Chapter 5

Multilingual Dictionaries

Martine Adda-Decker and Lori Lamel

5.1 Introduction

Substantial progress in speech technologies over the past decade has led to a variety of successful demonstration systems and commercial products. In an international context where potential users speak different languages, speech-based systems have to be able to handle multiple languages as well as code-switching and nonnative accents. Multilingual environments are very common with a wide spectrum of potential applications. To increase the usability of speech systems, the challenges of multilinguality and nonnative speech must be addressed efficiently. Porting a given system to another language usually requires significant linguistic resources as well as language-specific knowledge in order to obtain viable recognition performance. Speech recognizers are often quite sensitive to nonnative speech, with notable performance loss when compared to native speech. The two main research directions taken to address this problem have been training acoustic models on nonnative speech to implicitly model the accents, and adapting pronunciation dictionaries to take into account some known characteristics for a given accent.

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

Research in multilingual speech recognition has been supported by the European Commission when dealing with multiple languages (there are now 20 official languages of the European Union, not counting regional languages) and, more recently, by the Defense Advanced Research Project Agency (DARPA) for a relatively limited number of languages (Mariani and Lamel, 1998; Armstrong et al., 1998; Chase, 1998; Mariani and Paroubek, l998; Culhane, 1996; Pallett et al., 1998). Over recent years, there has been growing interest in reducing the costs (in terms of effort and money) to bootstrap the development of technologies for previously unaddressed languages.

The vast majority of the approximately 6,900 languages in the world do not have an acknowledged written form. Only 5–10% of all languages use one of about 25 writing systems (see Chapter 2 for a classification of writing systems).

To date, speech processing has primarily addressed languages for which there is a commonly accepted written form, with the exception of recent studies on dialectal forms of Arabic (Vergyri and Kirchhoff, 2004) and minority languages such as Mapudungun within the Avenue project.[1] This is largely due to the need to represent the language in a written (normalized symbolic) form for further downstream processing. Particularly for automatic speech recognition, the core functionality of a system is the automatic generation of a written representation of speech. However, for other tasks, such as speech-to-speech translation, the written form of a language can be considered less crucial, and ongoing research in this field will show to what extent and under what conditions it will be possible to bypass a written form of the language.

In this chapter, only languages for which written resources are available are considered. For relatively close dialects of standard written languages (for example, some Arabic dialects), automatic transcription may be able to bootstrap off the standard form. For other spoken languages, automatic processing tools may offer help to linguists working to define phonemic and morphological systems, aiding progress toward definition of a writing system.

From the speech recognition point of view, there is generally the need for at least a minimal knowledge of the linguistic characteristics of the language of interest and the means to obtain the necessary linguistic resources.

[1] http://www.cs.cmu.edu/∼avenue

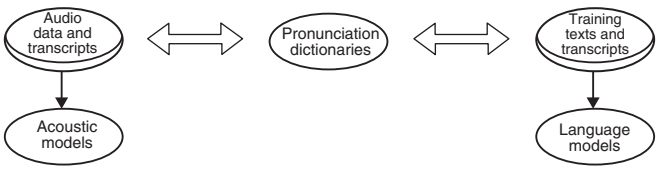


Figure 5.1: Language-dependent resources for transcription systems.

As shown in Figure 5.1, there are typically three primary language resources required for system development: texts for training language models; audio data for training acoustic models; and a pronunciation dictionary. While there is a tendency to treat these related activities as separate research areas (acoustic modeling, pronunciation modeling, and language modeling), there are close links between all three. The transcriptions of the audio data and the language model training texts are typically used in defining the recognition vocabulary; the pronunciation dictionary is the link between the acoustic and language models.

The main focus of this chapter is on multilingual dictionaries for use in automatic speech recognition. There are some common aspects with issues discussed in several other Chapters—in particular Chapter 7, concerning multilingual speech synthesis; Chapters 4 and 6, on multilingual acoustic and language modeling; and Chapter 9, on nonnative speech.

5.2 Multilingual Dictionaries

What is meant by multilingual dictionaries? Can we build a super-dictionary as the union of monolingual dictionaries? What is the intersection of two monolingual dictionaries of two different languages? When looking first at monolingual dictionaries, one can find, in variable proportions, entries from other languages. For example, it is known that the vast majority of words in French are derived from Latin. However, approximately 13% of the French vocabulary is imported from other languages (Walter, 1997). Imported items may be subject to some graphemic assimilation transformations, as shown by the Italian and Germanic examples in

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

ED: Please check if the figure 5.2 is OK.

Figure 5.2: About 13% of French's entries are imported from other languages, mainly English, Italian, and Germanic (after Walter, 1997).

Figure 5.2. More recently adopted English words mainly keep their original orthography, even though it is quite different from the French system. Any living language continually imports items from languages in contact.

Similar observations of multilingual permissiveness in monolingual dictionaries hold for other languages (e.g., Spanish or French entries in English dictionaries, French and English entries in German dictionaries). The adoption of foreign words can even imply a change in the writing system (e.g., Japanese). An interesting problem is raised by proper names for all languages. The same person or the same location may be designated by many different surface forms. For instance, in English news texts the following spellings of Muammar Kadhafi are found: *Kadafi, Kadaffi, Kaddafi, Khadafi, Khaddafi, Khadafy, Khaddafy, Khaddaffy, Qaddafi, Qadhafi, Qadaffi, Qadafi, Qhadafi*.

If languages are close cousins within a family of languages (e.g., Italian and Spanish in Romance languages), a number of identical words can be found. For example, the pairwise intersection of the N in most frequent words in the Indo-European languages French, Spanish, and German results in an overlap of less than 1% for the most frequent 1,000 words in

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

each language, but 10% for the most frequent 20,000 words. Although the vast majority of shared words are proper names and place names, some of the other words (ignoring accents) are *union, region, club, normal,* and *via.*

Concerning the current state of the art in large vocabulary speech recognition, multilingual pronunciation dictionaries are generally collections of monolingual dictionaries that are selectively applied, depending on the identity of the language hypothesized for the speech signal. However, there is ongoing research in speech recognition, speech synthesis, and speech-to-speech translation on how to couple dictionaries from different languages more tightly.

There are three main considerations when designing pronunciation dictionaries: the definition of words in the particular target language, selection of a finite set of words, and determining how each of these words is pronounced. Each of these aspects will typically require a variety of decisions that may be more or less language dependent and have a set of consequences that are interrelated with the two other considerations. The three main aspects of the global language-independent design process are represented in Figure 5.3.

While in many languages the definition of a word may at first appear to be straightforward for written texts (e.g., a sequence of alphabetic characters separated by a space or some other specified marker), for other languages, this is not the case (e.g., Chinese, cf. Chapter 2). Automatic procedures have been successfully used to propose a world-like splitting of the continuous character flow. These are also relevant in the context of language modeling, as further described in Chapter 6.

Word definitions for speech recognition needs to meet two contradictory requirements. On one hand, the number of distinct entries needs to be within reasonable limits, such that good coverage of the system's vocabulary is guaranteed while still enabling the reliable estimation of language model probabilities. This condition favors smaller units. On the other hand, there is a tendency to prefer longer items in order to provide context for pronunciation and acoustic modeling. For these antagonistic criteria, a trade-off is sought, which depends on the amount of available training texts, the limit for the vocabulary size, and the speaking style of the data to be handled. To overcome fixed-size vocabulary limitations, there has been growing interest in open-vocabulary speech recognition, for example, dynamic adaptation of the recognizer vocabulary (Allauzen and Gauvain, 2005b, 2003), which can help reduce errors caused by out-of-vocabulary (OOV) words.

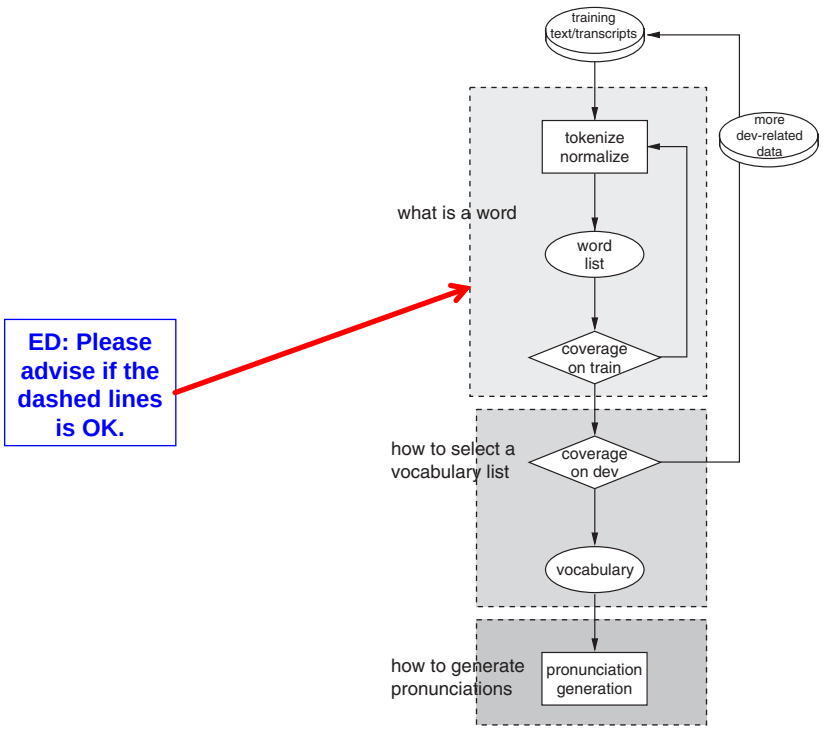


Figure 5.3: Language independent processing steps for pronunciation dictionary generation.

Pronunciation generation can be carried out automatically for languages with a close-to-phonemic writing system (e.g., Italian, Spanish); for others, preexisting pronunciation dictionaries or manual dictionary development are required (e.g., English). The writing system of languages such as Arabic, which specifies only consonants and long vowels, underspecifies the pronunciations, resulting in a high degree of ambiguity. Recent research has addressed using graphemes directly for speech recognition without using an explicit phonemic or phonetic representation of word pronunciations (Billa et al., 2002; Kanthak and Ney, 2003; Killer et al., 2003; Stüker and Schultz, 2004). An important criterion, whatever the adopted method of pronunciation generation, is consistency: if different lexical

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

entries share (partly) the same pronunciation, the phonemic transcription should be (partly) identical.

Some of these considerations are discussed in more detail below.

5.3 What Is a Word?

The question of what is a word seems to be trivial when we are considering languages with a well-established orthography. In general, each of these languages implies a clear sociocultural status, with an education system promoting and relying on a writing standard, and a high production of written language resources. For many spoken languages, however, there are no established writing conventions. For example, in Amharic, any written form that reflects what is said is acceptable (Yacob, 2004). The same applies to Luxembourgish before its writing reform some 20 years ago. For such languages, the orthographic variability and its manifestation at the acoustic level are major challenges for automatic speech recognition. The question of how to define a word for spoken languages with no established orthography is primarily a concern of linguistic research. However, automatic processing may contribute to future progress in addressing such problems.

The following discussion focuses on how to define a word for automatic speech processing in languages with widely adopted writing conventions. Even in this situation, nontrivial questions arise for adopting appropriate lexical units.

A lexicon's word list generally consists of a simple list of lexical items observed in running text.[2] A straightforward definition of a lexical item as a graphemic string between two blanks is too simplistic to be applied without generating a large number of spurious items. Depending on the form of the training texts, different word lists can be derived.

The sample lexicon shown in Figure 5.4 may be obtained from texts like:

> Mrs. Green is a member of the Greens Garden Club.
> Bob Green's car is green.

[2]It can also be a list of root forms (stems) augmented by derivation, declension, and composition rules. This approach is more powerful in terms of language coverage, which is a desirable quality for recognizer development but more difficult to integrate in present state-of-the-art recognizer technology.

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

Normalization 1		*Normalization 2*	
graphemic form	*phonemic form*	*graphemic form*	*phonemic form*
green	gri:n	green	gri:n
Green	gri:n	greens	gri:nz
Green′s	gri:nz	′s	z
greens	gri:nz		
Greens	gri:nz		

Figure 5.4: Sample word lists obtained using different text normalizations, with standard base-form pronunciations.

The Greens all like eating greens.
Green's her favorite color.

Various text forms can be generated by applying different normalization steps to the text corpora (Adda et al., 1997), which may then result in different word lists and pronunciation lexica.

While for many years a case-insensitive text form has been used for large vocabulary conversational speed recognition (LVCSR) in American English (in part due to the availability of common language models in this form) (Paul and Baker, 1992), there has been a move toward maintaining (or re-adding) case to avoid loss of syntactic and semantic information, which can be important for further downstream processing—in particular, named-entity extraction and indexing.

If during text processing case is ignored and the apostrophe (or single quote) is considered as a word boundary mark, the sample lexicon is reduced to the forms shown in Figure 5.4. In English, the apostrophe has a limited impact on lexical variety. It is mainly used to build the genitive form of nouns, but it can also be found in contracted forms, such as *won't, you'd, he'll, she's,* and *we've,* or in proper names (*D'Angelo, O'Keefe*). In general, English lexicons are represented without considering apostrophe as a boundary. For example, in a lexicon containing 100,000 entries, only 4% of the words contain an apostrophe. In contrast, in French, the *apostrophe* is very frequent, occurring in word sequences such as *l'ami, j'aime,* and *c'est*. If all forms containing the apostrophe are included as separate lexical-entries, there is a huge expansion in the lexicon size. Therefore, different text normalizations have to be considered depending on the language's characteristics.

5.3.1 Text Normalization

A common motivation for normalization in all language is to reduce the lexical variability so as to increase the coverage for a fixed-size task vocabulary. In addition, more robust language models can be estimated; however, normalization may entail a loss in syntactic or semantic resolution. Whereas generic normalization steps can be identified, their implementation is to a large extent language-specific. In the following, the most important normalization steps implemented for processing languages such as English, French, German, Spanish, and Arabic are highlighted, and case studies are presented for French and German.

The definition of a word in a language is carried out iteratively, as was illustrated in Figure 5.3. After relatively generic text normalization steps (formatting punctuation markers, numbers), the appropriateness of the resulting words can be measured as the lexical coverage for a fixed-size vocabulary. Depending on these measures, more or less language-specific normalization can be identified and added to the processing chain. Taking, for example, the 65,000 most frequent words in the available processed training data yields a lexical coverage close to 100% for English, but only about 95% for German. This means that with the same type of normalization procedures, German has a much higher lexical variety. The sources of this variety must be identified to efficiently address this problem.

For large-vocabulary conversational speech recognition (LVCSR) applications, some of the most readily available sources of training texts are from electronic versions of newspapers.[3] Much of the speech recognition research for American English has been supported by DARPA and has been based on text materials that were processed to remove case distinction and compound words (Paul and Baker, 1992). Thus, no lexical distinction is made between, for instance, *Gates, gates or Green, green.* In the French *Le Monde* corpus, capital letters are kept distinct for proper names, resulting in different lexical entries for *Pierre, pierre* or *Roman, roman,* for example (Adda et al., 1997). In German, all substantives are written with a capitalized first letter, and most words can be substantivized, thus generating a large lexical variety and homophones. Even so, the overall impact of

[3]While not the subject of this discussion, the text data contain errors of different types. Some are due to typographical errors, such as misspellings (MILLLION, OFFICALS) or missing spaces (LITTLEKNOWN); others may arise from prior text processing.

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

this kind of variability remains small. The out-of-vocabulary (OOV) rate reduction when going from a case-sensitive to a case-insensitive word list in German is only about 0.2% (from 4.9% to 4.7%) with a 65,000 word lexicon.

Different types of text normalization may be explored, depending on the characteristics of the language under study. In order to illustrate problems in word definition, case studies are given for the French and German languages in which a greater variety of graphemic forms are observed than for English. For French, an extensive study on different types of normalization was reported in Adda et al. (1997), using a training text set of about 40 million words from the *Le Monde* newspaper (years 1987–1988).[4] For German, the effect of word compounding on the vocabulary has been studied, and a general corpus-based decomposition algorithm is described.

Case Study I: Effect of Normalization Steps on French Vocabulary

French lexical variety stems mainly from gender and number agreement (nouns, adjectives) and from verb conjugation. A given root form produces a large number of derived forms, resulting in both low lexical coverage and poor language model training. The French language also makes frequent use of diacritic symbols, which are particularly prone to spelling, encoding, and formatting errors. Some of the normalization steps can be considered part of the process of establishing a baseline dictionary. These include the coding of accents and other diacritic signs (in ISO-Latin 1); separation of the text into articles, paragraphs, and sentences; preprocessing of digits ($10\,000 \rightarrow$ 10000) and units (kg/cm^3), as well as the correction of typical newspaper formatting and punctuation errors; and processing of unambiguous punctuation marks. These are carried out to produce a baseline text form. Other kinds of normalization generally carried out include the following:

ED: Please check if this should be 10,000.

N_0: processing of ambiguous punctuation marks
(hyphen -, apostrophe ') not including compounds
N_1: processing of capitalized sentence starts

[4] Evaluating n different types of text normalization entails producing (at least temporarily) n times the training text volume. For this reason, the study has been carried out on a limited subset (40 million words) of the complete training text material available at the time (200 million words).

N_2: digit processing (110 → *cent dix*)
N_3: acronym processing (*ABCD* → *A. B. C. D.*)
N_4: compounding punctuation (*arc-en-ciel* →*arc en ciel*)
N_5: remove case distinction (*Paris* → *paris*)
N_6: remove diacritics (*énervé* → *enerve*)

N_0, N_2, N_3, and N_4 can be considered as "decompounding" rules, which change tokenization (and thus the number of words in the corpus). N_1, N_5, and N_6 keep word boundaries unchanged but reduce intraword graphemic variability.

The elementary operations $N_0 \ldots N_6$ can be combined to produce different versions of normalized texts. Eight such combinations based on the normalization operations $N_0 \ldots N_6$ are shown in Table 5.1. Only the baseline normalizations are used to produce the reference text V_0. The N_0 and N_1 normalizations make use of two large French dictionaries: BDLEX (Pérennou, 1988) and DELAF (Silberztein, 1993) to produce V_1 and V_2 texts. A more detailed description of the normalizations can be found in Adda et al. (1997).

While any normalization results in a reduction of information, the amount of information loss varies for the different types of normalizations. It is straightforward to recover a V_0 text (or an equivalent form) from a V_5 text using some simple heuristics. It is nearly impossible to recover the original V_0 forms from the V_6 and V_7 texts without additional knowledge sources. Furthermore, the V_7 texts seem poorly suited for speech recognition, since a high level of lexical ambiguity is introduced.

Table 5.1 For each version V_i ($i = 0, \ldots, 7$) of normalized text, the elementary normalization steps N_j ($j = 0, \ldots, 6$) are indicated by 1 in the corresponding column.

	N_0	N_1	N_2	N_3	N_4	N_5	N_6	Comment
V_0	0	0	0	0	0	0	0	baseline normalizations
V_1	1	0	0	0	0	0	0	V_0 + ambiguous punctuations
V_2	1	1	0	0	0	0	0	V_1 + capitalized sentence starts
V_3	1	1	1	0	0	0	0	V_2 + digits
V_4	1	1	1	1	0	0	0	V_3 + acronyms
V_5	1	1	1	1	1	0	0	V_4 + decompounding
V_6	1	1	1	1	1	1	0	V_5 + case-insensitive
V_7	1	1	1	1	1	1	1	V_6 + no diacritics

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

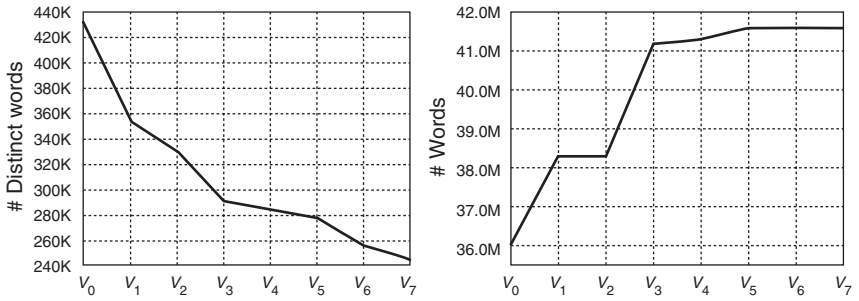


Figure 5.5: Number of distinct (left) and total number (right) of words in the training data for different normalization combination V_i.

Using these different normalization combinations, the number of distinct words and the lexical coverages for each text version can be compared with the training data to help evaluate the relative importance of each step. In the plot on the left-hand side of Figure 5.5, three regions can be distinguished: (1) the region between V_0 and V_3, in which the curve indicates a strong decrease in lexical variety, dropping from 435,000 to 290,000 word forms, (2) the middle region between V_3 and V_5 in which the curve is relatively flat, and (3) the region on the right of V_5, in which the curve gently decreases toward 245,000 word forms. The most important normalization steps are N_0 and N_2, which are case-independent "decompounding" rules. These account for 65% of the gain achieved by the best version V_7. The impact of the decompounding rules on the total number of words in a given text is shown in Figure 5.5 (right); an increase is observed for text versions V_1, V_3, and V_5, where the difference with the previous version is an additional normalization of type N_0, N_2, and N_4, respectively. Figure 5.6 shows the corresponding OOV rates (complementary measure of lexical coverage) of the training data using 64,000 entry lexica (containing the most frequent 64,000 words in the corresponding normalized training data). The OOV rate curve is seen to parallel the #-distinct-word curve of Figure 5.5. A large reduction in OOV rate is obtained for the V_1, V_2, and V_3 text versions, which correspond to the processing of ambiguous punctuation marks, sentence-initial capitalization, and digits. Subsequent normalizations improve coverage, but to a lesser extent.

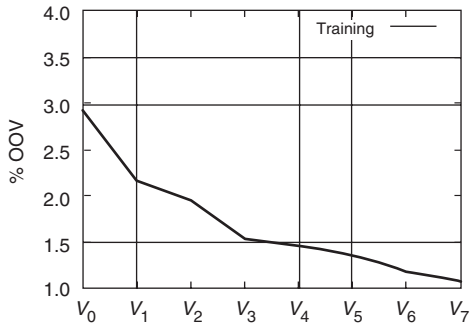


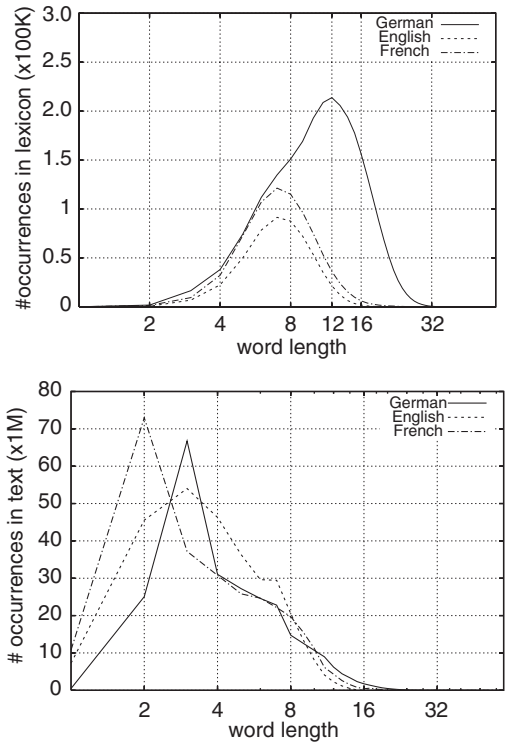
Figure 5.6: OOV rates for different normalization versions V_i on the training data using 64,000 word lists.

This shows the importance of processing punctuation marks and numbers prior to word-list selection. Generally speaking, the influence of ambiguous punctuation-mark processing can be considered as language-specific, whereas number processing is probably important for all languages. If a case-sensitive text output is a desirable feature of the recognizer, capitalized sentence-start processing also has a significant impact on lexical coverage; the other normalization steps turn out to be less important. The V_5 text form achieves a good compromise between standard correct French system output and lexical coverage (Adda et al., 1997; Gauvain et al., 2005). The final choice of a given normalization version has to be chosen as a compromise between the best possible graphemic form and the highest lexical coverage. This compromise is largely application driven.

Case Study II: Effect of Compounding on German Vocabulary

German lexical variety is mainly due to declensions and word compounding. In order to gain a deeper insight into the relative importance of both mechanisms several word-length measures can be compared. Compounding can have a multiplicative effect on word length, whereas declensions typically add just two or three characters.

Figure 5.7 shows the lexical distribution of word lists obtained for German, English, and French in a text corpus of 300 million words per



ED: Please advise if this should be top and bottom.

Figure 5.7: Number of words as a function of length (in characters) for German, English, and French from 300 million words running texts in each language. Number of **distinct entries** in the full **lexicon** (left). Number of **occurrences** in the **corpus** (right).

language. The left curves show the number of lexical entries as a function of word length in the three languages. The English and French curves are quite similar and have a maximum number of entries with a word length around 7. The French curve is higher than the English one, which can be explained by a larger number of distinct surface forms for French verbs. For example, the verb *faire* (to do) has about 40 distinct forms, whereas one of the most productive verbs in English is *to be,* with 8 distinct forms (*be*, *am*, *are*, *is*, *was*, *were*, *been*, and *being*). In German, verb conjugation as well as noun and adjective declension add significant variety to the word lists.

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

For instance, more than 10 distinct inflected forms can be found for the adjective *schnell* (fast) among the 65,000 most frequent entries, including the comparative and superlative forms (faster, fastest). The general slope of the German curve is quite different, with an inflection point around length 8. This characteristic can be related to **word compounding** and suggests that for German, compounding generates a large number of additional lexical entries. The curves on the right show the distribution on the text corpus for each language by word length. It can be seen that French has, in general, the shortest words, with a sharp peak at 2 characters compared to 3 for German and a broader distribution (2–4 characters) for English.

In order to study more closely the importance of compounding as a function of part of speech, the German word list was separated by case, since the words starting with a capital letter are mainly nouns (or proper names). This division clearly demonstrates that compounding involves mainly nouns, and that while nouns account for an overwhelming percentage of the lexicon, their occurrence in the text corpus is much more limited (Adda-Decker, 2003).

As an example, of the 65,000 most frequent words in the corpus, about 100 distinct entries start with the noun *Stadt* (town), for example: *Stadtamt, Stadtarcheologen*, *Stadtautobahn, Stadtbahn, Stadtbaurat*. In the total word list, there are more than 4,000 compounds beginning with *Stadt.*

Automatic Language-Independent Decomposition Algorithm

When developing a speech recognizer for a previously unseen language, it is necessary to assess the lexical variety of the available texts or transcripts: How many different units can be extracted from a given amount of data? How long are these units? If a high lexical variety is measured with a large proportion of long units, there are several reasons to consider reducing the variability. Smaller units will provide better lexical coverage for a given sized word list, easier development of pronunciations, more efficient spelling normalization, and more reliable *N*-gram estimates for language modeling.

German is a well-known example of a language that makes intense use of compounding to create new lexical units—a characteristic shared with other Germanic languages, such as Dutch and Luxembourgish; for the

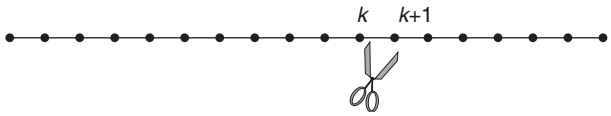


Figure 5.8: Can a word be decompounded after letter *k*.

latter, there is significantly less written material available, so decomposition can be an important processing step. In 1955, Zellig Harris described an algorithm to locate morph boundaries in phonemic strings (Harris, 1955) based on a general characteristic of spoken language: the number of distinct phonemes that are possible successors of the preceding string of phonemes reduces rapidly with the length of that string unless a morph boundary is crossed. This feature is easily transposable to written language: the number of potential distinct letters that are possible successors of a given word start reducing rapidly with the length of the word start.

The written language decomposition problem as illustrated in Figure 5.8 can be stated as: given a word of length K, is there a morpheme boundary between letters k and $k + l$? A straightforward solution is to check whether the decompounded items exist in a language's baseline vocabulary. Table 5.2 gives some example words with multiple possible decompositions. When the boundary is ambiguous, more information is required to make a decision. This information can be easily extracted from the corpus

Table 5.2 Example words with ambiguous decompositions.

compound	⇒	*decompounded*
Fluchtorten	⇒	Flucht-Orten (right)
Fluchtorten	⇒	Fluch-Torten (wrong)
Musikerleben	⇒	Musik-Erleben (right)
Musikerleben	⇒	Musiker-Leben (right)
Regionalligatorjäger	⇒	Regional- Liga-Tor-Jäger (right)
Regionalligatorjäger	⇒	Region-Alligator-Jäger (wrong)
Gastanker	⇒	Gas-Tanker (right)
Gastanker	⇒	Gast-Anker (wrong)
weiterdealt	⇒	weiter-dealt (right)
weiterdealt	⇒	weit-erde-alt (wrong)

Table 5.3 Given a word start *Wbeg*(*k*) of length *k*, the number of character successors #*Sc*(*k*) generally tends toward zero with *k*. A sudden increase of #*Sc*(*k*) indicates a boundary due to compounding. #*Wend*(*k*) indicates the number of words in the vocabulary sharing the same word start.

k	*Wbeg*(*k*)	#*Wend*(*k*)	#*Sc*(*k*)	*Examples*
1	K	147731	62	K**l**asse, K**o**pf, K**r**itik, K**i**nd, K**ö**ln, K**u**rs
2	Ka	29068	41	Ka**m**pf, Ka**n**zler, Ka**i**ro, Ka**u**f, Ka**p**pe
3	Kap	2131	25	Kap**u**ze, Kap**r**iolen, Kap**i**tän
4	Kapi	1281	14	Kapi**e**lski, Kapi**l**laren, Kapi**t**el
5	Kapit	1218	8	Kapit**ä**n, Kapit**e**l, Kapit**u**lation, Kapit**o**l
6	Kapita	974	4	Kapita**l**, Kapita**i**n, Kapita**n**
7	Kapital	968	27	Kapital**i**smus, Kapital**s**, K-**e**rhöhung

by organizing the word list in grapheme-node-based lexical trees. For a given grapheme node at depth k, the higher its branching factor (number of successor nodes), the more reliable is its boundary hypothesis.

As can be seen in the last entry of Table 5.3, which gives the successor information for the word start *Kapital*, at the location of a lexical (or morphemic) boundary, the number of successors significantly increases. This general behavior is schematically depicted in Figure 5.9.

Using this type of analysis, the boundary location for the following two examples can be resolved. The second example is ambiguous. If a word start has more than 10 distinct successor letters, the successor number is displayed.

Pfirsichtorten	$P_{58}f_{25}i_{14}rsich_{17}torten$	Pfirsich-Torten
Fluchtorten	$F_{60}l_{22}u_{25}ch_{15}t_{27}orten$	Flucht-Orten

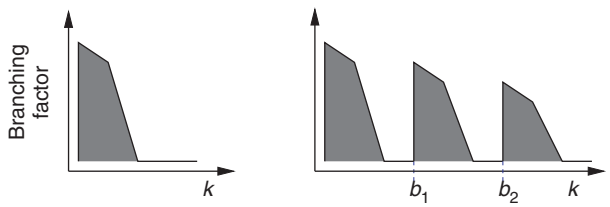


Figure 5.9: Goëlette profile for decomposition: branching factor as a function of length k for a simple word (left) and a three-word based compound (right).

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

Table 5.4 Examples of decomposition rules, including composita with imported English and French items; the number of occurrences of the decompounded items is given in parentheses.

3-word composita			
Theateraufbruchstimmung	→	Theater - Aufbruch•stimmung	(29205 - 788)
Schmerzensgeldanspruchs	→	Schmerzens•geld - Anspruchs	(1001 - 336)
Schlechtwettereinbruchs	→	Schlecht•wetter - Einbruchs	(65 - 253)
German-English composita			
Programmhighlights	→	Programm - High•lights	(38977 - 653)
Lightgetränk	→	Light - Getränk	(638 - 562)
Imagezuwachs	→	Image - Zuwachs	(7279 - 3890)
English-English composita			
Streetlights	→	Street - Lights	(6522 - 97)
Lightdesigns	→	Light - Designs	(638 - 232)
Breakthrough	→	Break - Through	(811 - 59)
German-French composita			
Weltraumrendezvous	→	Welt•raum - Rendez•vous	(1815 - 622)
Avantgardezeitungsprojekt	→	Avant•garde - Zeitungs•projekt	(2559 - 61)
Luxusboutiquen	→	Luxus - Boutiquen	(2564 - 410)

For the first word, the compound boundary can be unambiguously placed after the word *Pfirsich* with the locally highest branching factor. Similarly, in the second example, there is a large increase in branching factor after the letter *t,* indicating that the boundary should be placed after the word *Flucht*.

Table 5.4 gives examples of decompounded items for some typical German 3-word compounds and for compounds mixing German, French, and English items. The results of a one-step decomposition are shown; that is, only one boundary is located. Remaining boundaries are indicated with a •. The decompounding algorithm can be applied iteratively. After each iteration, the word lists, lexical trees, and successor-node information are updated. Decomposition rules already extracted for shorter items can be applied to partially decompounded items. The resulting decomposition can be represented hierarchically, as shown in Figure 5.10. It provides a semantic structuring, which may be useful for certain applications, such as translation and indexing.

Table 5.5 highlights the importance of decomposition as a normalization step for compounding languages. The larger the vocabulary size, the higher,

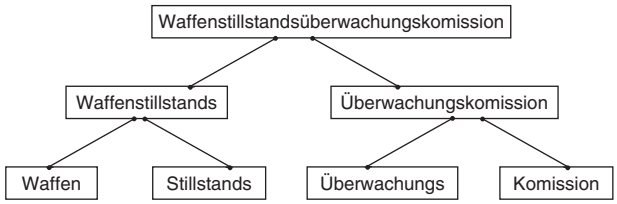


Figure 5.10: Hierarchical representation for a complex decomposition.

Table 5.5 Lexical coverage and complementary OOV rates measured for different-size vocabularies on a 300-million word German text corpus. Measures are given with and without decomposition. The last two columns indicate the absolute and relative gains in OOV reduction rates.

Vocab.	$\%cov_{orig}$	%OOV	$\%cov_{decomp}$	%OOV	Δabs	Δrel
65k	94.8	5.2	96.0	4.0	1.2	23
100k	96.1	3.9	97.2	2.8	1.1	28
200k	97.6	2.4	98.5	1.5	0.9	37
300k	98.3	1.7	99.0	1.0	0.7	41
600k	99.0	1.0	99.5	0.5	0.5	50

the relative OOV reduction rate. The OOV rate of a 300,000 vocabulary on 300 million words of training data is about 1% with the decompounded text version, whereas the original text OOV rate is close to 2%.

The decomposition algorithm presented here is language independent and only requires large corpora. It can thus be straightforwardly adapted to any language to minimize the impact of compounds on lexical coverage. Beyond its utility for tuning lexical coverage in a given language, the knowledge of lexeme (and morpheme) boundaries may be important for pronunciation generation. This point is addressed later in the chapter.

5.4 Vocabulary Selection

Recognizer vocabularies—that is, word lists for LVCSR—are generally defined as the N most frequent words in training texts. This guarantees

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

optimal lexical coverage on the training data. As the resulting word list depends heavily on the content of the training material, it is not necessarily optimal under testing conditions. For large vocabulary speech recognition, a major requirement for the word list is high lexical coverage during testing. In order to achieve this, the training text materials should be closely related (in time and topics) to the test data. In the following, the dynamic properties of living languages are discussed, and some measures highlighting the importance of epoch adequacy between training and test data from similar sources are presented. This is followed by a discussion of spoken language specific problems, the differences between spoken and written language, and how word-list development can accommodate these. Finally, prospective developments for multilingual dictionaries are presented.

5.4.1 Vocabulary Changes Over Time

Research on automatic transcription of broadcast news speech has highlighted the importance of word lists keeping pace with language usage across time. Diachronic word list and language model adaptation is a research area of its own (Allauzen and Gauvain, 2005a; Federico and Bertoldi, 2004; Khudanpur and Kim, 2004; Chen et al., 2003, 2004). The usage of a word can decay or increase with time, and completely new items may appear. An existing word can, at a given moment, be boosted by an important personality (*abracadabrantesque*, used by French President Chirac), by new techniques (*toile*, "*net*"), or by its usage in another language (the English words *road map* has boosted the usage of the translation *feuille de route* in France, which has become very popular in political speeches but also in everyday conversations). New items (neologisms and proper names) may also be introduced. In the last ten years, new items such as *européiste, solutionner*, *cédérom*, *Internet*, and *cyber-café* have appeared, which originated either in morphological combinations of existing items or as a result of new technological developments. However, in a system's word lists, most new items correspond to proper names.

5.4.2 Training Data Selection

It is common practice to use a set of development data in order to select a word list representing the expected test conditions. In practice, the selection

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

of words is done so as to minimize the system's OOV rate by including the most useful words. In this context, *useful* refers to (1) being an *expected input to the recognizer*, and (2) being *trainable for* ***LMs*** *given the training text corpora.* In order to meet the latter condition, one option is to choose the *N* most frequent words in the training corpora. This criterion does not, however, guarantee the usefulness of the lexicon, as stated by the first requirement. Selection or weighting of the training data can be a step in this direction.

For transcription of general news, problems of lexical coverage can appear if the training corpora are either too small or too remote in time from the test data. To illustrate this problem, French is again used as an example, although similar behavior could be expected for any other processed language. In order to measure the lexica1 coverage under similar training and test conditions a development set (dev_{96}) of about 20,000 words was extracted from the *Le Monde* newspaper from the month of May 1996. The impact of training corpus size and epoch on lexical coverage was measured by defining two additional training corpora: $T_{87\text{–}95}$ and $T_{91\text{–}95}$. The training text sets compared are:

$T_{87\text{–}88}$: 40 million words from 1987–1988
$T_{94\text{–}95}$: 40 million words from 1994–1995
$T_{87\text{–}95}$: 185 million words from 1987–1995
$T_{91\text{–}95}$: 105 million words from 1991–1995 of more recent data

Figure 5.11 compares OOV rates using 64,000 word lists (containing the most frequent words) obtained on the $T_{87\text{–}88}$ and $T_{94\text{–}95}$ training sets to the OOV rates on the dev_{96} data. For the word list derived from the $T_{87\text{–}88}$ training texts, the OOV rates for the dev_{96} set are significantly higher than those for the training data for all text versions. In contrast, for the word list derived from the $T_{94\text{–}95}$ training texts, the OOV rates on the training and dev_{96} sets are quite similar. This comparison measures the impact of training data epoch using a constant amount of training material, and illustrates the need for up-to-date data. As mentioned before, an important proportion of the word list consists of proper names related to current events, which are strongly time and topic dependent.

Figure 5.12 shows that the use of larger and more recent training texts ($T_{87\text{–}95}$ or $T_{94\text{–}95}$) significantly reduces OOV rates on test data. The OOV

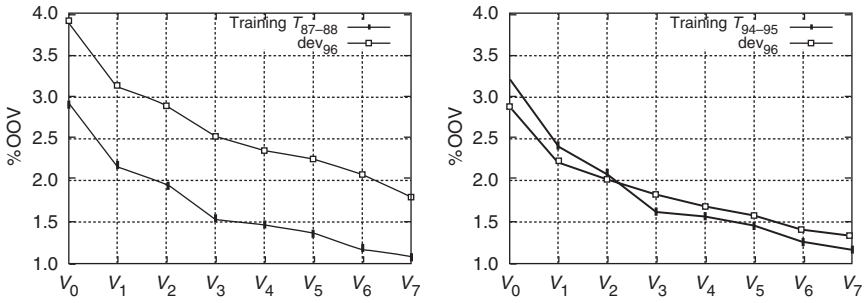


Figure 5.11: OOV rates on training and dev_{96} data for different normalization versions V_i and 64,000 most frequently words from 40 million training data highlighting the importance of training epoch. **Left:** $T_{87–88}$ training data (years 1987–1988). **Right:** $T_{94–95}$ training data (years 1994–1995).

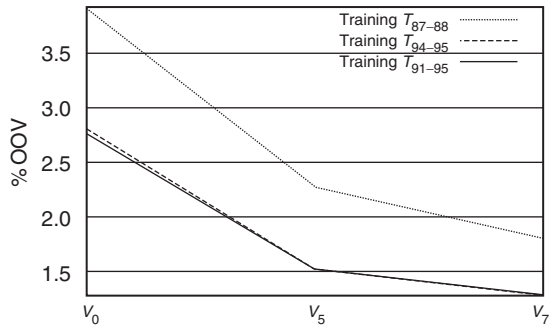


Figure 5.12: OOV rates for normalization versions V_0, V_5, and V_7 on dev_{96} text data, using 64,000 word lists derived from different training test sets.

rates for the $T_{94–95}$, $T_{87–95}$, and $T_{91–95}$ curves are close to 1.5% for V_i, versions ($i > 0$), with the absolute difference from the $T_{87–88}$ curve of close to 1%. The time proximity between the training and test data is more important than the use of additional older data to minimize the OOV rate. Appropriate selection of training material for a given test condition is also seen to be more important for reducing the OOV rate than some of the elementary normalization steps (compounding, punctuation, case sensitivity).

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

Optimized training-data selection can be carried out by weighting recent training texts more than the older text material. This optimization can even eradicate the effects of some minor normalizations (Adda et al., 1997).

5.4.3 Spoken Language-Specific Vocabulary Items

When processing spontaneous speech, there are additional considerations that should be taken into account. Since it is generally easier to have access to written texts than to have transcriptions of spoken language, such texts constitute the vast majority of language model training material. However, there are important differences between written and spoken language, including the vocabulary items, their frequencies, and reduced forms found essentially only in speech (Boula de Mareil et a1., 2005). For example, first- and second-person (singular and plural) pronouns and related verb forms are much more frequent in speech transcripts than in written texts. Thus, among the top words in speech transcripts are words such as *I, we, you, am,* and *are* but also hesitations and discourse markers. However, numbers and acronyms have a relatively low usage in speech transcripts as compared to news text sources. Speech transcripts are therefore particularly important for adapting language models for spontaneous speech.

There are also a number of abbreviated words or clipped words in spoken language that are rarely found in standard written texts. The use of clipped words seems to be quite language dependent, and is frequent in vernacular French. Some examples of general clipped words in French are *appart* (*appartement*), *aprèm* (*aprèsmidi*), *cata* (*catastrophe*), and *compèt* (*compétition*) (Huot, 2001). Similar items, though less frequent, can be found in spoken English, e.g., function words such as *cuz* (*because*) but also some content words, such as *vet* (*veterinarian*), *corp* (*corporation*), and *deli* (*delicatessen*).

Some of the processing specific to spoken language concerns frequent word sequences, corresponding to function words, letters of acronyms, or words composing a date or a complex number. Important temporal restructuring can be observed here (Adda-Decker et a1., 2005); in particular, short words may change significantly or even disappear. Hence it is common practice to represent a limited number of acronyms as distinct lexical entries (as opposed to a sequence of individual letters) and to

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

represent some frequent word sequences subject to reduction as compound words (Gauvain et al., 1996, 1997, 2002; Finke and Waibel, 1997; Ma et al., 1998; Stolcke et al., 2000). There are different strategies for selecting these items, ranging from simply including the most frequent *N*-word sequences, to including all *N*-word sequences containing a certain set of words, to complete manual specification based on linguistic knowledge or observed reductions.

Speech transcripts, if produced thoroughly, also contain disfluencies, such as hesitation items and word fragments (incompletely uttered word starts). Depending on the application, it can be interesting to map all hesitations or filler forms such as *uh, hmm, hum*, and *uhm* to a single unique form under the hypothesis that only the fact that there is a hesitation is important, and that the particular manifestation is a personal choice and therefore unimportant. In other situations, this information may be of interest since it can be indicative of the speaker or of the language being spoken or even of the native language of a speaker using a different language (Candea et al., 2005). Such forms may also serve as back channels during communication and in some languages (e.g., English) indicate agreement (*uh-huh*) or disagreement (*uh-uh*). Word fragments are generally ignored in word lists as well as singletons here (words occuring only once), which are likely to be errors.

These examples illustrate some major differences between spoken and written language, and are meant to underline the crucial importance of using speech transcripts as training data.

5.4.4 Multilingual Considerations

For all languages, recognition vocabularies are generally composed of:

- function words (hundreds)
- general content words (thousands)
- technical content words (thousands)
- proper names (millions)
- foreign proper names (millions)

These word classes are listed by decreasing frequency of occurrence and an appropriate number of types in each class is shown in parentheses.

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

These different types of words raise different problems for modeling and particularly for pronunciation modeling, which is addressed in the following section. Whereas function words and general content words are strongly language-specific, technical words and proper names tend to be shared more easily (after accounting for some writing convention adaptations across languages). In order to give an idea of the number and types of lexical entries shared among languages, the number of common entries among the top *N* words in recognizer word lists were compared in pairs, for the French, Spanish, German, and English languages. Results are shown in Figure 5.13. If for the top 50,000 words 10,000 words are shared, this represents 20% of the word list. This proportion is almost achieved for the English-French and the English-Spanish pairs. As can be expected with a higher top *N* limit, the share word percentage increases since the proportion of technical items and proper names becomes larger. Of course, shared proportions depend on the language pairs and the type of corpus. For the same type of speech corpora, English and French shared more words than German and Spanish (see Figure 5.13, left). When full word forms are compared, the German language shares the lowest number of entries with the other languages. A 50,000 word list is not large enough here to include many technical words or proper names, as declension, conjugation, and—more importantly—word compounding produce many distinct general language entries. Figure 5.13 (right) compares a Luxembourgish word list, extracted from parliamentary debates, to French, German, and English. The curves are relatively similar to the left part of the figure except for French, which is known to be largely used in official speech. However, the curves indicate that the proportion of shared proper names is smaller: the Luxembourgish corpus's proper names mostly refer to national personalities, whereas for the other languages, proper names taken from broadcast news data include more international names.

Some frequent words with common orthography in French and English are *but, or, son, me, mine*, *met, as, fond, sale, sort, note, type, charge, moment, service*, and *occasion.* Shared entries may be identical only in their surface forms or may share some of their meanings. Shared entries with some common meanings are *me, charge, moment, type, service,* and *occasion*, but others have entirely different meanings. The word *sale* in French means "dirty," the equivalent of the English *sale* being *soldes*; the French word *son* means "his," the English to French translation of *son* being *fils*.

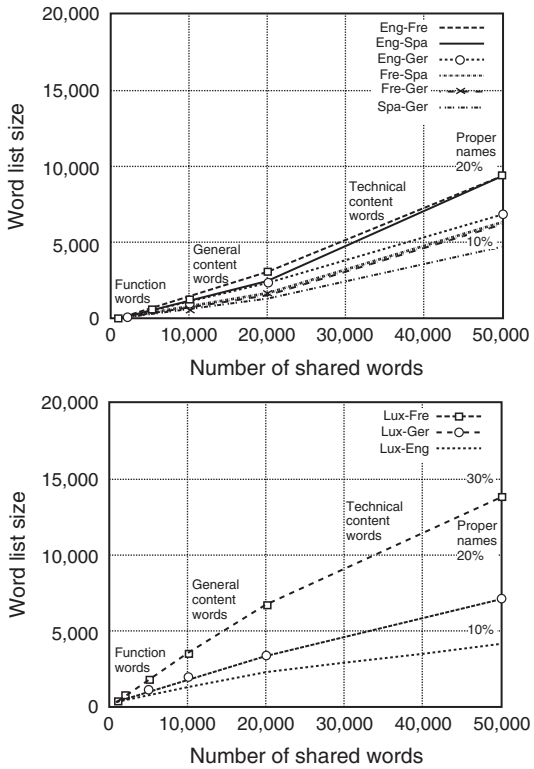


Figure 5.13: Word list comparisons between pairs of languages. The number of shared words is shown as a function of word list size (including for each language its N most frequent items). **Left:** language pairs are among English, French, Spanish, and German. **Right:** language pairs are Luxembourgish versus French, English, and German.

ED: Please advise if this should be top and bottom.

To date, multilingual dictionaries have been investigated for a few applications in which the language of the user may not be known in advance (Shozakai, 1999; Micca et al., 1999; Übler et al., 1998). Typically these applications are very task-specific, which entails relatively small vocabularies. Languages can be processed separately by different language-specific systems in parallel, or by one single "polyglot" system applying multilingual acoustic models as further discussed in Chapter 4.

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

It is likely that future automatic speech and text processing algorithms will be based on huge multilingual word lists of millions of entries (for languages from a given family that share similar writing conventions). Depending on the type of text normalizations (e.g., removal of accents and diacritics), a significant part of the vocabulary will then be shared among languages. Language-specific lexical entries can typically be limited to some tens of thousands of items. These aspects, though not yet addressed—in the context of open-domain speech recognition—offer new perspectives for multilingual automatic processing.

5.5 How to Generate Pronunciations

In order to correctly recognize an utterance of a given word, the corresponding acoustic word models must take into account the observed variations in the acoustic signal. Acoustic feature extraction and acoustic modeling techniques (see Chapter 4) provide powerful means either to reduce the variability or to take it properly into account. ASR is not, however, a bottom-up process, and the contribution of language models is very important in ranking acoustically similar candidates (see Chapter 6). The more discrimination among words is provided by the language model, the less discrimination needs to be provided by the acoustic word models.

Word pronunciations specified in the recognition dictionary provide the link between sequences of acoustic units (phones) and words as represented in the language model. Whereas spoken and written sources are relatively easy to collect for major languages, pronunciations are generally not directly available. Generating word pronunciations for ASR requires a modicum of human expertise and thus cannot be carried out fully automatically. However, if the primary purpose is to transform the acoustic signal into a word string without additional annotations, thus the question arises as to whether acoustic models can be directly linked to graphemic units, rather than using phone-based acoustic models. Given the close relationship in many languages between the graphemic and phonemic form, there has been growing interest in bypassing the explicit step of pronunciation generation in favor of using graphemes directly for speech recognition. However, the relation between graphemes and phonemic forms may be at least locally ambiguous. French and English are examples of languages

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

with a high proportion of ambiguous grapheme-phoneme relations. For example, the English grapheme sequence *ough* can be pronounced as /ʌf/, /o/, or /u/ depending on the carrier word (*rough, thorough, through*). Word context can help in resolving this ambiguity to a certain extent. Conversely, the English phoneme /f/ can be written as either *f*, *ff*, *ph*, or *gh*. French writing conventions include letters that carry information about words' etymological origins and that are mute with respect to pronunciations. Mute constonants in French word endings are very common. The sound /o/ can be written as *o*, *au*, *eau*, *ô*, *oh*, *aux*, *ault*, *eaux*. The word *est* ("is" or "east") can be pronounced as /e/, /ɛ/, or /ɛst/ or even /ɛstə/; the letter sequence *ent* can either be mute or be pronounced as IPA schwa symbol (ə)/ã/ or /ɛ̃/. The corresponding grapheme-based acoustic models need to implicitly include all these variants and share parts of them among different graphemic units. Grapheme based acoustic modeling has been successfully adressed for different languages by several teams, including Bisani and Ney (2003), Billa et al. (2002), Killer et al. (2003), Kanthak and Ney (2003), and Schultz (2004). It has the clear advantage of being straightforward and fully automatic and is of particular importance for rapid porting of an existing recognition system to a new language. However, ambiguous letter sequences necessarily produce ambiguous acoustic models, which is a drawback if the lexicon and the language model information are not able to solve the ambiguity. As seen in previous sections, a language is a living entity composed of a relatively small kernel of language-specific items (function and general content words). A huge number of items with low occurrences in the language are composed of grapheme sequences that escape language-specific regularities (thus, the importance of the Onomastica Project discussed later in the chapter). For a more detailed discussion of grapheme-based acoustic modeling in a multilingual setting and a comparison languages across, the reader is referred to Chapter 4, Section 4.3.

For languages with a close correspondence between writing and pronunciation conventions, ASR can be carried out without pronunciation generation. However, explicitly specified pronunciations allow spoken language to be modeled more accurately. A pronunciation-based approach includes the potential for reducing the ambiguity of a given language's writing system. Beyond its importance for speech recognition, a pronunciation-based approach contributes a finer tuning of oral dialogue components; to the development of educational and medical services (L2 acquisition,

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

orthophony) (Seneff et al., 2004; Flege, 1995); and to research in linguistics (phonetics, phonology, dialectology, sociolinguistics) (Gendrot and Adda-Decker, 2005; Durand and Laks, 2002).

Most state-of-the-art ASR systems use phone-based representations for acoustic modeling.[5] The strength of phone-based approaches is that acoustic word models can be built for any word, even if it has never been observed in the acoustic training data. The weak point is that phone-based pronunciations are fixed a priori, meaning they are not optimally integrated in the acoustic-model training process. Other units have been explored to model some well-known contextual factors that affect the acoustic realization of phones, ranging from demiphones to demisyllables, syllables, and automatically selected subword units (Holter and Svendsen, 1997; Jones et al., 1997; Marino et al., 1997; Pfau et al., 1997; Tsopanoglou and Fakotakis, 1997; Bacchiani and Ostendorf, 1998; Kiecza et al., 1999). Studies have addressed factors such as syllable position, lexical stress, and coarticulatory influences of the neighboring phones (Schwartz et al., 1984; Chow et al., 1986; Lee, 1988; Shafran and Ostendorf, 2003; Lamel and Gauvain, 2005).

Once a recognition vocabulary has been selected, it is necessary to generate a pronunciation for each entry. This process can be decomposed into several independent steps, as shown in Figure 5.14. The first step consists of producing **canonical pronunciations** for each lexical entry. This can be done by (1) relying on master dictionaries for the language under consideration, by (2) an automatic letter-to-sound module if the language has a relatively unambiguous writing system with respect to pronunciations, or by (3) manual specification of pronunciations by a human expert. In practice, a combination of different methods are chosen.

Next, pronunciation **variants** are added, both the canonical pronunciations and their variants being specified as phone sequences. Different types of variants may be introduced depending on the precision of the phone set. The choice of the phone set is also important for acoustic modeling as discussed in Section 5.5.3. When adding variants, one has to consider the types of speech that will be processed in order to add relevant pronunciation variants for genre and style. Is the speech formal in style (e.g., broadcast

[5]The following discussion focuses on phone-based pronunciation dictionaries. The notion of phone and phone inventory is described in the section addressing pronunciation variants.

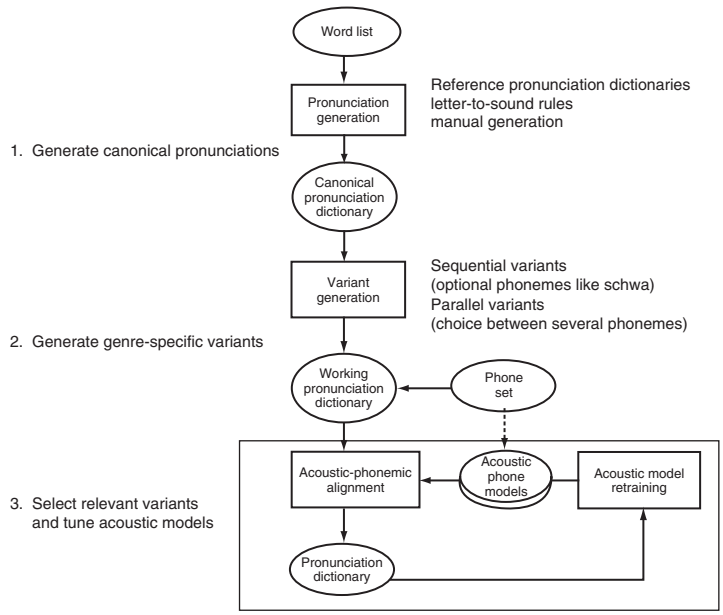


Figure 5.14: Pronunciation dictionary development for ASR system.

speech)? If so, the pronunciations will tend to remain close to canonical forms, and few variants are required. Is the speech vernacular? In this case, phonetic changes can be observed both within words and across word boundaries (Labov, 1966), which implies the need for a higher proportion of pronunciation variants, and in some cases, even a change of lexical unit definition. ASR researchers working on both styles of speech in different languages have become aware of the significant differences between both speaking styles at all modeling levels: word lists, pronunciations, and acoustic and language models.

After generating a basic set of pronunciations, an **acoustic corpus-based validation process** can be carried out, which aims at selecting the variants that are useful given the acoustic models and the type of speech under consideration. Different variants can be assigned probabilities, e.g., based on their occurrence in a training corpus. However, as word occurrences follow a Zipf distribution, which seems to be language universal, only a few words will have reliable estimates of pronunciation variants.

The problem is then to generalize pronunciation probabilities among similar words.

In the following discussion, the three steps of canonical pronunciation generation are discussed further for phone-based pronunciation dictionaries, as well as the addition of variants and corpus-based validation. While in practice these three steps are not necessarily separated, they provide a language independent methodological guideline for pronunciation dictionary development.

5.5.1 Canonical Pronunciations

To generate initial canonical pronunciations, one of the following approaches is generally used:

- **Completely manual:** The developer (often an expert in linguistics or phonetics) types in the phone sequence for each lexical entry. This approach is only viable for relatively small vocabulary tasks and poses the problem of pronunciation consistency.
- **Manually supervised:** Given an existing pronunciation, dictionary, rules are used to infer pronunciations of new entries. This requires a reasonably sized starting dictionary and is mainly useful to provide pronunciations for inflected forms and compound words.
- **Grapheme-to-phoneme rules:** These are usually developed for speech synthesis and work well for many languages. Special care needs to be taken to ensure that the text normalization is consistent with the pronunciation rules.
- **Manually supervised grapheme-to-phoneme rules:** Manual supervision is particularly important for languages with ambiguous written symbol sequences. For any language, proper names—particularly those of foreign origin—may not be properly spelled or may require multiple pronunciations.

In practice, it is common to use a combination of the above approaches, in which an existing pronunciation dictionary is used to provide pronunciations for known words, and new entries are added in a semiautomatic manner with possible pronunciations provided by rule. The resulting entries can be simply scanned into or to be displayed by a text editor,

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

or can be presented to a human via a specially adapted tool. Such a tool was developed for American English (Lamel and Adda, 1996), in which grapheme-to-phoneme conversion is often ambiguous, and straightforward rule application can produce erroneous phone transcription even for function and common words. For example, a rule to generate a pronunciation for the word *used* will derive a correct pronunciation from the root *use* but an incorrect one from the word *us*. An example tool to propose pronunciation for American English is shown in Figure 5.15, which applies rules as illustrated in Table 5.6. The rules try to remove either prefixes (P) or suffixes (S) from the word, and specify ordered actions (strip, strip+add), which are applied to the words (letter), and context dependent actions to modify the resulting pronunciations.[6] For example, if the word *banned* is unknown, the letter sequence *ed* is removed and the *n* undoubled. If the word *ban* is located in one of the source dictionaries, the phone /d/ is added to the returned pronunciation. If multiple rules match, all possible pronunciations are returned along with their source.

This type of tool is useful for deriving inflected forms of words already in the pronunciation dictionary. But once a reasonable sized master dictionary is available, new words tend to be proper names and acronyms, which fall out of the scope of a morphologically based pronunciation tool. If multiple monolingual dictionaries or a multilingual dictionary

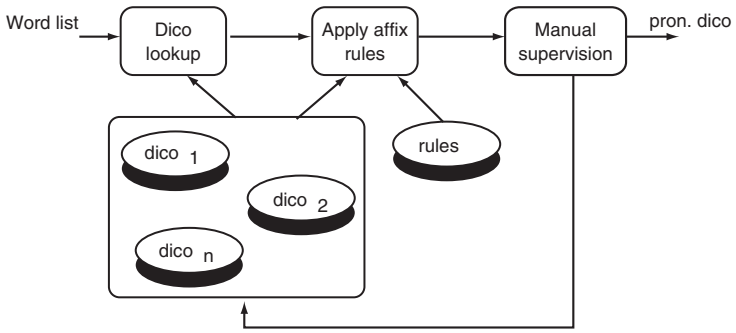


Figure 5.15: Pronunciation generation tool.

[6]The algorithm was inspired by a set of rules written by David Shipman when he was at MIT in the 1980s.

Table 5.6 Some example rules to strip and add affixes used by a pronunciation generation tool. Affix types are P (prefix) and S (suffix).

Affix type	Rule type	Remove affix	Add affix	Add phonemes	Context A/V/UV/C	Example word
P	strip	anti	-	/æn{t}[Iʊʲ]/	any	
P	strip	pre	-	/pri/	any	*preconceived*
S	strip+add	ier	y	/i/	any	*dirtier*
S	strip+add	iness	y	/nIs/	any	*happiness*
S	strip	ness	-	/nIs/	any	*carelessness*
S	strip	ally	-	/[l̩l]i/	any	*critically*
S	strip+	ed	-	/əd/	t,d	*admitted*
	undouble			/d/	V	*banned*
S	strip+add	ed	e	/əd/	/t,d/	*acted, ceded*
				/d/	V	*praised*
				/t/	UV	*faced*

is available, this tool can be used to propose spellings for proper names of different origins, with appropriate mapping of the graphemic form and phone spelling (Schultz and Waible, 1998a).

In languages with a close or at least regular correspondence between the written and spoken forms (such as French, German, Italian, Portuguese, and Spanish), an initial set of base-form pronunciation can be generated using grapheme-to-phoneme rules. The rules can be derived manually or can be developed in a data driven manner. (See Chapter 7 for references on letter-to-sound conversion.) Such letter-to-sound conversion systems typically make use of several hundred rules and a list of exceptions that may contain thousands of items. Spanish and Portuguese can be processed using about 100 rules, whereas German and French require about 300 and 500 rules, respectively. Some example rules and exceptions for French, German, and English are shown in Figure 5.16. The problem of homographic heterophones (words spelled alike but pronounced differently) is more or less important depending on the languages and will be addressed in the following section. For example, the Arabic language is particularly challenging since in Modern Standard Arabic, written texts are produced without vowel diacritic marks, there can be many vowelized forms (each which can have multiple possible pronunciations)

French	example	"l" Letter : /l/j/ / *mute*/sound ambiguity	
ctx letters ctx	sound	regular examples	Some exceptions
vi 11	l	ville, Albertville, bougainvilliers	pavillon, Chevilly
mi 11	l	mille, millier, millésime, millilitre	millet, Millon
Ci 11	j	fille, billet	pillule, bacille
^ V 11	l	elle, aller, illustre	
[aeu] il $	j	détail, pareil, seuil	
1	l	le, loi, alors, filet, total, vil,	gentil, fusil, outil
			mute letter

German	example	"s" letter: /z/s/ʃ/	sound ambiguity
ctx letters **ctx**	sound	regular examples	some exceptions
ss	s	Masse, Fluss, gefasst	Abgassteuer
			morpheme bound
sch	ʃ	schön	Volkscharakter
			morpheme bound
sh	ʃ	shoot, shirt	geisteshell
		English import	*morpheme bound*
s**[pt]**	ʃ	spassen, streiten	alterspassende
			morpheme bound
s**C**	s	skrupellos, slawish, snobistisch	
		import words	
s**V**	z	sieben, Faser	service *(imports)*
			preisaggressiv
			morpheme bound

English	example	"oo" letter: /ə/ʊ/ /ɔ/ʌ/	sound ambiguity
ctx letter **ctx**	sound	regular examples	some exceptions
oo **r**	ɔ	door, floor	coordinator,
			hooray, zoology
l oo **d**	ʌ	flood, blood	
oo	[əʊ]	room, football, balloon,	
		childhood	

Figure 5.16: Example of letter-to-sound rules standard French, German, and English, and related exception. Rule precedence corresponds to listed order; **ctx** specifies letter contexts; **C** is a constant; **V** is a vowel; and ^ and $ signify the word start and end.

associated with each lexical entry (Messaoudi et al., 2004). Each entry can be thought of as a word class containing all observed (or even all possible) vowelized forms of the word. However, if a vowelized written form is available, it is straightforward enough to derive reasonable canonical pronunciations.

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

In any case, even when grapheme-to-phoneme conversion is a viable solution for a given language, the huge class of words belonging to proper names (either domestic or imported) needs to be treated separately, as standard pronunciation rules often do not apply. Here, dictionaries from other languages can serve as initial knowledge sources.

Some available pronunciation dictionary resources

The Linguistic Data Consortium (see Chapter 3) lists pronunciation dictionaries for American English, German, Japanese, Mandarin Chinese, Spanish, Egyptian Colloquial Arabic, and Korean. The dictionaries range in size from 25,000 to over 300,000 words. Pronunciations for words in these dictionaries were either derived by letter-to-sound conversion with some manual supervision or derived from other resources. For example, the German dictionary is based on the Celex (http://www.ru.nl/celex/) lexical resources for which large dictionaries are available for the Dutch, English, and German languages. Also important to mention are the pronunciation dictionaries developed for the Eurom and SpeechDat family of corpora distributed by ELDA (see Chapter 3) and the Carnegie Mellon University Pronouncing Dictionary (CMUdict) (CMU, 1998) for North American English.

The difficulty in generating reasonable full-form pronunciations for proper names is a well-known problem: the potential list of proper names is unbounded, and pronunciation rules are certainly less reliable here than for general language words. The generation of proper name pronunciations for speech technology has been explicitly addressed by the Onomastica Project and at Bellcore Spiegel (1993) with the development of the Orator speech synthesizer for American English names. The Onomastica Project, funded by the European commission under the LRE program, developed pronunciation dictionaries of proper names and toponyms in 11 languages (Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish, and Swedish). The Orator system uses rules to generate name pronunciations with a classification of the ethnic origin of the name, complemented by an exceptions dictionary (Spiegel, 1993; Spiegel and Macchi, 1990). The availability of pronunciation dictionaries in many languages contributes to fostering multilingual speech technologies for numerous applications.

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

5.5.2 Pronunciation Variants

A given lexical entry may be assigned multiple pronunciations for different reasons: ambiguous writing conventions; phonological variants, which may be due to coarticulation, speaking style, dialect, or accent; loan words; and proper names. Adding appropriate variants seems particularly useful if significant differences can be observed in the acoustic realizations of a given word, and if these differences are unlikely to be represented properly by the acoustic models.

Variants need to be added for homographic heterophones, the proportion of which is to a large extent language dependent. Text normalizations may also contribute to this type of ambiguity. As mentioned earlier, Modern Standard Arabic written texts are produced without vowel diacritic marks, which means that each consonantal root form can be associated with a large number of vowelized forms (Messaoudi et al., 2004). Each consonantal root can be considered a word class, representing all possible vowelized forms of the root. As an example, the transliterated word *ktAb* (*book)* corresponds to four vowelized written forms: *kitaAb, kitaAba, kitaAbi*, and *kutaAbi*. The French language produces mainly morphosyntactic homographs (*président* /prezidã/ [noun] or /prezid/ [verb], *désertions* /dezɛʁsjõ/ [noun] or /dezɛʁtjõ/ [verb]) but also some homographs with unrelated lemmas (*as* /a/ [(*you*) *have*] or /as/ [*ace*], *est* /ɛ/ [(*he*) *is*] or /ɛst/ [*east*]). For Arabic the explicit writing of short vowels eliminates the need for variants, whereas in French, the explicit morphosyntactic information would be required.

Among very common phonological variants is the optional schwa vowels in many languages (French, German, Dutch). Other common phonological variants are due to assimilation. Some example alternate pronunciations for American English and French are given in Figure 5.17 using IPA symbols. For each word, the base-form transcription is used to generate a pronunciation graph to which word-internal phonological rules are optionally applied to account for some of the phonological variations observed in fluent speech. The pronunciation for *counting* allows the /t/ to be optional, as a result of a word-internal phonological rule. The second word, *interest*, may be produced with 2 or 3 syllables, depending on the speaker; in the latter case, the /t/ may be omitted and the [n] realized as a nasal flap. For the next word, *excuse*, the different pronunciations reflect different parts of speech. The suffix *-ization* can be pronounced

counting	kɑʷntɨŋ kɑʷnɨŋ	
interest	IntrIst IntəɹIst InəɹIst	
excuse	Ekskjuz Ekskjus	
amortization	əmɔrtəzeSxn æmɔrtəzeSxn əmɔrtɑʲzeSxn æmɔrtɑʲzeSxn	
company	kʌmpəni kʌmpni	
coupon	kjupɑn kupɑn	
republique	repyblik repyblikə	word-final optional schwa
les	le lez lɛ lɛz	liaison phoneme /z/, vowel harmony
prendre	prɑ̃drə prɑ̃dr prɑ̃d	word-final consonant cluster reduction
dix	dis dis{ə} di diz	variants on numbers
DM	dœtSmark deɛm	abbreviations
Morgan	mɔrgɑ̃ mɔrgɑn	multilingual proper name

Figure 5.17: Examples of alternate valid pronunciations for American English and French.

with a diphthong (/ɑʲ/) or a schwa (/ə/). Another well-known variant is the palatalization of the /k/ in a /u/ context, such as in the word *coupon* (/ku/ versus /kju/). In the spectrogram on the left of Figure 5.18, the word was pronounced /kjupan/, whereas on the right, the pronunciation was /kupan/. If the correct pronunciation is not predicted, during acoustic model training the speech frames will be aligned to the standard pronunciation unless the recognizer can handle pronunciation variants and applies a flexible alignment. If the recognizer can only handle single pronunciations, the variant /kju/ will be implicitly modeled by the model sequence /ku/ and that all pronunciations /kju/ can be decoded as /ku/, which is not always desirable (e.g., *Cooper* versus *cue*).

The French examples illustrate major variant phenomena: word-final optional schwa; vowel harmony; consonant cluster reduction; liaison consonants, which may be optionally produced before a vowel; ambiguous written forms (abbreviations and proper names).

For Spanish, rules can be used to generate multiple pronunciations of the grapheme *acci*—as in *accidental* and *acciones*—allowing for a realization as /akz/ or /az/, and for the grapheme *cion*, to be realized as an /s/ or a /z. Similarly, rules can be used to propose the deletion or insertion of schwas in certain contexts, vowel reduction to schwa, voicing assimilation, or stop deletion, just to mention a few common phenomena in many languages.

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

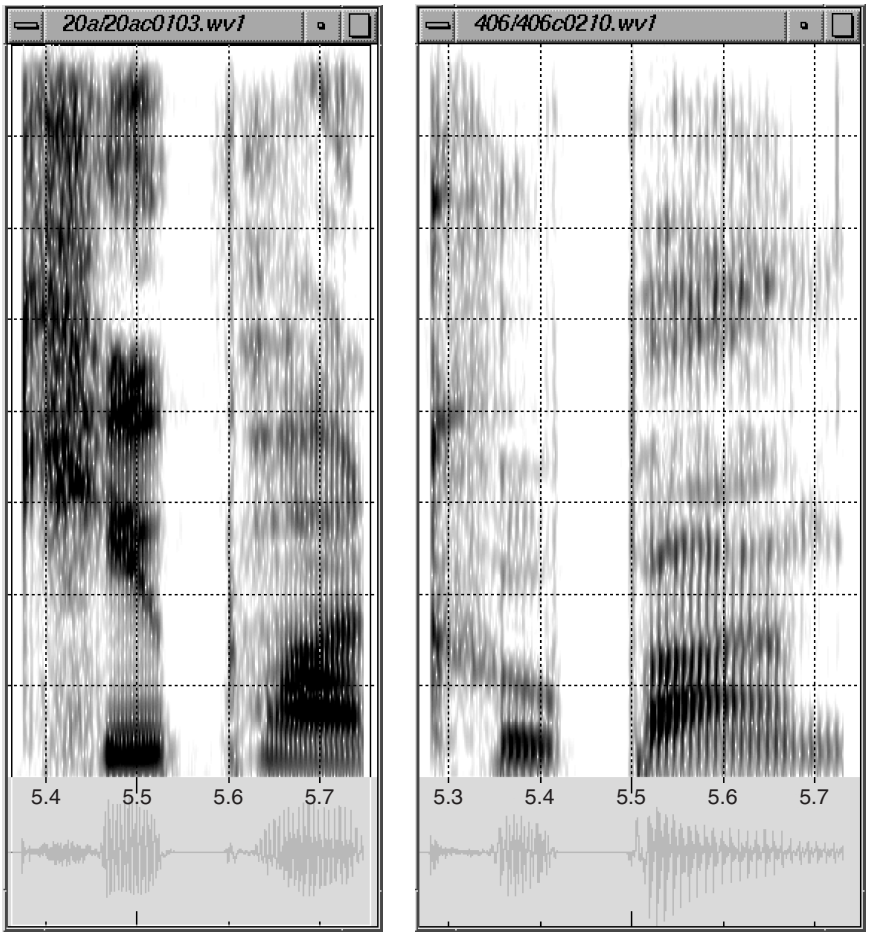


Figure 5.18: Two example spectrograms of the word *coupon:* (left) /kjupɑn/ and (right) /kupɑn/. The grid 100 ms by 1 kHz.

Including variants into the pronunciation dictionary becomes particularly important when severe temporal mismatches are likely to occur between the full-form pronunciation and the produced utterance. This is more frequent in casual speech than in formal speech. The part of the vocabulary that is shared by vernacular and formal speech (function words, common verbs, nouns, and idiomatic expressions) is more prone

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

to phonological variants than are technical items or proper names, for instance. This is a general observation that can be made for all languages for which we have developed LVCSR systems (English, French, German, Arabic, Spanish, Mandarin Chinese, and Portuguese). An explanation can be proposed on an information theoretic level: since these words are very frequent, their information content is low. They are very often highly predictable from their context. Similar observations hold for morphological units (subwords) corresponding to recurrent morphemic items, such as declension specifications, prefixes, and affixes. On an articulatory level, simplified articulations (pronunciations) can be favored as a result of the repetitive production of these words. On a perceptual level, one can also hypothesize an accelerated activation, which is due more to context than to objective acoustic observations.

Dates and numbers are subject to pronunciation simplifications since they are frequent and contain a fair amount of redundant information. This is particularly true when the contextual information is sufficient for understanding. For example, for the number 88 (*quatre-vingt-huit* in French), the /v/ is often essentially deleted. Similar observations can be made for numbers in German (99, nominally *neun'-'n-neunzig*, is frequently pronounced as display as /phone sequence/) and English (150, nominally *one hundred and fifty*, where the word *and* can be heavily reduced or even disappear).

ED: Please advise if the change made is OK.

Simplified pronunciations can also be observed across word boundaries for function-word sequences. For example, the German word sequence *haben wir* (*do we have*), with a full-form pronunciation /habən viɐ/, can be reduced in vernacular speech to approximate pronunciations such as /ham vɐ/ or even /hamɐ/. In French, the sequence *c'est quelque chose* (*it's something*), which has a canonical pronunciation (/sɛkɛlkəʃoz/), can be severely reduced, keeping only six phonemes /sekʃoz/.

As already mentioned, proper names are particularly difficult to handle, since their pronunciation can be quite variable, depending on the speaker's general knowledge, the origin of the name, and influence of other languages. For example, Worcester—a city in Massachusetts—should be pronounced /wustɝ/, but those not familiar with the name often mispronounce it as /wɔʧɛtɝ/. Similarly, the proper names *Houston* (the street in New York is pronounced /hɑ^{w}stən/ and the city in Texas is /hyustən/), *Pirrone*, and *SCSI* may be pronounced differently depending on the speaker's experience.

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

Experiments with automatically transcribing different styles of speech (public speech from broadcast news, conversational speech over the telephone) have highlighted the important differences in pronunciations between formal and casual speech, particularly concerning its temporal structure. Pronunciation modeling will contribute to a better knowledge of these spontaneous speech-specific phenomena.

5.5.3 Phone Sets and Acoustic Modeling

Typically a pronunciation dictionary will use a specific phone set. Different dictionaries for the same language may have slightly different numbers of units. For example, commonly used phone sets range from about 25 for Spanish to about 50 for English, French, Arabic, and German to about 80–100 for Mandarin when tone is explicitly modeled. For speech modeling, pronunciations are expressed using a phonemic alphabet, which is then shared by the acoustic models. This alphabet can allow for more or fewer distinctions with more or less detailed IPA (International Phonetic Alphabet) symbols. For automatic speech processing, it is important to consider at what level speech variation should be modeled. As pronunciation generation (base-forms and variants) involves human expertise, it is desirable to limit its relative importance in the overall system. This implies that it is preferable to model variation implicitly within the acoustic models rather than explicitly in the pronunciation dictionary. Different parameters must be taken into account when choosing a phone set:

- For the purpose of acoustic modeling, the granularity of the phone set, phone frequencies, and temporal modeling capacity must be considered.
- For pronunciation dictionary development, the granularity of the phone set, the consistency of the resulting pronunciations, and the level of human effort required must be taken into account.
- Finally, for multilingual applications, the portability of the phone set to different languages should be a criterion.

Frequency of occurrence of phones is an important criterion. In order to obtain reliable estimates of the properties of a sound, especially of a sound in different phonemic contexts, it is vital to have enough observations. This is the reason why xenophones (i.e., phones from different languages) are

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

generally not added to phone lists in a monolingual setup. Both during training and recognition, each phone needs to be aligned with an acoustic segment. In phone-based hidden Markov model (HMM) recognizers (see Chapter 4), it is common to use a 3-state left-to-right model with a minimum duration (typically 3 frames, corresponding to 30 ms), as illustrated in Figure 5.19. Beyond frequency of occurrence, this minimum duration constraint can also have an impact on the definition of the system's phone inventory. Complex phonemes such as affricates (ʧ, ʤ, ʦ) and diphtongs (ɑʲ, ɑʷ, ɔʲ, ʲu) can be represented by either one or two phone symbols, which implies modeling with one or two HMMs, as shown in Figure 5.20. On the one hand, a possible advantage of using a single unit is that the minimum duration is half that required for a sequence of two phones, and may be more appropriate for fast speaking rates or casual speech. On the other hand, a representation using two phones may provide more robust training if the two component phones also occur individually and diphthongs only occur infrequently in the training data.

Most contextual variation is implicitly taken into account by training multiple models for a given phone, depending on its left and right phone contexts (Schwartz et al., 1984; Chow et al., 1986; Lee, 1988). The selection of the contexts to model usually entails a trade-off between resolution

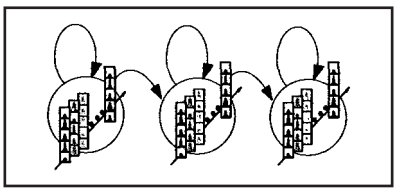


Figure 5.19: An acoustic phone like segment is temporally modeled as a sequence of 3 states, each state being acoustically modeled by a weighted sum of Gaussian densities (see Chapter 4 for more details on acoustic modeling).

Figure 5.20: Impact on acoustic/temporal modeling depending on the choice of one or two symbols for affricates or diphthongs.

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

and robustness, and is highly dependent on the available training data. Different approaches have been investigated, from modeling all possible context-dependent units, to using decision trees to select the contexts (see Chapter 4), to basing the selection on the observed frequency of occurrence in the training data. In all cases, smoothing or back-off techniques are used to model infrequent or unobserved contextual units. Numerous ways of tying HMM parameter have been investigated (Young et al., 1994; Gauvain and Lamel, 1996).

If the phone symbol set makes fine distinctions (such as between different stop allophones—unreleased, released, aspirated, unaspirated, sonorant-like) many variants will be needed to account for the different pronunciation variations. This raises the problem of completeness and consistency of the pronunciation dictionary, and increases the amount of human effort in pronunciation-dictionary generation. If the basic phones set remains close to a phonemic representation, pronunciation variants are necessary only for major deviations from the canonical form or words for which there are frequent alternative pronunciation variants that are not allophonic differences.

When porting a recognizer from one language to another, standard practice is to use acoustic models from already modeled languages as initial seed models. In doing so, a mapping must be made between the phones in the target language and those in the other language(s). Generally there is a preference to use more generic context-independent models to reduce the influence of the original language. Sometimes it is interesting to use a particular context-dependent model in order to better approximate a phone in the target language that does not exist in any of the other languages. There also exist language-independent and cross–language acoustic modeling techniques to port recognition systems from one language to another without language-specific acoustic data. However, these data remain valuable for acoustic model adaptation. See Chapter 4 for more details.

5.5.4 Corpus–based Validation

For many years, the use of pronunciation variants was considered risky, since too many variants could potentially increase the number of homophones; therefore, they were only sparsely introduced into pronunciation dictionaries. The availability of very large transcribed speech corpora

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

enables exploration of a new approach to introduce variants by using rules to overgenerate pronunciations in a preliminary working dictionary and validating their selection on a large amount of data (see Figure 5.14) (Adda-Decker et al., 2005). The corpus-validation step aims at a coupled tuning of acoustic and pronunciation models in order to minimize speech recognition errors.

Even if multiple pronunciations can be hypothesized for a given lexica1 entry, they are not equally useful. Whereas multiple pronunciation lexicons are often (at least partially) created manually, several approaches have been investigated to automatically learn and generate word pronunciations and to associate probabilities with the alternative pronunciations (Cohen, 1989; Riley and Ljojley, 1996; Cremelie and Martens, 1998). The estimation of pronunciation probabilities commonly relies on pronunciation variants in large corpora. As an example, Table 5.7 gives the pronunciation counts for different variants of four inflected forms of the word *interest* in 150 hours of broadcast news data and 300 hours of conversational telephone speech. It can be seen that there is a larger proportion of reduced pronunciations (fewer phones, nasal flap) in conversational telephone speech (CTS) than in broadcast news (BN). For a given style of speech, longer entries (*interesting*) tend to have more reduced variants than shorter entries (*interest*). As word occurrences follow Zipf's law, pronunciation probabilities can be reasonably estimated for several thousand lexical entries if several hundred hours of transcribed speech data are available.

Table 5.7 Pronunciation counts for inflected forms of the word *interest* in 150 hours of broadcast news (BN) data and 300 hours of conversational telephone speech (CTS).

Word	*Pronunciation*	*BN (150h)*	*CTS (300h)*
interest	ɪntrɪst	238	488
	ɪntəˠɪst	3	33
	ɪnəˠɪst	0	11
interested	ɪntrɪstəd	126	386
	ɪntəˠɪstəd	3	80
	ɪnəˠɪstəd	18	146
interesting	ɪntrɪstɨŋ	193	1399
	ɪntəˠɪstɨŋ	8	314
	ɪnəˠɪstɨŋ	21	463

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

The problem is then to estimate probabilities for pronunciation variants of words that are not sufficiently observed in the training corpus, which is the case for most pronunciations in any large dictionary. In practice, for frequent words (mainly function words but also some content words and idiomatic items), unobserved variants are removed or given a minimal count. By default, all pronunciation variants of infrequent words are considered as equiprobable. One major outstanding challenge is a realistic generalization of the observed pronunciation probabilities via phonological rules and variant types (used to generate the variants). Information about stressed and unstressed syllables in polysyllabic words is certainly a factor to take into account for probability generalization.

While modeling of pronunciation variants and estimation of pronunciation probabilities has attracted much research attention over the years, recent work reported in Hain (2005) proposes using only a single pronunciation for LVCSR. Recognition tests demonstrate that there is no loss in performance, and in certain cases, performance can be improved. This is an interesting case of a corpus-based selection process pushed to its limit. First, a large set of possible pronunciation variants is generated using varying degrees of human supervision. Then, the single most representative pronunciation is selected from all variants for a given word using appropriate acoustic training data. Experience shows that the most representative variant is likely to change with speaking style of the training data, at least for the most frequent items. This implies that for a given language, different pronunciation dictionaries are used depending on the speaking situation. Future multilingual pronunciation dictionaries will certainly need such a validation step on multilingual acoustic data.

5.6 Discussion

For most automatic speech recognition systems, multilingual pronunciation dictionaries are still collections of monolingual dictionaries. However, as was observed for the different languages examined in this chapter, the proportion of imported words—that is, words shared with other languages—increases with vocabulary size. For example, for word lists containing the most frequent 50,000 words in broadcast news texts, 10–20% of the lexical entries are shared between language pairs. Proportions

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

are particularly high for languages of smaller linguistic communities, which tend to more easily incorporate words from other languages. Almost 30% of the lexical entries are shared between French and Luxembourgish. French is often used in addition to Luxembourgish in official situations. This means that monolingual word lists naturally evolve toward multilingual ones with increasing vocabulary size. If common word lists are developed for processing different languages, pronunciations and acoustic models need to be adapted appropriately. The size of the vocabularies used in state-of-the-art speech recognition systems has been growing, and it is likely that system word lists will contain over 500,000 words in the near future.

This chapter has addressed the various steps in lexical development, including the normalization, choice of word items, the selection of a word list, and pronunciation generation. Tokenization and normalization were first addressed in the context of written sources, which often form the basis of language modeling material and are required to ensure viable lexical coverage and word *N*-gram estimates, particularly concerning the treatment of punctuation, numbers, abbreviations, acronyms, and capitalization. Lexical coverage measures on training data give a good indication of whether the normalizations are properly addressed, For compounding and agglutinative languages (e.g., German, Dutch, Turkish, and Finnish), decomposition techniques are required to optimize lexical coverage. A simple corpus-based and mostly language-independent decomposition algorithm was presented, based on Zellig Harris's algorithm for finding morphemes from phonemic strings (developed half a century ago).

To select efficient word lists for a given speech recognition application, it is important to tune the system's word list using appropriate development data with respect to epoch (in those cases where the data is time-sensitive, such as broadcast news text) and topics (see Chapter 6 on language modeling for more details). For automatic speech recognition, research has shown the importance of using speech transcripts in addition to written sources. Speech-specific items, such as disfluencies, discourse markers, respirations, and fragments do not exist or are only weakly present in written sources. Therefore their observation probabilities can only be reasonably estimated from speech transcripts. Here, the problem of weighting different types of sources for vocabulary list and language model definition is an important issue. Moreover, to achieve reasonable accuracy in acoustic modeling of reduced word sequences, *multiwords* (function word sequences,

Typeset by: CEPHA Imaging Pvt. Ltd., INDIA

idiomatic expressions, frequent acronyms) may be necessary for casual speech.

Concerning word pronunciations, general guidelines can be given for choosing a phone symbol set that produces a reasonable compromise between pronunciation and estimation accuracy: in a monolingual setup, rare symbols and xenophones need to be eliminated. A multilingual setup may at least partially solve this problem. If fine distinctions are allowed by the phone symbol set, more human effort may be required to ensure dictionary completeness and consistency. Even if acoustic phone models can potentially model a lot of implicit acoustic variation, pronunciation variants should be added for ambiguous written forms (homograph heterophones) and for the most important phonological variants. For all languages, adding variants is very important for the N most frequent words (with N lower than 1,000). Frequent function words often have pronunciation variants with different temporal structures, which do not necessarily generalize to less frequent items. As a result, implicit modeling within triphones can be harmful for the global acoustic modeling accuracy. In our view, explicitly modeling pronunciations contributes to a finer tuning of multilingual models, which can in turn be useful in the development of educational and medical services (L2 acquisition, orthophony) and for linguistic research (phonetics, phonology, dialectology, sociolinguistics). With the recent availability of very large spoken corpora, corpus-based explorations may develop into an important research direction.

Future work will include automatic processing based on huge multilingual word lists of millions of entries (for languages from a given family that share similar writing conventions). Depending on the type of text normalizations (e.g., removing accents and diacritics), a significant part of the vocabulary can then be shared among languages such that language-specific lexical entries can typically be limited to some tens of thousands of items. These aspects, though not yet addressed—at least in the context of open-domain transcription—offer new perspectives to multilingual automatic processing.