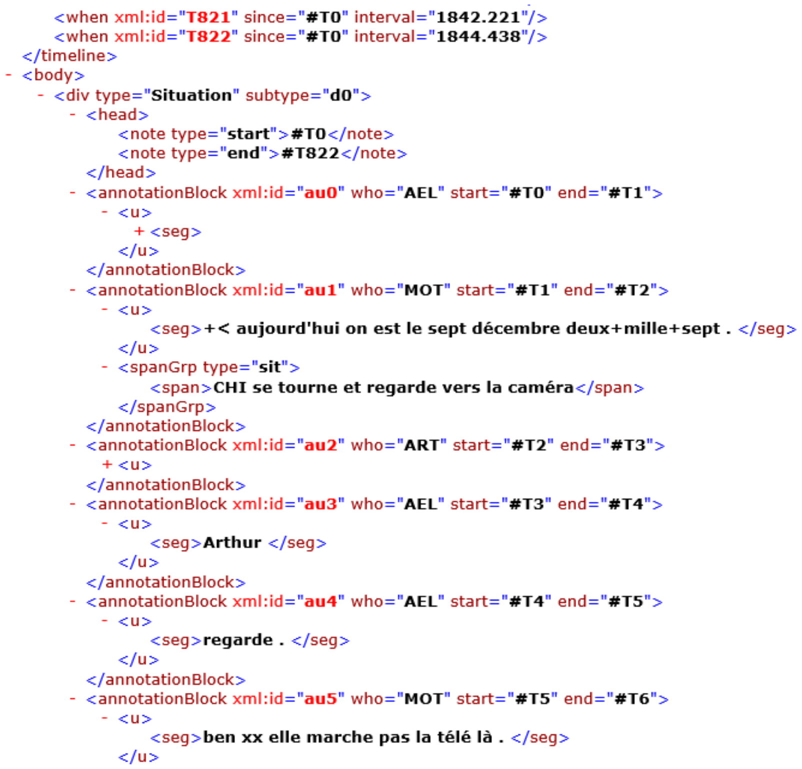
```
- <alignment type="segmental">
      <ag length="0"/>
      <ag length="0"/>
      <ag length="0"/>
      <ag length="0"/>
      <ag length="0"/>
  </alignment>
  <segment unitType="ms" duration="65303.0" startTime="0.0"/>
- <groupTier tierName="Morphology">
  - <tg>
        <w>conj|et</w>
    </tg>
  - <tg>
        <w>adv:place|là</w>
    </tg>
  - <tg>
        <w>pro:dem|ce$v:exist|être&PRES&3s</w>
    </tg>
  - <tg>
        <w>pro:rel|quoi</w>
    </tg>
  - <tg>
        <w>adv:place|là</w>
        <w>?</w>
    </tg>
  </groupTier>
</u>
```

Figure 6: Hunkeler: Camille corpora format XML



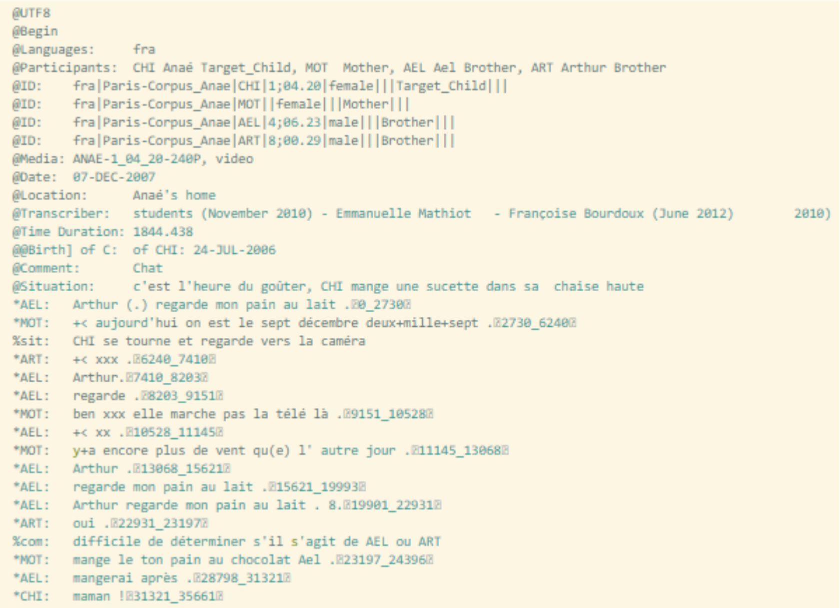
```
  <?xml version="1.0" encoding="UTF-8"?>
  <!DOCTYPE TEI SYSTEM "http://ct3.ortolang.fr/tei-corpo/tei_all.dtd">
- <TEI xml:lang="fra" version="0.9" xmlns="http://www.tei-c.org/ns/1.0">
  - <teiHeader>
    - <fileDesc>
      - <titleStmt>
          <title>Fichier TEI obtenu à partir du fichier CLAN ANAE-01-1_04_20.cha</title>
        </titleStmt>
      - <publicationStmt>
          <distributor>tei_corpo</distributor>
        </publicationStmt>
      - <notesStmt>
        - <note type="COMMENTS_DESC">
            <note type="@Birth">of CHI: 24-JUL-2006</note>
            <note type="scribe">coder - students (November 2010) - Emmanuelle Mathiot - Françoise Bourdoux (June 2012) 2010)</note>
          </note>
        - <note type="TEMPLATE_DESC">
          - <note>
              <note type="type">-</note>
              <note type="parent">-</note>
              <note type="code">CHI</note>
            </note>
          - <note>
              <note type="type">-</note>
              <note type="parent">-</note>
              <note type="code">MOT</note>
            </note>
          - <note>
              <note type="type">-</note>
              <note type="parent">-</note>
              <note type="code">AEL</note>
            </note>
          - <note>
              <note type="type">-</note>
              <note type="parent">-</note>
              <note type="code">ART</note>
            </note>
```

Figure 7: Colage example 1: Anae corpora format XML



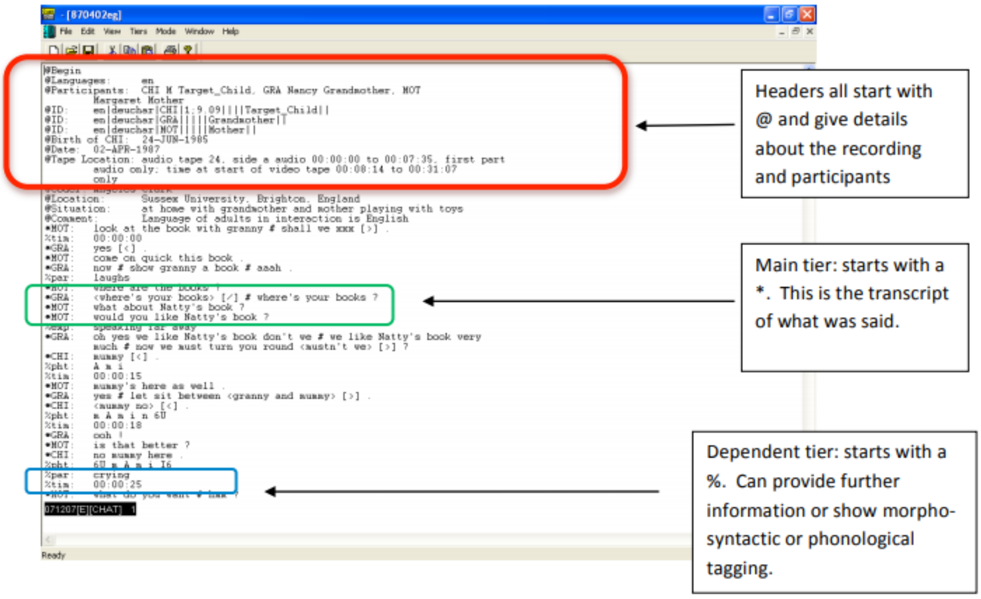
```
      <when xml:id="T821" since="#T0" interval="1842.221"/>
      <when xml:id="T822" since="#T0" interval="1844.438"/>
    </timeline>
  - <body>
    - <div type="Situation" subtype="d0">
      - <head>
          <note type="start">#T0</note>
          <note type="end">#T822</note>
        </head>
      - <annotationBlock xml:id="au0" who="AEL" start="#T0" end="#T1">
        - <u>
          + <seg>
          </u>
        </annotationBlock>
      - <annotationBlock xml:id="au1" who="MOT" start="#T1" end="#T2">
        - <u>
            <seg>+< aujourd'hui on est le sept décembre deux+mille+sept . </seg>
          </u>
        - <spanGrp type="sit">
            <span>CHI se tourne et regarde vers la caméra</span>
          </spanGrp>
        </annotationBlock>
      - <annotationBlock xml:id="au2" who="ART" start="#T2" end="#T3">
        + <u>
        </annotationBlock>
      - <annotationBlock xml:id="au3" who="AEL" start="#T3" end="#T4">
        - <u>
            <seg>Arthur </seg>
          </u>
        </annotationBlock>
      - <annotationBlock xml:id="au4" who="AEL" start="#T4" end="#T5">
        - <u>
            <seg>regarde . </seg>
          </u>
        </annotationBlock>
      - <annotationBlock xml:id="au5" who="MOT" start="#T5" end="#T6">
        - <u>
            <seg>ben xx elle marche pas la télé là . </seg>
          </u>
```

Figure 8: Colaje example 2: Anae corpora format XML



```
@UTF8
@Begin
@Languages:     fra
@Participants:  CHI Anaé Target_Child, MOT  Mother, AEL Ael Brother, ART Arthur Brother
@ID:    fra|Paris-Corpus_Anae|CHI|1;04.20|female|||Target_Child|||
@ID:    fra|Paris-Corpus_Anae|MOT||female|||Mother|||
@ID:    fra|Paris-Corpus_Anae|AEL|4;06.23|male|||Brother|||
@ID:    fra|Paris-Corpus_Anae|ART|8;00.29|male|||Brother|||
@Media: ANAE-1_04_20-240P, video
@Date:  07-DEC-2007
@Location:      Anaé's home
@Transcriber:   students (November 2010) - Emmanuelle Mathiot   - Françoise Bourdoux (June 2012)       2010)
@Time Duration: 1844.438
@@Birth] of C:  of CHI: 24-JUL-2006
@Comment:       Chat
@Situation:     c'est l'heure du goûter, CHI mange une sucette dans sa  chaise haute
*AEL:   Arthur (.) regarde mon pain au lait .0_2730
*MOT:   +< aujourd'hui on est le sept décembre deux+mille+sept .2730_6240
%sit:   CHI se tourne et regarde vers la caméra
*ART:   +< xxx .6240_7410
*AEL:   Arthur.7410_8203
*AEL:   regarde .8203_9151
*MOT:   ben xxx elle marche pas la télé là .9151_10528
*AEL:   +< xx .10528_11145
*MOT:   y+a encore plus de vent qu(e) l' autre jour .11145_13068
*AEL:   Arthur .13068_15621
*AEL:   regarde mon pain au lait .15621_19993
*AEL:   Arthur regarde mon pain au lait . 8.19901_22931
*ART:   oui .22931_23197
%com:   difficile de déterminer s'il s'agit de AEL ou ART
*MOT:   mange le ton pain au chocolat Ael .23197_24396
*AEL:   mangerai après .28798_31321
*CHI:   maman !31321_35661
```

Figure 9: Colaje example 3: Anae corpora format CHAT



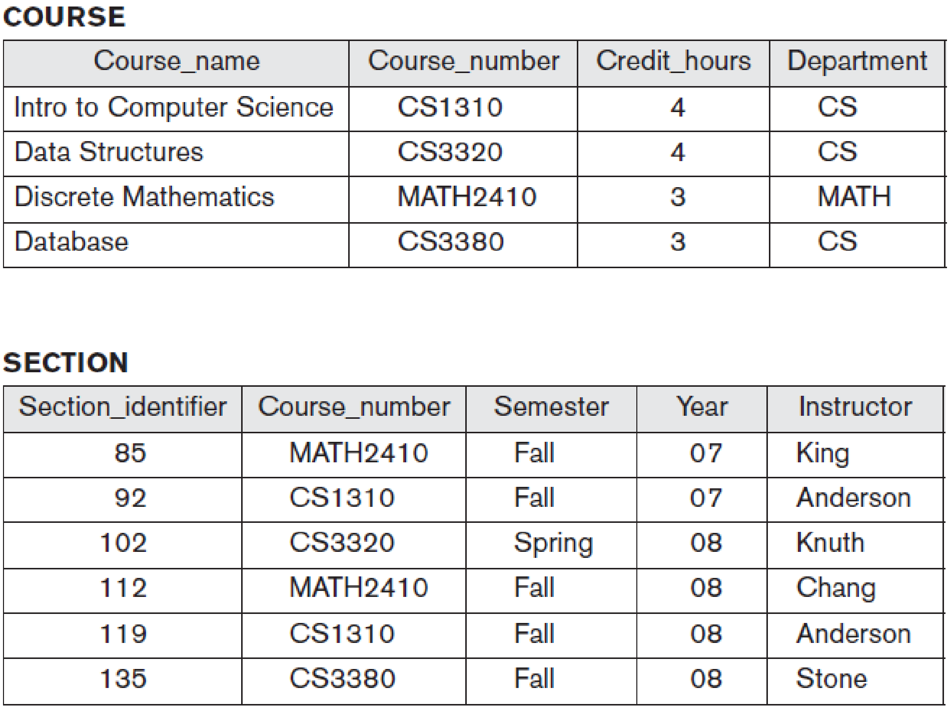
```
@Begin
@Languages:     en
@Participants:  CHI M Target_Child, GRA Nancy Grandmother, MOT
        Margaret Mother
@ID:    en|deuchar|CHI|1;9.09||||Target_Child||
@ID:    en|deuchar|GRA|||||Grandmother||
@ID:    en|deuchar|MOT|||||Mother||
@Birth of CHI:  24-JUN-1985
@Date:  02-APR-1987
@Tape Location: audio tape 24, side a audio 00:00:00 to 00:07:35, first part
        audio only; time at start of video tape 00:08:14 to 00:31:07
        only
@Coder: Angelica Clark
@Location:      Sussex University, Brighton, England
@Situation:     at home with grandmother and mother playing with toys
@Comment:       Language of adults in interaction is English
*MOT:   look at the book with granny # shall we xxx [>] .
%tim:   00:00:00
*GRA:   yes [<] .
*MOT:   come on quick this book .
*GRA:   now # show granny a book # aaah .
%par:   laughs
*MOT:   where are the books ?
*GRA:   <where's your books> [/] # where's your books ?
*MOT:   what about Natty's book ?
*MOT:   would you like Natty's book ?
%exp:   speaking far away
*GRA:   oh yes we like Natty's book don't we # we like Natty's book very
        much # now we must turn you round <mustn't we> [>] ?
*CHI:   mummy [<] .
%pht:   A m i
%tim:   00:00:15
*MOT:   mummy's here as well .
*GRA:   yes # let sit between <granny and mummy> [>] .
*CHI:   <mummy no> [<] .
%pht:   m A m i n 6U
%tim:   00:00:18
*GRA:   ooh !
*MOT:   is that better ?
*CHI:   no mummy here .
%pht:   6U m A m i I6
%par:   crying
%tim:   00:00:25
*MOT:   what do you want # hmm ?
```

Headers all start with @ and give details about the recording and participants

Main tier: starts with a *. This is the transcript of what was said.

Dependent tier: starts with a %. Can provide further information or show morpho-syntactic or phonological tagging.

Figure 10: CHAT format[23]



COURSE

Course_name	Course_number	Credit_hours	Department
Intro to Computer Science	CS1310	4	CS
Data Structures	CS3320	4	CS
Discrete Mathematics	MATH2410	3	MATH
Database	CS3380	3	CS

SECTION

Section_identifier	Course_number	Semester	Year	Instructor
85	MATH2410	Fall	07	King
92	CS1310	Fall	07	Anderson
102	CS3320	Spring	08	Knuth
112	MATH2410	Fall	08	Chang
119	CS1310	Fall	08	Anderson
135	CS3380	Fall	08	Stone

Figure 11: A Database Sample

Bibliography

References

[1] D. Bassano, Production naturelle précoce et acquisition du langage. l'exemple du développement des noms, Lidil (2005) 61–84doi:10.4000/lidil.136.

[2] Les réseaux d'éducation prioritaire, (Accessed December 15, 2020). URL https://mallettedesparents.education.gouv.fr/parents/ID223/les-reseaux-d-education-prioritaire

[3] E. V. Clark, The lexicon in acquisition, Cambridge University Press, 1993.

[4] M. L. Rice, Children's language acquisition, American Psychologist 44 (2) (1989) 149–156. doi:https://doi.org/10.1037/0003-066X.44.2.149.

[5] B. MacWhinney, The child language data exchange system, Journal of Child Language 12 (2) (1985) 271–295. doi:https://doi.org/10.1017/S0305000900006449.

[6] B. Lété, L. Sprenger-Charolles, P. Colé, Manulex: A grade-level lexical database from french elementary school readers, Behavior Research Methods 36 (1) (2004) 156–166.

[7] B. MacWhinney, The CHILDES Project: Tools for Analyzing Talk, 3rd Edition, Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[8] B. MacWhinney, The childes project: Tools for analyzing talk, Behavior Research Methods 3 (2000) 156–166. doi:https://doi.org/10.21415/3mhn-0z89.

[9] Childes, (Accessed December 15, 2020).
URL https://childes.talkbank.org/

[10] A. Morgenstern, Le projet colaje, (Accessed December 15, 2020).
URL http://anr-leonard.ens-lsh.fr/article.php3?id_article=333

[11] Anr; programme formes et mutations de la communication : processus, compétences, usages(document de soumission b), (Accessed December 15, 2020) (2008).
URL http://anr-leonard.ens-lsh.fr/IMG/pdf/Document_B_COLAJE.pdf

[12] Colage corpus, (Accessed December 15, 2020).
URL http://colaje.scicog.fr/index.php/corpus

[13] Leonard. anr. acquisition du langage et grammaticalisation (2005).
URL http://anr-leonard.ens-lsh.fr/

[14] H. Gregory, Y. Rose, Phon 3.1 [computer software], (Accessed December 17, 2020).
URL https://www.phon.ca/phon-manual/getting_started.html

[15] What is a database?, (Accessed December 15, 2020).
URL https://www.oracle.com/database/what-is-database/

[16] Flat file, (Accessed December 15, 2020).
URL https://techterms.com/definition/flatfile

[17] Benefits of databases, (Accessed December 15, 2020).
URL https://www.nibusinessinfo.co.uk/content/types-database-system

[18] relational-databases, (Accessed December 15, 2020).
URL https://www.ibm.com/cloud/learn/relational-databases

[19] R. Elmasri, S. Navathe, Fundamentals of Database Systems, 6th Edition, Addison-Wesley Publishing Company, USA, 2010.
URL https://dl.acm.org/doi/book/10.5555/1855347

[20] What is crud? explaining crud operations, (Accessed December 15, 2020).
URL https://www.sumologic.com/glossary/crud/

[21] Tables - sql server, (Accessed December 15, 2020).
URL https://docs.microsoft.com/en-us/sql/relational-databases/tables/tables?view=sql-server-ver15

[22] About database tables, (Accessed December 15, 2020).
URL https://www.quackit.com/database/tutorial/about_database_tables.cfm

[23] V. Rogers, Childes workshop (2015). doi:https://viviennerogers.info/wp-uploads/2015/04/Childes-workshop.pdf.