Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

Automatic Pronunciation Dictionary Generation from Wiktionary and Wikipedia

Studienarbeit am Cognitive Systems Lab
Prof. Dr.-Ing. Tanja Schultz
Fakultät für Informatik
Universität Karlsruhe (TH)

von

cand. inform.
Qingyue He

Betreuer:
Dipl.-Inform. Tim Schlippe
Prof. Dr.-Ing. Tanja Schultz

Tag der Anmeldung: 01. Juni 2009
Tag der Abgabe: 31. August 2009

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karlsruhe, den 30. November 2009

Abstract

In this work we show that dictionaries from the World Wide Web which contain phonetic notations may represent a good basis for the rapid pronunciation dictionary creation within the speech recognition and speech synthesis system building process. As a representative dictionary, we selected *wiktionary.org* [1] since it is available in multiple languages, and in addition to the definitions of the words many phonetic notations in characters of the International Phonetic Alphabet (IPA) are detectable. We checked the quantity of the pronunciations to vocabulary lists in five languages. Furthermore, a quality check was performed by comparing pronunciations of the dictionary from the World Wide Web to the pronunciations of dictionaries from the *GlobalPhone* project [2] which are commonly employed by the speech community. Paradigm languages are English, French, German, Spanish and Vietnamese. French *wiktionary.org* achieved best results as it included 92.580% pronunciations for GlobalPhone vocabulary as well as 33.333% and 76.119% for lists of international cities and countries. Finally, we are planning to integrate our work into the *Rapid Language Adaptation Toolkit* (RLAT). RLAT is a web based toolkit enabling naive users to create speech recognizers in any language [3].

Table of Contents

1. Introduction

1.1 Motivation

With some 6,900 languages in the world, speech data such as text files, transcribed data or pronunciation dictionaries are not available in all but the most economically viable languages. Therefore, for rapid adaptation of speech recognition systems to new languages, the World Wide Web is used as a source of text data: Websites are crawled to collect texts that are used to build language models. Moreover, prompts which are read aloud by native speakers to receive transcribed audio data are extracted from the crawled text [4]. The production of pronunciation dictionaries is usually time consuming and expensive as they are manually produced by language experts. Thus our intention is to discover if the World Wide Web can be used as a data source of phonetic notations to build dictionaries as well. We also intend to investigate the value of this data resource with regard to quantity and quality of the pronunciations.

If even phonetic notations can be extracted together with the corresponding written words from the World Wide Web, the following economical and automated scenario is imaginable: Text belonging to the domain where the speech recognition system should be adopted is crawled from the World Wide Web and a text normalization is performed. Subsequently, a language model is derived from the collected text. Besides, the vocabulary of the normalized crawled text is extracted automatically, and on the basis of a dictionary from the World Wide Web a pronunciation dictionary with words located in the crawled text is created. The detected words together with their pronunciations are the basis to train a data-driven grapheme-to-phoneme converter which will create pronunciations for words which cannot be found in the World Wide Web. Thereby, dictionary and language model are automatically created without manual effort. Furthermore, prompts which will be read by native speakers to receive transcribed audio data to train the acoustic model are extracted from the normalized crawled text. Figure 1.1 shows the way of generating such a pronunciation dictionary.

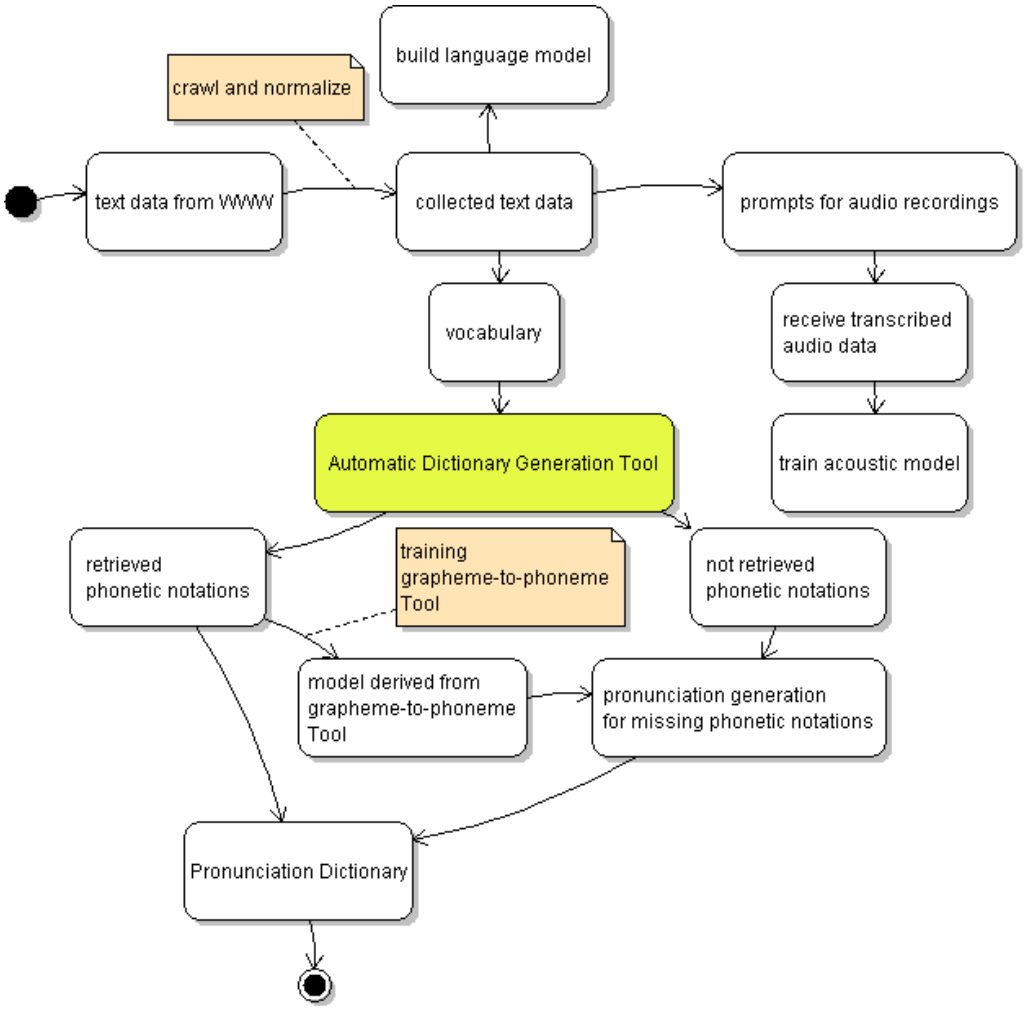

Figure 1.1: Generating a Pronunciation Dictionary in an Automatic Framework for the Building of Speech Recognizers

1.2 Purpose of this Thesis

For our experiments English, French, German, Spanish, and Vietnamese were selected as representative languages since we possess *GlobalPhone* dictionaries in the first four languages which we use as reference pronunciation dictionaries. *GlobalPhone* is a database collection that provides transcribed speech data for the development and evaluation of large speech processing systems in the most widespread languages of the world. In this project speech recognition in 20 languages is investigated [5]. To diagnose the coverage of proper names we also checked the IPA (International Phonetic Alphabet) occurrences for lists of international cities and countries as described in Section 2.4. Proper names can be of very diverse etymological origin and can surface in another language without undergoing the slow process of assimilation to the phonologic system of the new language [6]. One example is the word *Paris*: In French it is pronounced like [paʁi] and English speaking people say /ˈpærɪs/.

We analyzed *wiktionary.org*, a wiki-based open content dictionary, since this free lexical database is available in many languages, and the dictionary entries are checked by an Internet community frequently. The phonetic notations are written in IPA. The

International Phonetic Alphabet is based on the Roman alphabet, devised by the International Phonetic Association as a standardized representation of the sounds of spoken language [7]. Through our experiments we intended to compare the dictionaries from the World Wide Web to the *GlobalPhone* dictionaries by the following checkings:

1. *Quantity Check*

   In order to do a quantity check, we upload a list of words. Our tool automatically looks for each of the words on *wiktionary.org* and checks if IPA pronunciations exist to each of the words. Then the located pronunciations will be extracted from the website and finally there will be a calculation of the number of words from the list to which a pronunciation could be found.

2. *Quality Check*

   For a quality check the located pronunciations from the websites are compared to the pronunciations of a reference list from the *GlobalPhone* dictionaries which we assume to have a reliable pronunciation presentation. Therefore, the pronunciation presentation of the *GlobalPhone* dictionaries needs to be normalized. This is necessary to make a comparison between IPA symbols possible which is explained in Section 4.2.3. In order to get a number of how many words of the compared pronunciations are identical, those are counted and in order to improve our results, we investigated the IPA symbols used on the Internet and in the *GlobalPhone* dictionaries. So we substituted or deleted some IPA symbols which had barely or no influence on the pronunciation of the words, for example the syllable break or the aspirated h.

1.3 International Phonetic Alphabet

As mentioned before, we chose the International Phonetic Alphabet in order to get pronunciation information from the World Wide Web. IPA was developed by a group of French and British language teachers in the late 19^{th} century and was published by Paul Passy in 1888 [8]. There are some reasons why we selected this phonetic alphabet: IPA is designed to represent one symbol for each verbalizable sound and vice versa in all of the languages in the world. This fact makes this alphabet really international as it can be used in every language and can be easily understood. Also, it can help people - whether they are learning or teaching a new language - by using IPA in a simple and intuitive way [9, 10].

Besides, IPA reduces ambiguity in denoting phonemes and is very useful as it is used in many dictionaries, e.g. dictionaries from UK publishers such as *Cambridge, Collins, Longman, Oxford* and also dictionaries from the German publisher *Langenscheidt*. Furthermore other phonetic alphabets are based on IPA, e.g. X-SAMPA. Moreover, it is still the best choice to represent the language independent mapping of sounds [11].

The International Phonetic Alphabet mainly consists of letters from the Roman alphabet, only if not avoidable new symbols were established (see Figure 1.2). The IPA chart contains vowels, consonants, different intonations and diacritics and "needs thus several hundred basic symbols to encompass all languages" [10, 11]. An example

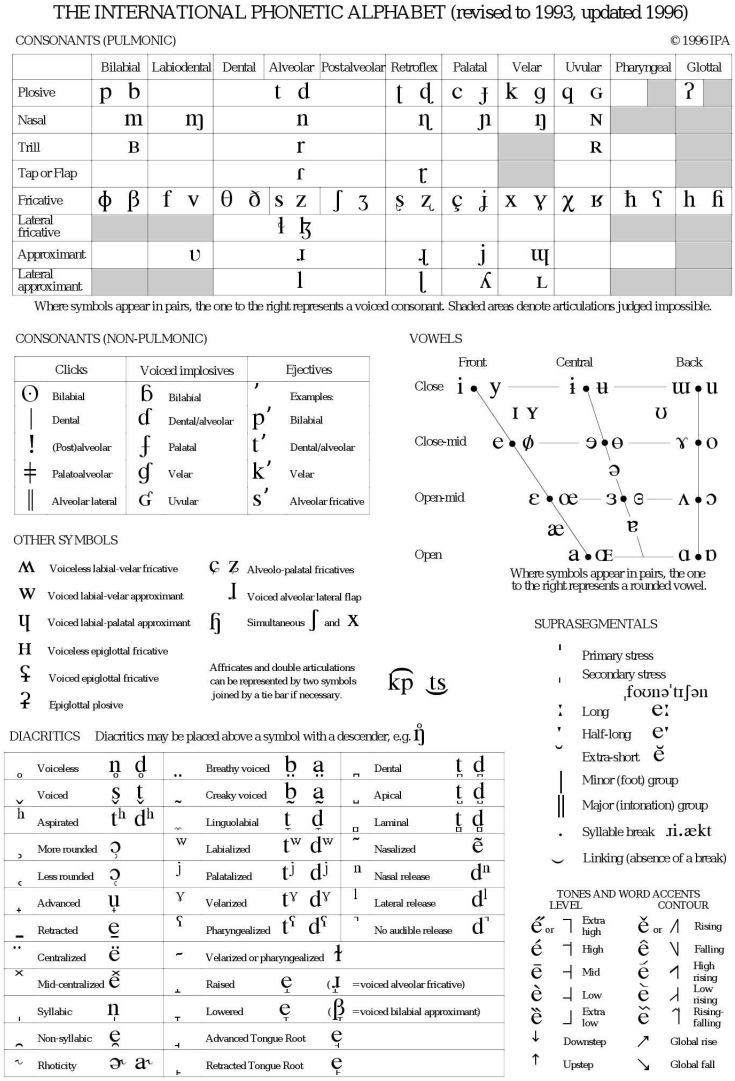

Figure 1.2: IPA chart revised 1993 and updated 1996.

for a word including IPA symbols is [paʁi]. It is an IPA representation of the word *Paris*. For more information about the IPA chart, please visit the official website of the International Phonetic Association [12].

1.4 Pronunciation Dictionary Production

There are two main methods of producing a pronunciation dictionary which are statistical or rule-based, as can be seen in Figure 1.3. The statistical method is either supervised or unsupervised and the rule-based method can be done either manually by linguists or automatically or alternatively semi-automatically such as the LexLearner. The LexLearner is a component in the RLAT which learns automatically with the help of the interaction of a user how words are pronounced [4]. The manual rule-based method is more common but expensive. If the vocabulary size is small the pronunciation production can be done completely manually. If it is a supervised type, a dictionary is already given and rules are made for new entries such

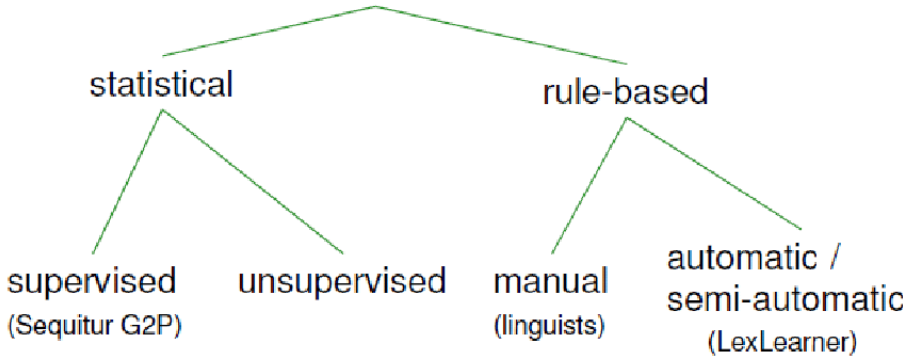

Figure 1.3: Representation of different types of pronunciation dictionary production

as grapheme-to-phoneme rules. These rules are used not only for all new words but also for ambiguously written symbol sequences or proper names which are spelled in different ways depending on the language origin and which may require multiple pronunciations.

1.5 Previous and Related Work

In the field of speech processing the World Wide Web has been used as a data source for improving the language model probability estimation as well as for obtaining additional training material [13].

Furthermore, several approaches to automatic dictionary generation have been introduced in the past. Besling proposes heuristical and statistical methods [14]. Black et al. apply letter to sound rules for the dictionary production [15]. Unfortunately, these methods still require post editing by a human expert or using another manually generated pronunciation dictionary.

In SPICE, a web-based tool for rapid language adaptation in speech processing systems, the Lexicon Learner component presents words to the user, who does not have to be a language expert but is able to provide the pronunciations [4]. Each word is accompanied by a suggested pronunciation, along with a synthesized wavefile. The prediction is based on letter to sound rules that the system infers from the user's answers, which are updated after each additional word. The rules are seeded during an initialization stage in which SPICE asks the user for the phoneme most commonly associated with each letter.

The latest approach which is very similar to our work was done by A. Ghoshal et al. [16]. They use Google's Web and news page repositories to generate pronunciation dictionaries and they focus on web pages in English and from non-European countries. In order to extract, validate and normalize pronunciation information they apply letter-to-phoneme, letter-to-letter and phoneme-to-phoneme rules. Besides, IPA and ad-hoc pronunciations were chosen for their examination, as they say these two kinds were the most commonly found on the Web. Both have their assets and drawbacks, which complement each other: IPA contains special symbols that can describe English phoneme unambiguously, but requires some skills; in contrast the ad-hoc transcription is simpler and does not need any special skills and it follows the rules of English orthography, but ad-hoc transcriptions does not offer phonemic

transcriptions. An example for an ad-hoc transcription for the word “bruschetta” is as follows: “broo-SKET-uh”. As reference an American English lexicon, called Pronlex was used. They concluded that they could build lexica which are comparable in quality to Pronlex, but larger in size [16]. In Section 5.3 we compare the results of Google’s approach and our Automatic Pronunciation Dictionary Generation Tool.

1.6 Structure of this Thesis

In the previous section an introduction in automatic pronunciation dictionary generation from the Internet has been given and the motivation to pronunciation dictionary generation has been explained as well as the reasons for choosing the International Phonetic Alphabet as a pronunciation representation. Also, previous and related work has been presented in Chapter 1.

In Chapter 2, an overview of the data we conducted our experiments with is given. They consist of websites from *wiktionary.org* and *wikipedia.org*, the *GlobalPhone* dictionaries, Vietnamese syllables and lists of words.

Chapter 3 describes the beginning of our researches on *wikipedia.org*, introducing the problems of using these websites.

Chapter 4 explains the different components and source codes of our Automatic Pronunciation Dictionary Generation Tool which we applied to *wiktionary.org*. Procedures of the quantity and quality checks are described.

The results of our quantity and quality checks are presented and interpreted in Chapter 5.

In Chapter 6 suggest how the integration of the Automatic Pronunciation Dictionary Generation Tool into the RLAT may be realized. RLAT is the rapid language adaptation toolkit which itself is also explained in this chapter.

We summarize this thesis with Chapter 7 where we conclude our work and suggest future work.

2. Data

This chapter introduces data and sources which we employed for our researches and presents the motives why we used them. As a major source we looked into *wiktionary.org* in order to get pronunciation information from the World Wide Web. Moreover, the *GlobalPhone* project is presented which provides pronunciation dictionaries. Especially, the pronunciation dictionaries of *GlobalPhone* were of relevance due to the fact that we applied them as vocabulary lists for the quantity check and as reference data in our quality check. Furthermore, different lists will be introduced such as Vietnamese syllables and lists of named entities.

2.1 wiktionary.org and wikipedia.org

For retrieving phonetic notations from the World Wide Web we decided to make use of *wikipedia.org* and *wiktionary.org*.

First, we looked into *wikipedia.org*, a free collaborative encyclopedia project on the Internet which is one of the largest reference websites on the web [17]. Since January 2001 it provides articles and news in over 230 languages with active editions. These are written and frequently updated by volunteers from everywhere in the world [18]. The top 10 of all wikipedias, ordered by number of articles are shown in Table 2.1 [19].

Next, *wiktionary.org* is introduced. The reasons for having the main focus on *wiktionary.org* is explained later in Section 3.

wiktionary.org is also a collaborative project and the companion of *wikipedia.org* for creating a free lexical database in every language, complete with meanings, etymologies, and pronunciations. A big community consisting of users and administrators checks the entries frequently as it is done on *wikipedia.org*. The ten largest Wiktionary language editions of September 2008 are illustrated in Table 2.2 [20]. As can be seen in Table 2.2 the numbers of administrators and users on the English websites are much higer than those on the French websites but nevertheless the French web pages have more vocabulary entries. We assume that this is because English words are more familiar in their pronunciation and spelling than French words and it seems

No.	Language	Articles	Admins	Users
1	English	2,933,299	1,666	10,019,946
2	German	926,052	329	784,867
3	French	823,627	180	629,521
4	Polish	617,338	154	285,053
5	Japanese	599,340	65	324,962
6	Italian	584,048	93	402,152
7	Dutch	545,643	83	246,585
8	Spanish	490,584	136	1,132,871
9	Portuguese	490,478	62	580,502
10	Russian	410,001	73	299,332

Table 2.1: The Top 10 Wikipedia Language Editions, ordered by number of articles (June 2009, meta.wikimedia.org).

that the French language has more words which are pronounced same but spelled different. As a result, the French wiktionary has much more vocabulary entries than other languages in wiktionary.

No.	Language	Vocab. Entries	Admins	Users
1	French	1,125,065	22	9,714
2	English	1,053,934	81	109,494
3	Turkish	251,797	8	4,929
4	Vietnamese	228,098	3	2,759
5	Russian	184,570	4	3,618
6	Ido	141,979	1	607
7	Greek	117,058	7	1,370
8	Chinese	116,358	6	6,091
9	Polish	105,430	18	4,403
10	Tamil	102,389	7	781

Table 2.2: The Top 10 Wiktionary Language Editions (September 2008, meta.wikimedia.org).

All characters in *wiktionary.org* which also include the phonetic notations are in UTF-8 format.

In addition to IPA, pronunciations are partially expressed by the Speech Assessment Methods Phonetic Alphabet (SAMPA) which is a computer-readable phonetic script using 7-bit printable ASCII characters, based on the International Phonetic Alphabet [21]. As many symbols as possible have been taken over from the IPA; where this is not possible, other characters that are available are used. As SAMPA occurs rarely and IPA is more common than SAMPA, we decided to investigate the pronunciations based on IPA.

Figure 2.1 has been created by the wiktionary community and illustrates the article growth within the last years. For some languages (e.g. French (fr), English (en)) the growth was almost exponential, for others it was slower (e.g. Chinese (zh)) and some seem to level off (e.g. Greek (el)). By increasing articles and community members

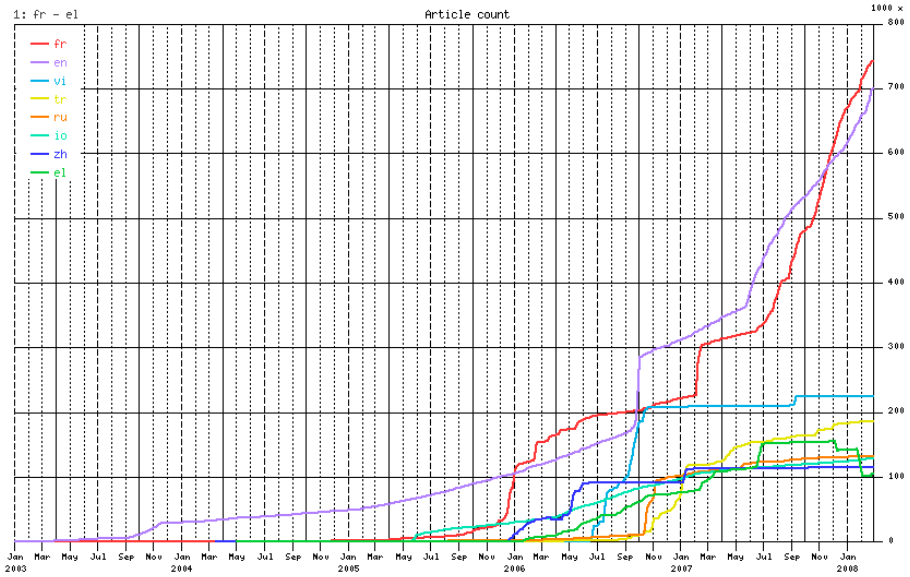

Figure 2.1: Wiktionary Article Growth from Jan 2003 till Jan 2008 (meta.wikimedia.org/wiki/List of Wiktionaries).

it can be expected that the vocabulary coverage will improve and wiktionary pages for more languages will be created.

2.2 GlobalPhone Pronunciation Dictionary

In the *GlobalPhone* project which started in 1995 pronunciation dictionaries in the 20 languages Arabic (Tunesian and Palestine), Bulgarian, Czech, Chinese (Mandarin and Shanghai dialect), Croatian, English (US), French, German, Japanese, Korean, Polish, Portuguese (Brasil), Russian, Spanish (Costa Rica), Swedish, Tamil, Thai and Turkish have been established so far [5]. Widely read national newspapers available on the Internet were selected as resources and texts from national and international political and economic topics were chosen to restrict the vocabulary.

We used the *GlobalPhone* pronunciation dictionaries as references for investigations within quality check which is explained in Section 4 as they cover a huge vocabulary. There already exists over 400 hours of audio data records, spoken by more than 1,800 native speakers [22] of different languages. For each language, 100 adult speakers were asked to read 3-5 articles from newspapers for 20 minutes. Speakers of both sexes, various age categories and different education levels participated in these recordings. As they were mainly generated by native speakers the dictionaries contain authentic pronunciations, which are important for our comparison to pronunciations derived from the World Wide Web. Due to different pronunciation variations the chance of having a match between *GlobalPhone* pronunciation and derived pronunciation from the Internet increases.

As in *wiktionary.org*, the encoding of the *GlobalPhone* pronunciation dictionaries is in UTF-8. UTF-8 is an octet lossless encoding of Unicode charaters.

2.3 Vietnamese Syllables

According to [23], the standard Vietnamese orthography can represent 6,200 syllables (tones included), whereas only about 4,500 to 4,800 are used depending on dialect. We used a list of 5,427 syllables to check the coverage in *vi.wiktionary.org*.

As we know that each Vietnamese word consists of Vietnamese syllables it was worth to do a quantity check with these syllables on *vi.wiktionary.org*, because we would be able to find every word on the Vietnamese websites assuming that we can find every syllable. In fact, currently no *GlobalPhone* pronunciation dictionary exists for Vietnamese. Hence, we exclusively have to perform a quantity check and are allowed to leave the quality check out for this language.

2.4 Lists of Words

In this section we introduce the lists of proper names which we applied to the automatic pronunciation dictionary generation in the studies of this thesis.

First, we employed a list of named entities which is utilized in the program *Spoken Language Communication and Translation System for Tactical Use* (TRANSTAC) of the DARPA (*Defense Advanced Research Projects Agency*) research area. One of the TRANSTAC goals is to present capabilities of two-way speech-to-speech translation systems, which enables English and foreign language speakers communicating with each other in real-world tactical situations where an interpreter is unavailable [24]. This huge list of named entities covers words of different categories, such as Arabic named entities from various classes like Arabic districts, cities, tribes or Arabic first and last names.

The second source was a list from the European project *Quaero* [25], provided by the company *Exalead*. Goal of the *Quaero* project is not only to digitalize European libraries and to translate texts automatically into the language of the searcher but also to retrieve images, videos and audio materials. We used a list containing name, occupation and number of queries about famous people. We left the number of queries out of consideration since we were interested in the names of the people. This list includes names of famous persons like authors, actors, musicians, painters or pianists.

We also decided to look for names of international cities and countries. Each list is translated from English into French, German, Spanish and Vietnamese as far as the words exist in the corresponding language. If the word did not exist in the language we needed, we tested the English word on the respective website instead. We chose 189 international city names from [26] and 201 Country names extracted from [27], as they were available on *en.wikipedia.org*.

As mentioned before, we took vocabulary lists from the *GlobalPhone* pronunciation dictionaries which were also used as references during the process of quality check.

3. First Steps of the Experiments

At the beginning of our experimental phase we looked into source codes of the free encyclopedia *wikipedia.org* [28] which contains pronunciations. Our first idea was to retrieve phonetic notations by grepping the line which contains a term that indicates the existence of a pronunciation, as can be seen in Figure 3.1. Such a term is defined as "tag" in the following text. The grepped line contains not only a tag but also phonetic notations. We deleted the context of phonetic noations which includes tags to get exclusively the pronunciation.

Proper name	Pronunciation + tag	Comment
Uganda	**Uganda** (deutsch: [uˈganda]; englisch: [jʊˈgændə])	Tag "deutsch" before IPA
Paris	**Paris** (frz. [paˈʀi])	Tag "frz."
Joseph Heinrich Beuys	**Joseph Heinrich Beuys** (Aussprache: [bɔɪs];	Tag "Aussprache" + only IPA for family name
Andy Warhol	**Andy Warhol** [ˈændi ˈwɑːɹhoʊl]	No tag
Simbabwe	**Simbabwe** [zɪmˈbapvə] (englisch: ***Zimbabwe*** [zɪmˈbɑːbwɪ];	More than one IPA
Kamerun	**Kamerun** [ˈkaməɾuːn, kaməˈɾuːn] (frz.: **Cameroun** [kamˈʀun]; engl.: **Cameroon** [ˈkæməɹuːn])	More than one IPA (between IPA-brackets) + tag
Schottland	**Schottland** (englisch und *Scots*: **Scotland**, schottisch-gälisch: **Alba** [ˈaləpə], Lateinisch-*keltisch*: **Caledonia**)	IPA does not match "Schottland"

Figure 3.1: Representation of discriminative tags and IPA occurences on *de.wikipedia.org*

We checked single proper names such as "Wolfgang Amadeus Mozart" or "Pennsylvania" and after some testings we spotted that the source codes of web pages on *wikipedia.org* were inconsistent in their syntax (see Figure 3.1 and 3.2). So, we were not able to get pronunciations by grepping tags. Our next idea of obtaining IPA was to get the pronunciations by searching for a set of IPA symbols. We stopped following this idea as we saw that in some cases the retrieved pronunciations did not

correspond to the word we were looking for. Due to the problems, as can be seen in the lists below, it was difficult to find IPA pronunciations on each website correctly. Depending on the language of the web page we looked at, there are different numbers or no pronunciations.

Those websites containing pronunciations are tagged in various order just like "German pronunciation: [ˈjoːzɛf ˈbɔʏs]" or "pronounced /ˈpærɪs/ in English", in any similar way or just without any marking. Some web pages even contained phonetic notations that looked like IPA symbols but were no part of the IPA set, instead those were normal characters from the keyboard. There were also cases, where the IPA on the website did not match the proper name we were looking for, e.g. the pronunciation for "Schottland", the German word for Scotland, matches "Alba" instead of "Schottland". With such irregularities it was not easy to get useful results. So we checked the web pages of *wiktionary.org* and detected thereby a better structure.

Proper name	Pronunciation + tag	Comment
Mozart	**Wolfgang Amadeus Mozart** (German pronunciation: [ˈvɔlfgaŋ amaˈdeus ˈmoːtsart],	Tag: "German pronunciation"
Tony Jaa	**Tatchakorn Yeerum** (Thai: ทัชชกร ยีรัมย์; or formerly **Panom Yeerum** (Thai: พนม ยีรัมย์; IPA: [pʰanom jiːrɑm])	IPA matches original name not searched term
Leonardo da Vinci	**Leonardo di ser Piero da Vinci** (pronunciation (help·info),	Only audio
Zimbabwe	**Zimbabwe** (pronounced /zɪmˈbɑːbweɪ/)	pronounced
Paris	**Paris** (pronounced /ˈpærɪs/ in English; [paʁi] (help·info) in French)	Tag + audio
Uganda	The **Republic of Uganda** (pronounced /juːˈgændə/ or /juːˈgɑːndə/)	More than one IPA
Chopin	**Frédéric Chopin** (Polish: *Fryderyk [Franciszek] Chopin, sometimes Szopen*; French: *Frédéric [François] Chopin*; surname pronounced /ˈʃoʊpæn/ in English; French pronunciation: [ʃɔp ɛ̃]	Polish, French, English
Sophie Marceau	**Sophie Marceau** (French pronunciation: [sɔfi maʁso];	French pronunciation

Figure 3.2: Representation of discriminative tags and IPA occurences on *en.wikipedia.org*

On the English *wiktionary.org* websites phonetic notations were explicitly titled by "**Pronunciation**" and if there was an IPA pronunciation, this was tagged by "**IPA:** ". This concept of structure is also true for the web pages of the other languages we analyzed. Therefore, the titles "**Pronunciation**" and "**IPA:** " have to be translated into the respective language, for example on the French web page phonetic notations are titled by "**Prononciation**" and on the German websites they are titled by "**Aussprache**". Figure 3.3 shows an excerpt of the website *en.wiktionary.org* by the example of the word *lemon* that is marked in the way described before. Thus, *wiktionary.org* has a better structure than *wikipedia.org* and *wiktionary.org* assign a greater importance to pronunciations. Another main difference which should be named here, is the fact that *wikipedia.org* provides more information for proper names. These are information about people, places and their histories, whereas

wiktionary.org represents the wiki-based dictionary with words used in daily life including grammatical background but fewer named entities.

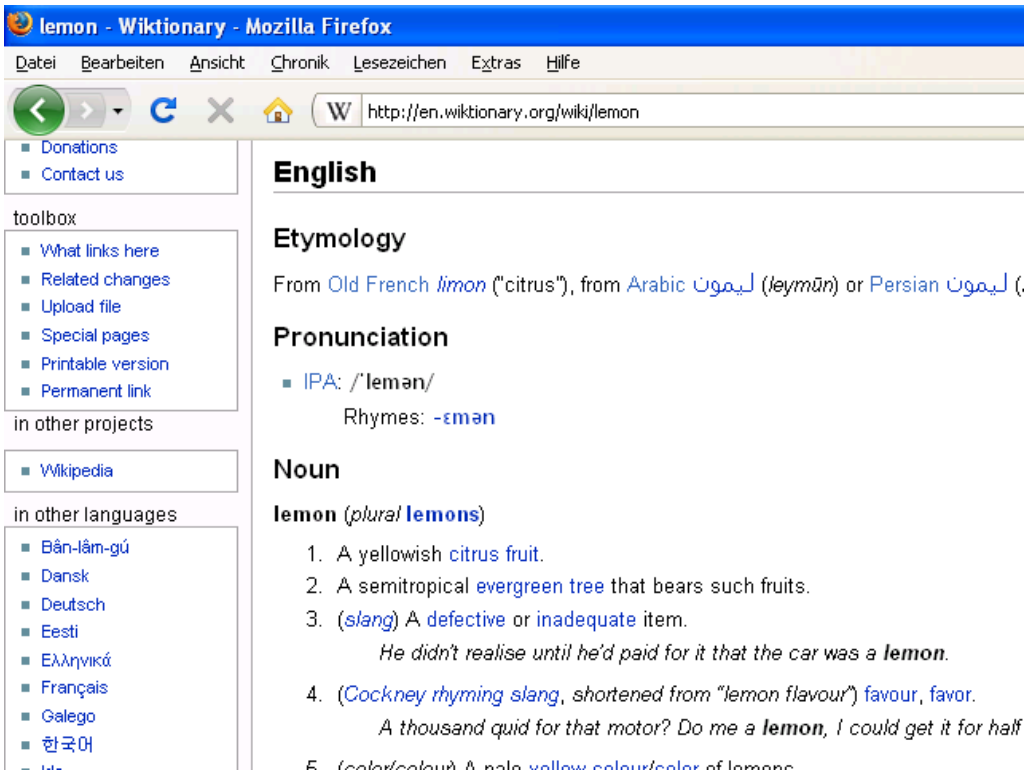

Figure 3.3: Representation of IPA on *en.wiktionary.org* by example "lemon"

Based on the coverage in our named entities checks which can be seen in Table 5.1 in Section 5.1.1, *wikipedia.org* would be the better choice, as we could find more pronunciations of named entities on *en.wikipedia.org* than on *en.wiktionary.org*. But having coverages mostly less than 7% or around 0% was not persuasive enough to continue the investigation on *wikipedia.org*. So we decided to look for more common words which are used in daily life. As such words are available in the *GlobalPhone* dictionaries and as *wiktionary.org* has more pronunciations for those kind of words, we concentrated on *en.wiktionary.org*. We tested *en.wiktionary.org*, *fr.wiktionary.org*, *de.wiktionary.org*, *es.wiktionary.org* as well as *vi.wiktionary.org*, the English, French, German, Spanish and Vietnamese web pages of *wiktionary.org*.

4. Automatic Pronunciation Dictionary Generation

Now we give detailed descriptions of each step of the Automatic Pronunciation Dictionary Generation Tool using the Internet. The steps are:

1. Read in a list of words
2. Crawl for pronunciation of each word in the list
3. Create the result list
4. Check the quantity
5. Check the quality

In the following the scripts of each step are described. Each step is implemented in a different script.

4.1 Scripts for the Automatic Pronunciation Dictionary Generation

1. Read in a list of words:

   `NormalizeWiktionary.pl`:

   Depending on the language the user chooses which is described in Section 4.2.1 the respective website corresponding to the word in the vocabulary list will be downloaded by the Linux function *wget*. At downloading phase the spelling will be normalized, as typing errors can occur and language-dependent letters, such as lower case umlauts at the beginning of a word have to be transformed into upper case. We also check upper, lower and capitalized case of the word as it can be seen in Figure 4.5 and explained in Section 4.2.1. We delete blanks, except one, which happen to appear more than one time when looking for words consisting of two parts, e.g. in words with first and last names. The one blank that is left over is substituted by an underscore "_".

2. Crawl for pronunciation of each word in the list:

   `grepIPA.sh`:

   After downloading the source code of the website `grepIPA.sh` is executed to search the corresponding IPA pronunciation. For each language a different regular expression will be used since the syntax of the website varies from language to language. If no IPA could be found after 4 steps of normalization and crawling, a label with "WARNING" is printed out. This procedure is illustrated in Figure 4.5 and explained in Section 4.2.1.

3. Create the result list:

   `ListRealResults.txt`:

   A file with the list of the words including the corresponding detected IPA or "WARNING" - label will be created. An example of such a list is given in Figure 4.7.

4. Check the quantity:

   `CoverageOfList.sh`:

   This script computes the absolute and relative frequencies of found phonetic notations in IPA by means of `ListRealResults.txt`.

5. Check the quality:

   `RunScript.sh`:

   This script normalizes *GlobalPhone* dictionaries as they do not contain pronunciation in IPA notation. Depending on the language, such as umlauts in German or special accentuation marks above vowels in Spanish are also not represented the way we know them, e.g. ü is represented as ∼u. Moreover, some tags are removed from the *GlobalPhone* dictionaries, for example the silent label "SIL". Those and several other normalizations are explained in Section 4.2.3.1.

   Besides the normalization, `RunScript.sh` performs mappings from the *GlobalPhone* pronunciation notation to IPA notation by the use of the scripts `makeIPA_Qingy.pl` and for French `makeIPAFrench.pl` as the French *GlobalPhone* dictionary has a special pronunciation notation, which is also shown in Section 4.2.3.1.

   `makeIPA_Qingy.pl`:

   This script is an extension of Prof. Tanja Schultz' script called `makeIPA.pl`. Its function is to map each *GlobalPhone* phoneme to IPA and vice versa. On the basis of `makeIPA.pl` we replaced the column "Broadclasses" by IPA symbols. As shown in Figure 4.1, the third column represents the IPA symbols.

   As the French *GlobalPhone* dictionary has some special mappings, these are mapped manually in an extra script called `makeIPAFrench.pl`, instead of using `makeIPA_Qingy.pl`. An excerpt of this mapping table can be seen in Figure 4.2. For example the IPA symbol ʃ is coded in `makeIPA_Qingy.pl` as M_S but on the IPA chart for the French it is FR_SH or another example would be FR_AX in the French IPA chart for ə but there exists no FR_AX or M_AX or ə in `makeIPA_Qingy.pl` as can be seen in Figure 4.3.

```
 "W_r("    => [ ["W_r(",    1],["ɾ",   1],["M_rf",    1],
["--",     0] ],
 "W_d("    => [ ["W_d(",    1],["ɾ",   1],["M_rfd",   1],
["--",     0] ],
 "W_F"     => [ ["W_F",     1],["Φ",    1],["M_F",     1],
["--",     0] ],
 "W_V"     => [ ["W_V",     1],["β",    1],["M_V",     1],
["--",     0] ],
 "W_f"     => [ ["W_f",     1],["f",    1],["M_f",     1],
["AR_F",     1] ],
 "W_v"     => [ ["W_v",     1],["v",    1],["M_v",     1],
["--",     0] ],
```

Figure 4.1: Excerpt of `makeIPA_Qingy.pl` where the column "Broadclasses" is replaced by IPA symbols

```
$line  =~ s/FR_AX/ə/g;#sampa @
$line  =~ s/FR_EU/ø/g;# sampa 2
$line  =~ s/FR_OE/œ/g;# sampa 9
$line  =~ s/FR_AE/ɛ/g;# sampa &
$line  =~ s/FR_A~/ɑ̃/g;# nasal A
$line  =~ s/FR_E~/ɛ̃/g;# nasal E
$line  =~ s/FR_o~/ɔ̃/g;# nasal o
$line  =~ s/FR_OE~/œ̃/g;# nasal OE
```

Figure 4.2: Excerpt of `makeIPAFrench.pl` where the *GlobalPhone* pronunciation notations are replaced by IPA symbols exclusively

4.2 Automatic Pronunciation Dictionary Generation Tool

Our Automatic Pronunciation Dictionary Generation Tool takes a vocabulary list with one entry in each line. For each line, we append the term as a parameter into the URL. If the website can not be found we investigate upper and lower case. Thus, even *wiktionary.org* pages for terms which are written differently with regard to the use of capital and small initial letters or fully capitalized words are located. Each web page is saved and browsed for IPA notations. Output of our tool are the detected IPA notations related to the input vocabulary list plus information about not detected pronunciations.

4.2.1 Upload and Crawl

First, we built a Graphical User Interface which is demonstrated in Figure 4.4. The user uploads a text file which is usually in UTF-8 format, containing the vocabulary list mentioned in the prior section. Before clicking on *uploadFile*, the user selects between five languages (English, French, German, Spanish or Vietnamese) to crawl on the respective *wiktionary.org* web page or otherwise the default setting is English.

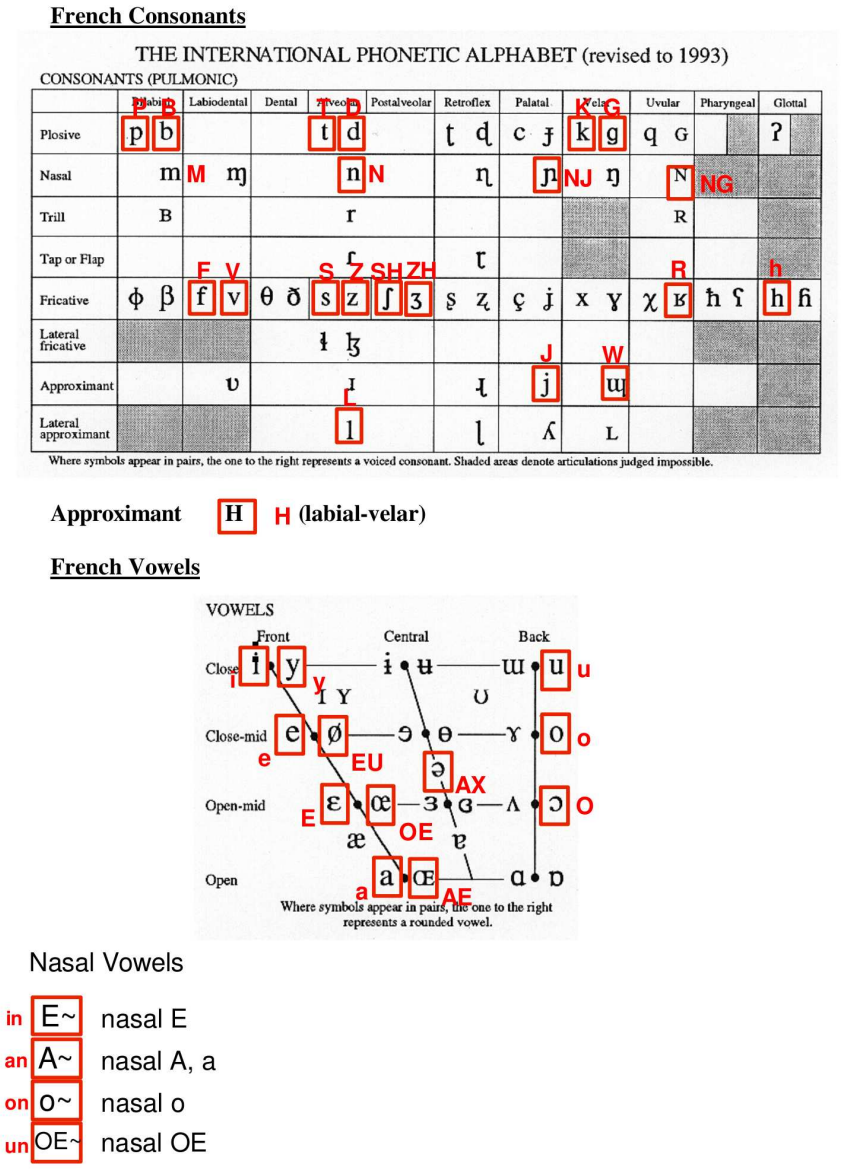

Figure 4.3: IPA chart for French

After chosen the language and pressed the upload button the upload starts. Possible control characters like tabs or blanks at the beginning of each line are deleted. Once the upload is done, the next step is automatically started by downloading each source code of the website of the respective word in the uploaded list. Each line of the uploaded list is processed sequentially. We apply the Linux function **wget** to download each web page and subsequently depending on the chosen language the respective normalization script is performed as each normalization script calls another *wiktionary.org* site.

The normalization takes into account that words in the uploaded list could be written as follows: A word may start with an upper or lower case or can be written completely capitalized or in mixed upper and lower case. This normalization is necessary as we search for the words by attaching the searched term to the web address.

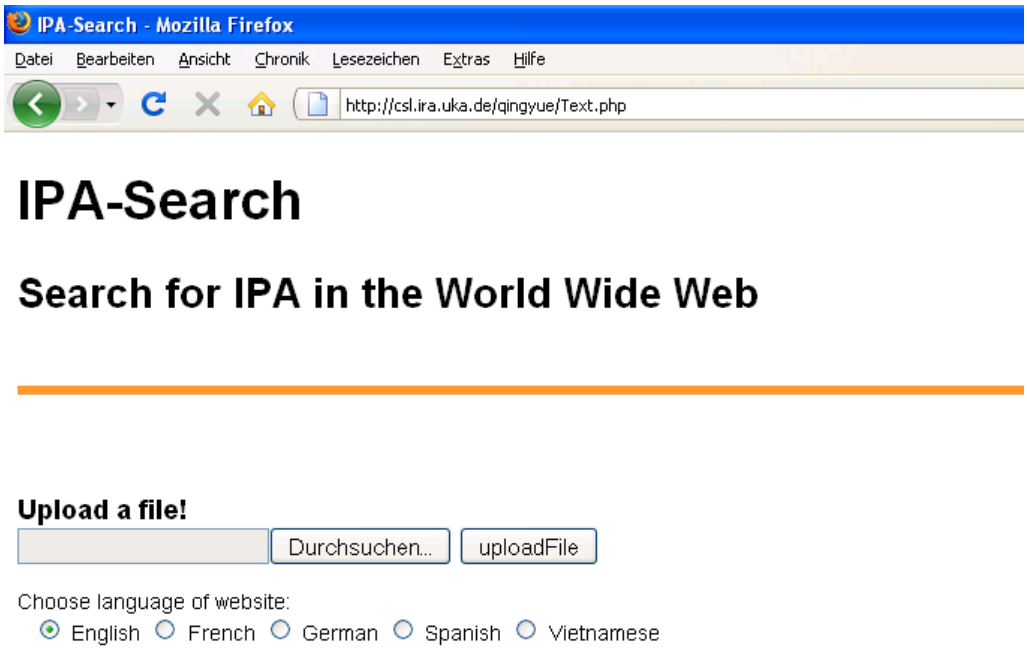

Figure 4.4: Interface of Automatic Pronunciation Creation

For example the web address of the country *San Marino* on the English wiktionary is: *en.wiktionary.org/wiki/San_Marino*.

The normalization is performed in the following order: There can be four steps depending on the right or wrong orthography of the word. First, the source code of the corresponding website is downloaded by using the unchanged orthography of the word from the uploaded file. If a website with IPA-pronunciation can be found by applying the unchanged orthography of this word, the IPA symbols are written into a file called `ListRealResults.txt` 4.7 and the next step of normalization will not be traversed. If not, the normalization script continues with the next orthographical variation: the word is changed into lower case, if this does not work the word is changed into a word with a starting letter in upper case and if even this fails the word is capitalized. If all of these steps fail or if no website exists or no pronunciation can be found on the website a labeling in `ListRealResults.txt` is printed, that outputs a warning (see illustration in Figure 4.5).

For each language different normalization scripts are used which call different *wiktionary.org* site depending on the chosen language and as the German normalization has to be checked for the occurrence of lower or upper case in the beginning of a word. Further, each language will look differently for the corresponding IPA in the source code. This is because the website's source code vary in syntax in each language and also because we need to look for the pronunciations by regular expressions. Hence, each normalization script calls an individual `grepIPA.sh`-script.

4.2.2 Quantity Check

In order to get the absolute and relative frequencies, we need to compute the detected terms with the help of words which are not found. If the file ListRealResults.txt

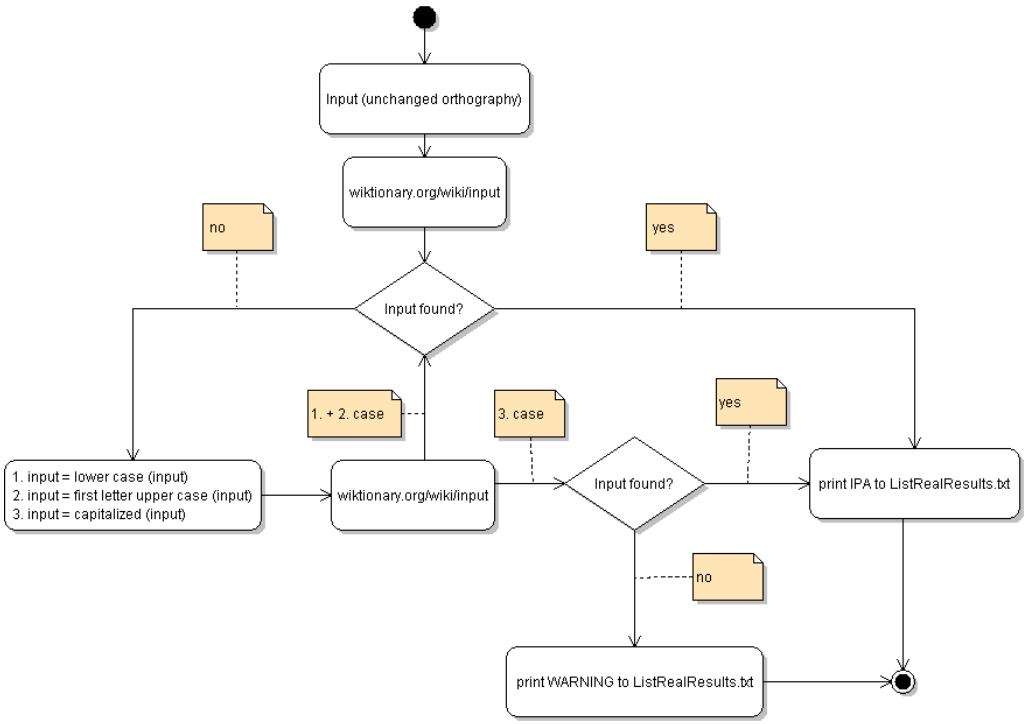

Figure 4.5: Activity diagram of grep IPA

exists, the number of existing lines in this file and the number of lines containing "WARNING" will be computed. Afterwards the number of lines labeled by "WARNING" is substracted from the entire number of all lines in the file ListRealResults.txt and the resulting number is multiplied by 100 to get the percentage. At last it is devided by the number of existing lines. The absolute and relative frequency is printed on the GUI (see example Figure 4.6). On the GUI we can see an output like: "**coverage: 8 / 22 = 36.363 %**", which means IPA notations for 8 of 22 uploaded named entities could be detected. Converted into percentage it is 36.363 % by leaving three decimal places.

4.2.3 Quality Check

After the quantity check we perform a quality check by comparing the pronunciations of the detected IPA notations from the websites of *wiktionary.org* with the pronunciations of the *GlobalPhone* pronunciation dictionaries.

The process of quality check is performed by the script `RunScript.sh` and is devided into two parts:

1. Various normalization steps of *GlobalPhone* phonetic notations which have to be done only once
2. Comparison between the normalized phonetic notations and the phonetic notations, located on *wiktionary.org*

4.2.3.1 Normalization

For quantity and quality checks the goal is to get the same representation for the pronunciations in *GlobalPhone* dictionaries and web-derived pronunciations. IPA

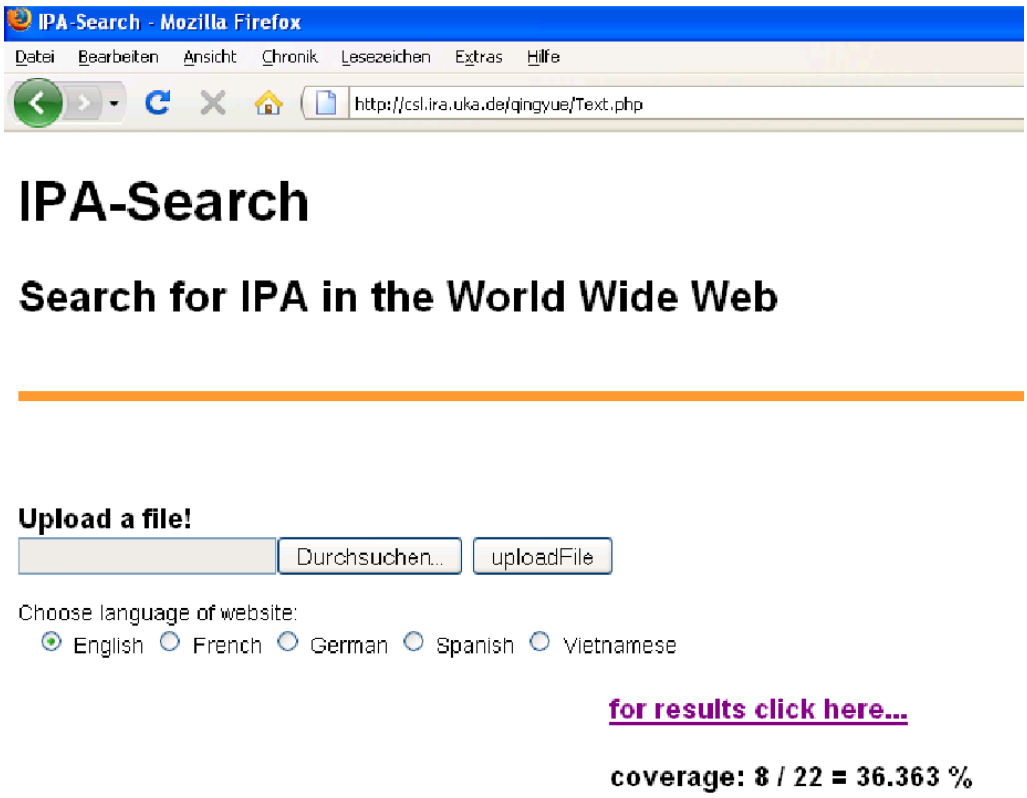

Figure 4.6: Interface of Automatic Pronunciation Retrieval showing link to results and coverage

is selected for this common representation. To achieve this we have to do some normalization steps as the pronunciations of the *GlobalPhone* dictionaries are not in IPA style, which you can see in the following Table:

Language	Word	Pronunciation
English	{FAST} {FAT}	{{M_f/EN WB} M_ale/EN M_s/EN {M_th/EN WB}} {{M_f/EN WB} M_ale/EN {M_th/EN WB}}
French	fragmentées(2) regardons	F R a G M A~ T e Z R AE G a R D o~
German	{Fu~s} {~Osterreich}	{{M_f WB} M_ul {M_s WB}} {{M_oel WB} M_s M_t M_etu M_r M_aI {M_C WB}}
Spanish	mama+ man5ana mie+rcoles	m a m a+ m a n~ a n a m j e+ rf k o l e s

Table 4.1: Examples of pronunciations from the English, French, German and Spanish *GlobalPhone* dictionaries

The first column in Table 4.1 describes the language of the *GlobalPhone* dictionary, the second column presents the words of the extracted examples in each language and the third column shows the pronunciations in a special notation of *GlobalPhone*. Also, you will already have noticed that the notations are not consistent, instead the English and the German lexicon differ from the other two. Therefore, we had to normalize each language individually.

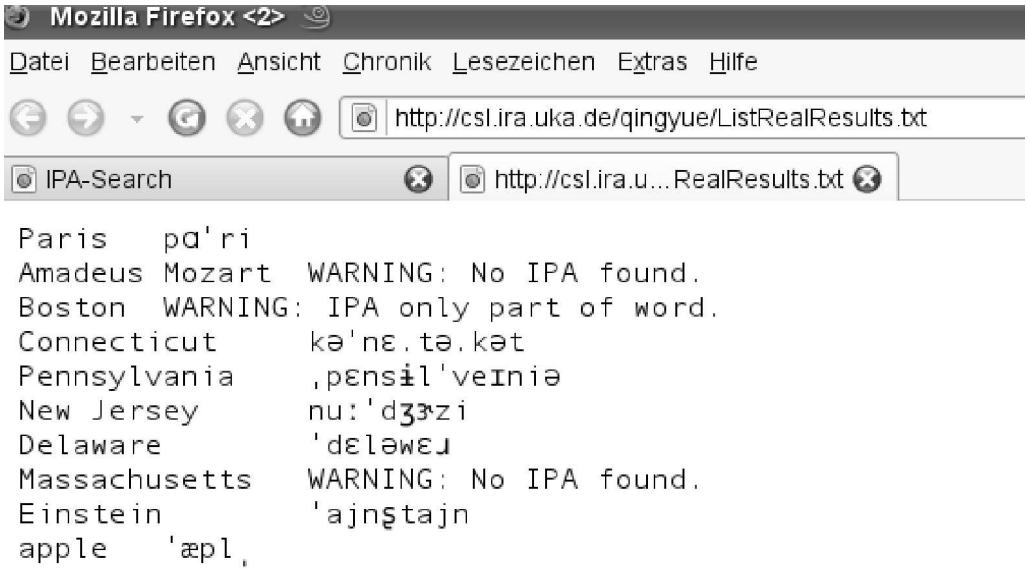

Figure 4.7: Example of `ListRealResults.txt` file

In German we had to transform umlauts from ∼a, ∼o, ∼u to ä, ö, ü (each also in upper case respectively) and the German letter ß which is presented as ∼s.

In Spanish we had to convert the vowels with a plus behind it (e.g. a+) into the stressed vowels á, é, í, ó, ú as well as n5 into ñ.

For French we had to consider the French *nasal vowels* as you can see in the *GlobalPhone* pronunciation "F R a G M A∼ T e Z" which is presented with a tilde ∼. While the IPA notations from the French IPA chart have an overlapping with those from `makeIPA_Qingy.pl` we had to do a special mapping for the French, which is done by the script `makeIPAFrench.pl` that only can run manually otherwise an encoding problem can occur. That means, if the script is executed automatically the identification of IPA symbols are not correct and so the mappings cannot be done. An example for an overlapping is: According to the IPA chart for the French the IPA sign **s** should be mapped to **FR_S** but in `makeIPA_Qingy.pl` the corresponding IPA of **FR_S** is ʃ.

Furthermore, some characters have to be removed from the *GlobalPhone* dictionaries as they are not needed for the mapping of the pronunciations to IPA symbols, such as brackets "{" and "}", "WB", silent marking "SIL" and other labels, like digits with brackets after a word. For example in: "FASTEST(2)" meaning a second pronunciation variant of *fastest*. Moreover, in the French *GlobalPhone* pronunciation dictionary the labeling **FR_** in front of each pronunciation letter is missing and has also to be added. You can see the normalized style of the English words *FAST* and *FAT* below:

```
FAST    M_f M_ale M_s M_th
FAT     M_f M_ale M_th
```

Having removed unused characters the normalized list can be mapped to IPA symbols applying the script `makeIPA_Qingy.pl` or `makeIPAFrench.pl` for French respectively, as you can see in the following example:

FAST f æ s t^h
FAT f æ t^h

After the process of mapping the IPA symbols are still separated by a blank which also has to be cleared away.

4.2.3.2 Comparison of IPA in GPDict and wiktionary.org

As we have obtained the normalized phonetic notations of the pronunciations in *GlobalPhone* dictionary (GPDict), we can compare these with the detected ones from *wiktionary.org* by the Linux command **fgrep -xf**. This command selects only those matches that exactly match the whole line in both files. If there is a match, this pronunciation will be printed to a new file, if not, later will be a new comparison by deleting special notations that usually do not occur either on *wiktionary.org* or in the *GlobalPhone* dictionaries. We choose to remove or substitute the most frequent detected diacritics and suprasegmentals, assuming to get an improvement in the coverage of same phonetic notations between extracted IPA and those from the *GlobalPhone* dictionaries. Such removals are for example the diacritics ^{h} and **:** or the symbol for syllable break "**.**" which nearly do not influence the way of pronouncing a word, especially the removal of syllable breaks. Then the number of matched IPA notations is computed and so we can get an absolute number of same pronunciations in IPA notation.

5. Results

This chapter shows the results of the experiments we conducted with the languages English, French, German, Spanish and Vietnamese with regard to quantity check and quality check. We performed a quality check exclusively for each of the first four languages as still no Vietnamese *GlobalPhone* dictionary exists which we could use as a reference dictionary.

As we started our experiments with lists of named entities on the websites of the English *wikipedia.org*, these results are also displayed in this chapter.

5.1 Quantity Check

In the following two subsections we present our results of the quantity check. The first part is about the researches with named entities on *en.wikipedia.org* which showed decent results for lists of author names, composer names, painter names and country names. We received quite bad results in the field of names of actors, singers, activists, artists and Arabic female first and last names as well as Arabic cities, districts, landmarks and tribes. For some realms proper names could not be found, which let the coverage number drop to zero. Due to this reason we chose to check *wiktionary.org* where we could test more common words such as those given in the *GlobalPhone* dictionaries.

5.1.1 Quantity Check on wikipedia.org

Some results of the quantity check on *wikipedia.org*, made with lists of named entities are shown in Table 5.1. As described in Section 2.4 we mainly had the two categories Arabic person names and place names in English or at least written in Roman letters. We divided the Arabic list into different lists like Arabic female first names or Arabic tribes. For our tests we also divided the list of *Exalead* into different occupation lists, such as authors or painters. In Table 5.1 **Arabic FFirst** means "Arabic Female First Names", **Arabic FLast** means "Arabic Female last Names" and **Arabic Lmarks** stands for "Arabic Landmarks". **No. Hits** is the abbreviation for "Number of IPA Hits" on the respective website and **Coverage** is the "Percentage Number" of **No.**

Hits, for example 4/60 means: Four IPA notations could be detected out of 60 proper names.

The first impression one might get is most of the numbers from *wikipedia.org* are higher than those of *wiktionary.org*. But the percentage numbers of *wikipedia.org* are not the best since they are ranging between 0 and 20 percent with an outlier at 64% for the list of Countries and some results at 0% where no word could be detected.

Since the numbers of detectable IPA notations on *en.wikipedia.org* were higher than those of *en.wiktionary.org* in every check listed below, except one case for *Arabic Female First Names*, we should have done more investigations on *wikipedia.org*. But we decided to do more testings on *wiktionary.org* due to the fact that *wikipedia.org* does not provide websites with a consistent syntax with terms which support us to detect pronunciations, as explained in Chapter 3. Also, the coverage number of most of the lists had results less than 7% which was not persuasive enough to have more investigations on *wikipedia.org*. We also assume that the results of quantity checks with proper names will improve in the course of time as *wiktionary.org* was established after *wikipedia.org* and has therefore not as much entries as *wikipedia.org*. As we noticed that the coverage of the list of country names was enormous high compared to the other lists, we performed quantity checks and quality checks with lists of country names and international cities.

Name of List	No. Hits *Wikipedia*	Coverage	No. Hits *Wiktionary*	Coverage
Countries	32/50	64.000%	31/50	62.000%
Authors	37/176	21.023%	3/176	1.705%
Composers	17/84	20.238%	0/84	0.000%
Painters	15/129	11,628%	0/129	0.000%
Activists	4/60	6.667%	0/60	0.000%
Arabic FFirst	11/200	5.500%	18/200	9.000%
Artists	3/61	4.918%	0/61	0.000%
Musicians	27/603	4.478%	5/603	0.829%
Arabic FLast	2/49	4.082%	2/49	4.082%
Singers	63/1589	3.965%	63/1589	3.965%
Actors	138/4291	3.216%	21/4291	0.489%
Arabic cities	21/1782	1.178%	14/1782	0.786%
Arabic Lmarks	5/1171	0.427%	4/1171	0.342%
Arabic districts	2/796	0.251%	1/796	0.126%
Arabic tribes	1/878	0.114%	0/878	0.000%

Table 5.1: Results of the Quantity Check for lists of Named Entities on *en.wikipedia.org* and *en.wiktionary.org*

5.1.2 Quantity Check on wiktionary.org

The results of our quantity checks on *wiktionary.org* are illustrated in Table 5.2, Table 5.3 and Table 5.4. The quantity check for English *GlobalPhone* dictionary resulted in 16.885% which surprises as English is the most spoken language in the world [22] and therefore might have achieved better results than French. Instead, *fr.wiktionary.org* outperformed the other languages with a 92.580% match. We

suppose the reason for having enormous good results in French is that the administrators of *fr.wiktionary.org* focus more on pronunciation than the administrators of the other languages. As a lot of words in French are written different but pronounced in the same way it might be important for French community to have the words and their pronunciation on the websites. Also, the pronunciation of French words is not as known in the world as for example the English vocabulary due to the reason that more people know how to speak English than French. For the lists of international cities and countries, we detected most phonetic notations in *fr.wiktionary.org*. Due to the fact that the number of vocabulary entries of Vietnamese are the smallest, Vietnamese performed quite well with 76.248% in quantity check. Unfortunately, *es.wiktionary.org* could not accomplish good results as the results of the Spanish quantity check are on the last position. The reason for the low result in Spanish may refer to the pronunciation rules in Spanish as the words are always pronounced the way they are written. Therefore, little motivation for adding pronunciations exist. *de.wiktionary.org* achieved the second best position in the test of international cities and countries.

No.	Language	Vocab. Entries	IPA Coverage
1	French	20,700	92.580%
2	German	41,800	17.624%
3	English	58,585	16.885%
4	Spanish	31,591	8.284%
5	Vietnamese	5,427	76.248%

Table 5.2: Results of the Quantity Check for *GlobalPhone* words (French, English, German, Spanish) and Vietnamese Syllables

No.	Language	IPA Hits	IPA Coverage
1	French	63	33.333%
2	German	52	27.513%
3	English	34	17.989%
4	Spanish	19	10.053%
5	Vietnamese	12	6,349%

Table 5.3: Results of the Quantity Check for 189 International City Names

No.	Language	IPA Hits	IPA Coverage
1	French	153	76.119%
2	German	139	69.154%
3	English	121	60.199%
4	Spanish	33	16.418%
5	Vietnamese	24	11.940%

Table 5.4: Results of the Quantity Check for 201 Countries

5.2 Quality Check

For the quality check we do the following preparations for each language: We extract words which are covered by both *GlobalPhone* and *wiktionary.org*. Then we select

the first pronunciation of each word from *wiktionary.org*. In order to have comparable phonetic notations a mapping from *GlobalPhone* pronunciation characters to IPA characters is performed. Then we compare the first *wiktionary.org* pronunciation for one word to all GlobalPhone pronunciations of the same word. Hits in the quality checks are the lines with the same phonetic notations. As already in our quantity checks French wiktionary shows best results (60.175%) in the quality checks.

English wiktionary shows that it may happen that some words in *wiktionary.org* have pronunciations from one region while other words are pronounced as in another region. For example, exclusively the American pronunciation of `Sri Lanka` (ʃriː.ˈlaŋ.kaː) is possible to find. In contrast, exclusively the British pronunciation of `Glasgow` (ˈglæz.gəʊ) is written. However, there are also web pages with pronunciations from different regions, such as for `vitamin`: ˈvɪtamɪn (UK) and ˈvaɪtamɪn (USA), ˈvaɪtamən (Australia). As we are taking the first IPA pronunciation if there are more than one pronunciation on the website, it can happen that the numbers of the quality check varies. This is due to the reason that the first IPA pronunciation might be not the same as the pronunciation in the *GlobalPhone* dictionaries. More thanone pronunciation can be useful for the automatic dictionary production, as further pronunciations for one word can be added as pronunciation variation to the existing pronunciation in the pronunciation dictionary.

We noticed that phonetic notations in *wiktionary.org* contain suprasegmentals such as stresses (ˈ) and syllable breaks (`.`) which are not regarded in our five *GlobalPhone* dictionaries. For instance, for the word `accepter` in *fr.wiktionary.org* the phonetic notation is "ak.sɛp.te", while in *GlobalPhone* it is "`a K S E P T e`" which is mapped "aksɛpte".

No.	Language	Quality Coverage
1	French	60.175%
2	English	3.761%
	ˈ deleted	4.246%
	. deleted	4.397%
	ˌ deleted	4.418%
	ʰ deleted	10.716%
	ə substituted by ɘ	14.608%
3	German	3.688%
	ˈ deleted	5.484%
	. deleted	5.497%
4	Spanish	3.516%
	ˈ deleted	6.878%
	ʰ deleted	6.916%
	. deleted	42.377%

Table 5.5: Results of the Quality Check

Therefore, we applied the tool *Sclite* which scores and evaluates the output of speech recognition systems with the minimal edit distance. This program compares the hypothesized text to the reference text [29]. In our case the derived IPA pronunciations from the World Wide Web correspond to the hypothesized text and the pronunciations from the *GlobalPhone* dictionaries correspond to the reference text. After the

alignment process statistics related to substitutions, deletions, insertions or correct aligned IPA symbols are computed. We decided to remove and substitute certain diacritics and suprasegmentals which occurred the most in the statistics. So we were able to reach greater matches in our quality checks, as can be seen in Table 5.5. The quality coverage for English increases from 3.761% to 14.608%, mostly by deleting the suprasegmental ʰ and by substitution of ə. There was a slight improvement of nearly 2% for the German check and an enormous increase from 3.516% to 42.377 % for the Spanish check by this procedure. We achieved this big increase by removing the syllable break "." which appears to be in nearly every pronunciation on the Spanish websites. It seems that more modifications of our mapping file are necessary, by adjusting it to the symbols which are used online. This means that on the Internet certain IPA symbols are used in that way they are corresponding to other IPA symbols in our mapping file. For example in the English quality check the symbol ə happens to be used more often on the Internet than the symbol ɘ from the mapping file and in the case of the Spanish quality check more suprasegmentals and tones need to be added in the mapping file or just be considered while mapping.

5.3 Comparison to Google's Web-Derived Pronunciations

In this section our results are compared to Google's approach which was introduced in Section 1.5 [16]. The following table shows the difference of the two approaches.

	Our Approach	Google's Approach
source	wikipedia.org, wiktionary.org	news page repositories, Google's web
language	English, French, German, Spanish, Vietnamese	English from non-EU countries
type of phonetic notation	IPA, GlobalPhone pronunciation notation	IPA, ad-hoc transcription
reference	GlobalPhone dictionaries	American Lexicon Pronlex

Table 5.6: Comparison of our approach to the approach of Ghoshal et al.

As can be seen in Table 5.6 the methods of both research groups are similar. Both approaches include extracting pronunciation information from the Internet and both check how many pronunciations can be found. But we also have some differences which are the selected websites, the analyzed languages and the referring lexica. We did multi-lingual research in English, French, German, Spanish and Vietnamese on the websites of *wikipedia.org* and *wiktionary.org* with the reference of *GlobalPhone* dictionaries by finding IPA pronunciations for lists of words. Google's approach focused on websites from Google's Web and news page repositories in English and from non-European countries when looking for IPA pronunciations and ad-hoc transcriptions (as it is explained in Section 1.5) and also the corresponding orthographic form to those pronunciations. They compared the pronunciations they retrieved from the Internet with an American English lexicon called Pronlex [30].

Google's approach is subdivided in three major phases which are **Pronunciation Extraction**, **Pronunciation Extraction Validation** and **Pronunciation Normalization**.

In the **Pronunciation Extraction** phase they retrieve IPA and ad-hoc pronunciations from Google's Web and news page repositories which are identified as in English and from non-EU countries. In the IPA case, detected IPA pronunciations consist of entirely legal English IPA characters, at least one non-ASCII character, tokens from the web with no punctuation and no HTML markups and the pronunciations are delimited by a pair of forward slashes ("/. . . /"), backward slashes ("\. . . \") or square brackets ("[. . .]"). Additionally, they look for the corresponding orthographic form of the retrieved IPA pronunciation on the same website by using an English phoneme-to-letter (L2P) model to estimate the probability that an orthographic string corresponds to the given phonemic string. With this method they were able to find 2.53M candidate orthographic and phonemic string pairs (309K unique pairs) which is at about six times more than we retrieved altogether both from Wikipedia and Wiktionary. With our lists of the *GlobalPhone* dictionaries we derived about 44.4K IPA pronunciations which is a lot less than those of Google's approach. They collected much more phonetic notations as they could find on a website together with the corresponding orthographic form, while we look exclusively for the IPA pronunciations of given word lists and extract one pronunciation to one given word per website. Google detected ad-hoc pronunciations which match to regular expressions and applied an English letter-to-letter (L2L) model to find the corresponding conventionally-spelled terms. Thereby, they could find 4.52M candidate orthographic and "pronunciation" pairs (568K unique pairs).

In the **Pronunciation Extraction Validation** phase they manually labeled 667 randomly selected pairs of (orthography, IPA pronunciation) and 1,000 pairs of (orthography, ad-hoc pronunciation) in order to check if the extracted items were correct. After a process of computing features and aligning IPA pronunciations with predicted pronunciations and different n-gram letter-pair models for the ad-hoc case, they used a five-fold cross validation on the selected labeled examples to get precision-recall results. They reported that the IPA extraction classifier had a precision of 96.2% when the recall was 88.2%. For the ad-hoc classifier they got a precision of 98.1% when the recall was 87.5%. We did not gain those good results, as our best coverage was in another language - French with 92.58% but we had a bigger set of pronunciations compared to the 667 randomly selected pairs. English with a coverage of 16.885% with words from the *GlobalPhone* dictionary performed not that well on *en.wiktionary.org*. We noticed that the coverage number varies depending on the category we were looking for. For example, our check with country names on *en.wiktionary.org* achieved a quite good coverage with 60.199%. Regarding our quality check we also computed an alignment of found IPA symbols to the reference IPA symbols from the *GlobalPhone* dictionaries with the tool *Sclite*. Having a closer look on the IPA symbol mapping and modifying the mapping after learning from the results of *Sclite* does improve in certain languages as can be seen in Table 5.5.

In the **Pronunciation Normalization** phase Google normalizes the extracted pronunciations to transform them into a standard phonemic form. This is due to the reason that the websites use idiosyncratic conventions and because the ad-hoc pronunciations still are in orthographic form. Therefore, they select a subset of the intersection of Pronlex words and words derived from the Internet as test and training data. Through pronunciation evaluations using alignment methods and computing "Phoneme Error Rates" (which is mostly similar to the tool *Sclite* we applied for quality check) they conclude that "models trained on web data are poor predictors

of Pronlex data, and vice versa". Regarding the website quality, they had the same observations as we did. The websites vary in the quality of pronunciations which can be caused by improper use of IPA symbols and because of other site-specific conventions. But they also state that the quality of the retrieved pronunciations from the Internet rises after the normalization step which lets "the letter-to-phoneme models trained on normalized Web-IPA pronunciations be as good as the models trained on comparable amounts of Pronlex". In comparison to our approach our normalization steps contained other aspects which were necessary to get a common representation between the web-derived pronunciations and the *GlobalPhone* pronunciations.

To summarize, the results of our approach compared to Google's approach could have been better if the number of extracted pronunciations was as high as their approach and therefore better comparable. We could have extracted more than one pronunciation per website to get more data for better comparing with Google's approach. For some languages it might happen that for one word the pronunciation of one region is entered in *wiktionary.org* while another word is pronounced as spoken in another region. This issue could be solved by extracting more pronunciation variants. Also, the language of the website as well as the category of to be searched pronunciations are influencing factors. Analyzing the IPA symbols which are used on the Internet and those which are manually mapped in the *GlobalPhone* dictionaries lead to higher coverages in our quality check which is shown in Table 5.5.

6. A Proposal for the Integration in RLAT

This chapter presents a proposal for the embedding of our Automatic Pronunciation Dictionary Generation Tool into the RLAT which can be found at

`http://141.3.34.5/rlat-dev/index.php`.

Figure 6.1 shows an excerpt of the phoneme selection where the user selects a phoneme set which will be used for speech recognizer creation by labeling the phonemes.

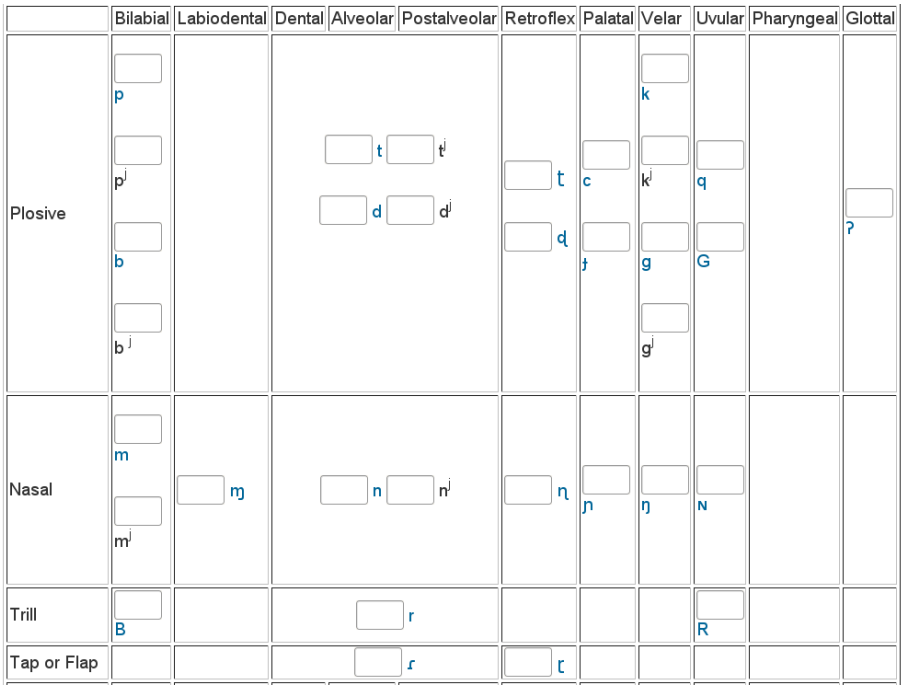

Figure 6.1: Phoneme Selection in RLAT

Figure 6.2 presents an excerpt of the initial grapheme-to-phoneme rules defined by the user.

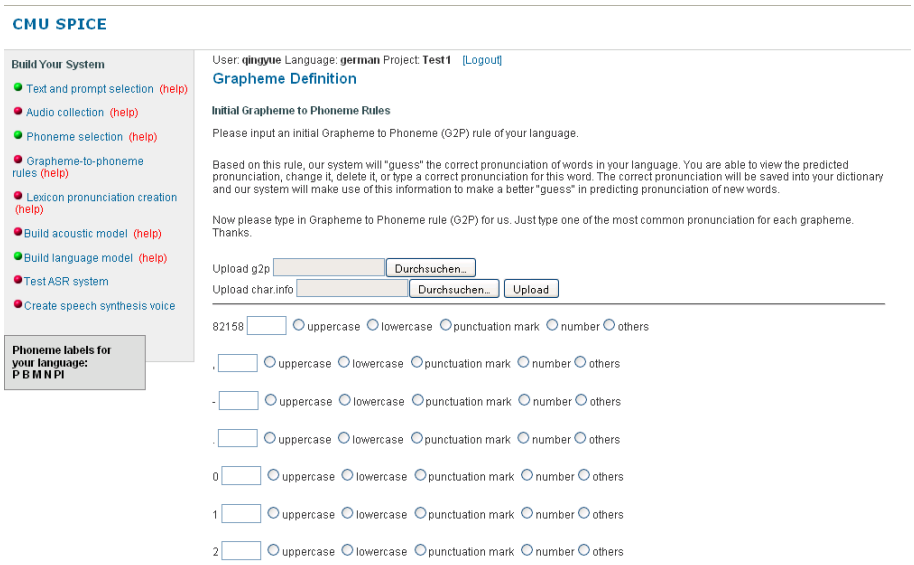

Figure 6.2: Initial Grapheme to Phoneme Rules in RLAT

After having named the phonemes and defined a first pronunciation for each grapheme our approach comes into operation. Our Automatic Pronunciation Dictionary Generation Tool uses IPA symbols as phonetic notations. Therefore, we need a mapping between the phoneme set which was selected and labeled by the user on the phoneme selection page and the IPA symbols detected on the Internet. Then the mapped pronunciation which was retrieved from the Internet by our tool will be presented to the user in his or her selected phoneme representation. The user can make corrections of the pronunciation if needed, for example with the LexLearner [4] and then it will be saved in the pronunciation dictionary. If a pronunciation of the considered word already exists but is not the same as the corrected one, the new pronunciation will be added as a further pronunciation variant into the dictionary.

One of the benefits will hopefully be a better grapheme-to-phoneme learning as existing pronunciations from the Internet are taken into account in the first step and not only the grapheme-to-phoneme rules for only each character which were created by the user in the beginning. Also, we hope the pronunciation dictionary gets bigger as more pronunciation variations are added and it gets more accurate, due to different pronunciations of one word. Pronunciation variants consider different ways people pronounce words influenced by their origin.

7. Conclusion and Future Work

In this thesis a data source from the World Wide Web for pronunciation dictionary creation has been proposed. With our Automatic Pronunciation Dictionary Generation Tool we developed a system which automatically extracts phonetic notations in IPA from the representative dictionary *wiktionary.org* and produces a pronunciation dictionary.

We reported various results for the five paradigm languages concerning quantity and quality checks. For *wiktionary.org* pages which include phonetic notations, pronunciations of a number of words can be extracted which do not need to be created manually.

We assume that the vocabulary coverage will increase by a growing community in the future and even *wiktionary.org* pages for more languages will be created. We also expect better results in quality check by deeper analysis of the retrieved IPA symbols and their mapping to *GlobalPhone* dictionary pronunciations as we definitely could achieve improvements by analyzing the differences with the tool *Sclite*.

After extracting a large number of words together with their pronunciations, we are able to use our data to train a data-driven grapheme-to-phoneme converter which will generate the pronunciations of further words in the corresponding language. The quantity checks with lists of international cities and countries demonstrated that even pronunciations of proper names which might not be found in the phonologic system of a language are detectable together with their phonetic notations in *wiktionary.org*.

Further work will be to investigate the influence of the dictionary generated from the World Wide Web to speech recognition systems.

Finally, we introduced our idea of the integration of our Tool in RLAT and presented the preparatory work that needs to be done before our Tool can be applied. We also explained the benefits of embedding our Tool in RLAT.

Bibliography

[1] "Wiktionary - a wiki-based Open Content dictionary," http://www.wiktionary.org.

[2] Tanja Schultz, "Globalphone: A Multilingual Speech and Text Database Developed at Karlsruhe University," in *Proceedings of the ICSLP*, 2002, pp. 345–348.

[3] "link to the RLAT website," http://141.3.34.5/rlat-dev/index.php.

[4] Tanja Schultz, Alan W Black, Sameer Badaskar, Matthew Hornyak, and John Kominek, *SPICE: Web-based Tools for Rapid Language Adaptation in Speech Processing Systems*, Proceedings of Interspeech, Antwerp, Belgium, August 2007.

[5] "GlobalPhone," http://csl.ira.uka.de/index.php?id=107.

[6] Ariadna Font Llitjós and Alan W Black, "Evaluation and collection of proper name pronunciations online," in *In Proceedings of LREC2002, Las Palmas, Canary Islands*, 2002.

[7] *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, International Handbook Association.

[8] "Omniglot," http://www.omniglot.com/writing/ipa.htm.

[9] "yourdict," http://www.yourdictionary.com/library/ipa.html.

[10] Dirk Van Compernolle, "Recognizing speech of goats, wolves, sheep and...non-natives," *Speech Commun.*, vol. 35, no. 1-2, pp. 71–79, 2001.

[11] Tanja Schultz, *Multilinguale Spracherkennung Kombination akustischer Modelle zur Portierung auf neue Sprachen*, Shaker Verlag, 2000.

[12] "IPA-Association," http://www.langsci.ucl.ac.uk/ipa/index.html.

[13] X. Zhu and R. Rosenfeld, "Improving trigram language modeling with the world wide web," 2001.

[14] Stefan Besling, "Heuristical and statistical methods for grapheme-to-phoneme conversion," in *Konvens*, Vienna, Austria, 1994.

[15] Alan W Black, "Issues in building general letter to sound rules," 1998, pp. 77–80.

[16] Arnab Ghoshal, Martin Jansche, Sanjeev Khudanpur, Michael Riley, and Morgan Ulinsky, "Web-derived pronunciations," in *Proceedings of the ICASSP*, 2009.

[17] "Top 500 sites on the web," http://www.alexa.com/topsites.

[18] "About Wikipedia," http://en.wikipedia.org/wiki/Wikipedia.

[19] "List of Wikipedias," http://meta.wikimedia.org/wiki/List_of_Wikipedia.

[20] "List of Wiktionary editions, ranked by article count," http://meta.wikimedia.org/wiki/List_of_Wiktionaries.

[21] J.C. Wells, "Sampa computer readable phonetic alphabet," *Gibbon, D., Moore, R. and Winski, R. (eds.), 1997. Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter. Part IV, section B.*, 1997.

[22] "Lecture of Tanja Schultz," http://csl.ira.uka.de/fileadmin/Vorlesungen/SS2009/MMMK/slides/PP14+15-MSP-SS2009.pdf.

[23] William Hannas, *Asia's Orthographic Dilemma*, University of Hawaii Press, 1997.

[24] "Intelligent Systems Division," http://www.isd.mel.nist.gov/projects/score/transtac.htm.

[25] "Quaero project and Exalead," http://de.wikipedia.org/wiki/Quaero.

[26] "International City Names," http://en.wikipedia.org/wiki/List_of_national_capitals.

[27] "Country Names," http://en.wikipedia.org/wiki/List_of_capitals_and_largest_cities_by_country.

[28] "Wikipedia - The Free Encyclopedia," http://www.wikipedia.org.

[29] "Sclite - tool for scoring and evaluating the output of speech recognition systems," http://www.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm.

[30] "CALLHOME American English Lexicon (Pronlex)," http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97L20.

[31] T. Schultz and A. Waibel, "Polyphone decision tree specialization for language adaptation," in *Proceedings of the ICASSP, Instanbul*, 2000.

[32] "The GlobalPhone Project," http://www.cs.cmu.edu/~tanja/GlobalPhone/index-g.html.