Machine-Readable Dictionaries in Text-to-Speech Systems

Judith L. Klavans†and Evelyne Tzoukermann *
†Columbia University, Department of Computer Science, New York, New York 10027
klavans@cs.columbia.edu
†* A.T.&T. Bell Laboratories, 600 Mountain Avenue, Murray Hill, N.J. 07974
evelyne@research.att.com

Abstract

This paper presents the results of an experiment using machine-readable dictionaries (MRDs) and corpora for building concatenative units for text to speech (TTS) systems. Theoretical questions concerning the nature of phonemic data in dictionaries are raised; phonemic dictionary data is viewed as a representative corpus over which to extract n- gram phonemic frequencies in the language. Dictionary data are compared to corpus data, and phoneme inventories are evaluated for coverage. A methodology is defined to compute phonemic n-grams for incorporation into a TTS system.

1 Introduction

The majority of speech synthesis systems use two techniques: concatenation and formant-synthesis. Building a comprehensive and intelligible concatenative-based speech synthesis system relies heavily on the successful choice of concatenative units. Our results contribute to the task of developing an efficient and effective methodology for reducing the potentially large set of concatentive units to a manageable size, and to chosing the optimal set for recording and storage.

The paper is aimed primarily at two audiences: one consists of those concerned with research on the automatic use of MRD data; the other are TTS system designers who require linguistic and lexicographic resources to improve and streamline system-building. Issues of morphological analysis and generation, as well as stress assignment based on dictionary data, are discussed.

2 Using MRDs in Text to Speech

Several problems are addressed in this paper; one concerns the subtle complexities and idiosyncrasies involved in parsing dictionaries and extracting data. Added to this is the lack of consistency both within the same dictionary and across dictionaries which often requires ad hoc procedures for each resource. Another issue relates to the structure of the modules of a TTS system, specifically in the grapheme-to-phoneme component; dictionary lookup depends on several factors including size, machine power and storage, factors that have important consequences for the extraction of concatenative units. Another consideration concerns the nature of the language itself: a language with irregular grapheme-to-phoneme mapping and lexically determined stress assignment (such as English) benefits most from the large exception list which a dictionary can provide. There is also the practical issue of dictionary availability, and of pronunciation field accuracy within an available dictionary. Thus, decisions on the use of MRD data depend on many factors, and can significantly impact efficiency and accuracy of a speech system.

Since a dictionary entry consists of several fields of information, naturally, each will be useful for different applications [1]. Among the standard fields are pronunciation, etymology, subject field notes, definition fields, synonym and antonym cross references, semantic and syntactic comments, run-on forms, conjugational class and inflectional information where relevant, and translation for the bilingual dictionaries. Each of these fields has proven useful for different applications, such as for building semantic taxonomies [3], [13] and machine translation [12]. The most directly useful for TTS is the pronunciation field [4], [11]. Equally useful for TTS, but less directly acces-

sible, are data from run-on fields, conjugational class information, and part-of-speech. [1]

To illustrate, the following partial entries from Webster's Seventh (W7) [15] illustrate typical pronunciation, definition, and run-on fields:

(1) **ha.ven** /ˈh.ā-vən/ *n* **1**: HARBOR, PORT **2** : a place of safety : ASYLUM – **haven** *vt*

(2) **bi.son** /ˈbīs-ən, ˈbīz-/ *n* ...

(3) **ho.mo.ge.neous** /-ˈjē-nē-əs, -nyəs/ ...

(4) **den.tic.u.late** /den-ˈtik-yə-lət/ *or* **den.tic.u.lat.ed** /-ˌlāt-əd/ *adj*

The entry for "haven" contains one full pronunciation. The entry for "bison" has one alternative, but the user must figure out that the /ən/ should be appended after ˈbīz-/, as in the first pronunciation, in order to obtain the correct variation. Correct pronunciation for "homogeneous" relies on the pronunciation of the previous entry,"homogeneity", and requires the user to separate and bring the prefix "homo-" from one entry to another. To complicate matters, the alternative pronunciation for the suffix /nē-əs/-nyəs/ must also be correctly interpreted by the user. Finally, "denticulate" has a morphologically related run-on form "denticulated" in the early part of the entry, and the pronunciation of that run-on is related to the main entry, but the user must decide how to strip and append the given syllables. [2] While these types of reasoning are not difficult for humans, for whom the dictionary was written, they are quite difficult for programs, and thus are not straighforward to perform automatically.

2.1 Using the MRD pronunciation field

Extracting the pronunciation field from an MRD is one of the most obvious uses of a dictionary. Nevertheless, parsing dictionaries in general can be a very complex operation ([16]) and even the extraction of one field, such as pronunciation, can pose problems. Similar to W7, in the Robert French dictionary [9], which contains about 89,000 entries, several pronunciations can be given for a headword and the choice of one must be made. Moreover, because of the rich morphology of French which has a rough ratio of eight morphologically inflected words for one baseform, Robert lists only the non-inflected forms of the lexical entries. However, if pronunciation varies during inflection of nouns and adjectives, the pronunciation field reflects that variation which makes the information difficult to extract automatically. For example, in (5) and (6), one needs to know the nature of the rule to apply in order to relate both forms of the adjective.

(5) **blanc, blanche** /blã, blãʃ/ adj. et n.

(6) **vif, vive** /vif, viv/ adj. et n.

In (5), the masculine /blã/ is obtained by removing the phoneme /ʃ/ from the feminine /blãʃ/ (blanche,"white"). In (6), the form masculine form /vif/ ("sharp, quick") is formed by stripping the affix /ve/ and substituting the phoneme /f/. Notice that the rules are different in nature, the first being a addition/deletion relation, and the second being a substitution.

In this project, the dictionary pronunciation field was used to start building the phonetic inventory of a speech synthesis system. For the French TTS system [?], the set of diphones was established by taking most of the thirty-five phonemes for French and coupling them with each other (35^2 = 1225 pairs). Then, the diphones were extracted from the pronunciation field for headwords in the Robert dictionary. A program was written to search through the dictionary phonetic field and select the longest word where the phoneme pairs would be in mid-syllable position. For example, the phonemic pair /lo/ was found in the pronunciation field /zoolɔʒik/ corresponding to the headword zoologique "zoologic."

Out of 1225 phonemic pairs, 874 words were found with at least one occurence of the pair. The pair [headword_orth, headword_phon] was extracted and headword_orth was placed in a carrier sentence for recording. For instance, the speaker would utter the following sentence: "C'est zoologique que je dis" where "C'est ... que je dis" is the carrier sentence. Due to the lack of explicit inflectional information for nouns and adjectives, only the non-inflected forms of the entries were extracted during dictionary lookup for building the diphone table. Similarly for verbs, only the infinitive forms were used since the dictionary does not list the inflected forms as headwords. This exemplifies the most simple way to use pronunciation field data, which we have completed. A pronunciation list of around 85,796 phonetic words was obtained from the original list of almost 89,000 entries, i.e. 96% of the entries. The remaining 4% consist primarily of prefixes and suffixes which are listed in the dictionary without pronunciations,

[1] Notice, however, that the full Collins Spanish-English dictionary [7], as opposed to the other bilinguals, does not contain any pronunciation information. Although this is rather surprising taking into account that the smaller versions such as the paperback and gem ([8], [10]) do have a phonetic field, it could be attributed to the fact that pronunciation rules in Spanish are relatively predictable.

[2] [2] reports on the need to resyllabify entries already syllabified in LDOCE [18], since syllable boundaries for written forms usually reflect hyphenation conventions, rather than phonologically motivated syllabification conventions necessary for pronunciation.

and which should not be used in isolation in any case.

2.2 Using the MRD for morphology

Even though an MRD may not list complete inflectional paradigms, it contains useful inflectional information. For example in the Collins Spanish-English dictionary, verb entries are listed with an index pointing to the conjugation class and table, listed at the end of the dictionary. Using this information, a finite-state transducer for morphological analysis and generation was built for Spanish [20]. From the original list of over 50,000 words, a few million words have been generated. These forms can then be used as the input to the grapheme-to-phoneme conversion module in a Spanish TTS system.

2.3 Using Run-on's

A run-on is defined as a morphological variant of a headword, included in the entry. Run-on's are problematic data in MRDs [16], and they can be found nearly anywhere in the entry. In example (4), the run-on occurs at the beginning of the entry, and consists of a full form with suffix. More commonly, run-on's occur towards the end of the entry, and tend to consist of predictable suffixation, that is, class II or neutral suffixes [19] , such as -ness, -ly, or -er, as in:

(7) **sharp** *adj*.... **sharp.ly** *adv*–**sharp.ness** *n*

(8) **suc.ces.sion** *n*.... **suc.ces.sion.al** *adj*
suc.ces.sion.al.ly *adv*

In cases where stress is changed with class I non-neutral suffixes, a separate pronounciation is given as in:

(9) **gy.ro.scope** /ˈjī-rə-ˌskōp/ *n* ....
gy.ro.scop.ic /jī-rə-ˈskäp-ik/ *adj*-
gy.ro.scop.i.cal.ly /-i-k(ə-)lē/ *adv*

The run-on form with part-of-speech is given inside the entry, so it could be used for morphological analysis. However, since pronunciation is usually predictable from the headword (i.e. there is usually no stress change, and if there is a change, this is explicitly indicated) the run-on pronunciation often consists of a truncated form, requiring some logic for reconstruction of the entire pronunciation. Again, this may be obvious to the human user, but rather complex to figure out by program. Thus, the run-on may be useful for morphology, but is not as useful for automatic pronunciation extraction.

3 Methodology and Results

3.1 Collecting Data

As stated above, out of almost 89,000 headwords in the dictionary, 874 phonemic pairs,which represents 71% of the total, were found. This is due to the fact that (a) the lookup occurs only on non-inflected words, thus a limited sample of the language, (b) because the dictionary consists of a list of isolated words, it does not account for inter-word boundary phenomena. Since French liaison plays such an important role in the phonology of French, a look at phonetic data from a corpus must be given in order to achieve full coverage. A portion of the Hansard French corpus (over 2.3 million words) was used for this purpose. Grapheme-to-phoneme software [14] was utilized in order to convert French orthography into phonemes. For the sake of comparison, both the phonetic transcription from the corpus and the one from the MRD were converted into a unique set of phonemes. Typical output from the dictionary looks like:

ABACA [abaka] n. m.
ABASOURDIR [abazuRdiR]; [abasuRdiR]
ABASOURDISSANT, ANTE [abazuRdisA, At
ABATTEUR, EUSE [abat8R, 7z] n.
ABCE'S [absE; apsE] n. m.
ABDOMINAL, ALE, AUX [abd>minal, o] adj.
ABDOMINO-
ABDUCTION [abdyksjO] n. f.

A small sample of the Hansard followed by the ascii transcription is shown below:

Pre'sident de la Compagnie d'Ame'nagement du barreau de.
X Monsieur X De'pute' ancien Ministre Pre'sident du Conseil.

prezidA d& la kOpaNi d amenaZmA dy baro d& iks m&sju dis depyte AsjI ministr prezid dy kOsEj

As an experiment, we compared triphones extracted from dictionary data and corpora. A greedy algorithm [3] to locate the most common coocurrences between orthography and transcription was run on the data sets. A sample of the corpus and dictionary results are given in the Table below. The table shows in the leftmost two columns the top twenty triphones and occurring frequencies extracted from the Hansard corpus, whereas the righthand columns show dictionary results. Notice the discrepancy between these lists; for the top twenty triphones, there are only

[3] We thank Jan van Santen for providing this software.

two overlaps, sjO and jO*. The levels of commonality between the triphones of the Hansard and the dictionary (5% of commonality for the top 100 triphones and 15% of commonality for the top 1000 triphones) is interesting to observe.

Hansard data				Robert data			
54580	sjO	38745	*Z&	3636	mA*	1725	ism
53948	jO*	38707	sEt	3324	ik*	1554	ite
47339	par	35052	k>m	3223	jO*	1492	je*
44328	asj	30389	*mE	2823	sjO	1462	sm*
44065	pr¿	39722	te*	2597	te*	1405	bl*
43288	tr&	29093	&pr	2202	*de	1391	ali
41356	&la	28784	ist	2105	5sj	1389	abl
40877	pur	26766	*s&	2086	ER*	1376	tik
39122	&mA	25997	Est	2067	aZ*	1341	*ka
38707	d&l	25378	mA*	1789	ist	1321	st*

Table 1: Twenty most frequent triphones

The preliminary results indicate that the coarticulatory effects derived from the corpus data will be useful, in particular for languages like French where liaison plays a major role. This remains to be tested in the TTS system.

3.2 Related Work

Although the statistical analysis of MRDs has focussed primarily on definitions and translations, [5] used the pronunciation field as data. A dictionary of over 110,000 entries containing 51,219 common words and 59,625 proper nouns, [17] was used for selecting candidate units that were further utilized in the set of concatenative units (diphones, triphones, and longer units) for synthesis. The phonemic string was split according to ten language-dependent segmentation principles. For example, the word "abacus" [ˈab-ə-kəs] was first transformed into cuttable units as follows: [#ˈa,ˈab,bə,ək,kə,əs,s#]. Once each dictionary word was split, the duplicates were removed and the remaining units formed the set of concatenative units. At the end of this operation, a rather long list was obtained that was pruned by methods such as reduction of secondary and primary stress into one stress in order to keep only one +stress/-stress distinction. Techniques were shown that allow the selection of a minimal set of word pairs for inter-word junctures; every candidate unit inside and across word sequence was included. The same strategy was replicated on the Collins Spanish-English dictionary by [6]. In this fashion, the dictionary was used as a sample of the language in the sense that it assumes that most of the phonemic combinations of the language were present.

4 Limitations of MRDs

The most straightforward way, but in the long run not the most flexible, is to parse the phonetic information out of the pronunciation field. The pronunciation field information can generally be consulted by a TTS system within the grapheme-to-phoneme module. Additional rules for processes such as inter-word assimilation, juncture, and prosodic contouring need to be added, since isolated word pronunciation could already be handled by look-up table. Although appealing, there are two major drawbacks to this approach:

(a) dictionary pronunciation fields are often not phonetically fine-grained enough for acceptable speech output. For example, the pronunciation for "inquest" is given in W7 as /ˈin-ˌkwest/, but of course the nasal will assimilate in place to the velar, giving /iŋ-kwest/. Without assimilation, the perceptual effect is of two words: "in quest" and would be misleading. Again, the human user will assimilate naturally, but a text to speech system must figure out such details, since articulatory ease is not a factor in most synthesis systems. One way to solve this problem is to impose such assimilation on input from the pronunciation field by a set of post-processing rules. Although this solution would be correct in the majority of cases, blanket application of such rules is not always appropriate for lexical exceptions. For example, assimilation is optional for words like "uncaring", in this case related to the morphological structure of the lexical item. A TTS system will probably already have such rules since they are inherent in the grapheme-to-phoneme approach. Thus, it could be argued that there is no need for the dictionary pronunciation, since with a complete and comprehensive grapheme-to-phoneme conversion system, a list which requires post-processing is simply inadequate and unnecessary. This is the approach taken, for example, by [14], who makes use of small word lists (the main dictionary being 25K stored forms) and several affix tables to recognize graphemic forms, which are then transformed into phonemic representations;

(b) only a small percentage of possible words are listed with pronunciations in a dictionary. For example, Webster's Seventh contains about 70,000 headwords, but is missing words like "computerize" and "computerization" since they came into frequent use in the language after the 1963 publication date. Two solutions to this problem present themselves. One is to expand the word list from the dictionary to include run-on's, as illustrated in examples (3) and (4), and discussed in Section 2.3. The other is to build a morphological generator,

using headwords, part of speech, and other information as input, discussed in Section 2.2 that would be invoked when the word does not figure in the headword list.

5 Final Remarks

Although limitations clearly constrain the use of MRDs in TTS, we have demonstrated in this paper that it is more cost efficient to post process underspecified dictionary information such as inflection, pronunciation, and part-of-speech, rather than generate rules from scratch to arrive at the same end point. For speech synthesis, the data is not always perfect, and often must be post-processed. This paper has demonstrated ways we have successfully used dictionary data in TTS systems, ways we have post-processed data to make it more useful, and ways data cannot be easily post-processed or used.

Of course, for any TTS system, the power of the dictionary data can be found at the lexical, phrasal, and idiom level. Although any word list such as a dictionary is by definition closed, whereas language is open-ended, dictionary data has proven to be useful from both a theoretical and practical point of view.

References

[1] Branimir Boguraev, Roy Byrd, Judith Klavans, and Mary Neff. From machine readable dictionaries to a lexical knowledge base. Detroit, Michigan, 1989. First International Lexical Acquisition Workshop.

[2] David Carter. Ldoce and speech recognition. In Branimir Boguraev and Ted Briscoe, editors, *Computational Lexicography for Natural Language Processing*, chapter 6, pages 135–152. Longman, Burnt Hill, Harlow, Essex, 1989.

[3] Martin S. Chodorow, Roy J. Byrd, and George E. Heidorn. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, pages 299–304. Association for Computational Linguistics, 1985.

[4] Paul Cohen. Spelling to sound conversion for text to speech. 1982.

[5] John Coleman. Computation of candidate synthesis units. In *11222-930719-07TM*, Murray Hill, N.J., USA, 1993. Technical Memorandum, AT& Bell Laboratories.

[6] John Coleman and Pilar Prieto. Accurate pronunciation rules for american spanish text-to-speech. In *11222-930719-06TM*, Murray Hill, N.J., USA, 1993. Technical Memorandum, AT& Bell Laboratories.

[7] *Collins Spanish Dictionary: Spanish-English.* Collins Publishers, Glasgow, 1989.

[8] P-H. Cousin, L. Sinclair, J-F. Allain, and C. E. Love. *The Collins Paperback French Dictionary: French-English. English-French.* Collins Publishers, London, 1989.

[9] Alain Duval et al. *Robert Encyclopedic Dictionary (CD-ROM).* Hachette, Paris, 1992.

[10] M. Gonzales. *Collins Gem Spanish Dictionary: French-English. English-French.* Harper Collins Publishers, London, 1990.

[11] Judith Klavans and Sara Basson. Documentation of letter to sound components of the WALRUS text to speech system. 1984.

[12] Judith Klavans and Evelyne Tzoukermann. The bicord system: Combining lexical information from bilingual corpora and machine readable dictionaries. In *Proceedings of the 13th International Conference on Computational Linguistics*, Helsinki, Finland, 1990.

[13] Judith L. Klavans, Martin S. Chodorow, and Nina Wacholder. From dictionary to knowledge base via taxonomy. Centre for the New Oxford English Dictionary and Text Research: Electronic Text Research, University of Waterloo, Canada, 1990. Proceedings of the Sixth Conference of the University of Waterloo.

[14] F. Marty. Trois systèmes informatiques de transcription phonétique et graphémique. *Le Francais Moderne*, LX, 2:179–197, 1992.

[15] Merriam. *Webster's Seventh New Collegiate Dictionary.* G.& C. Merriam, Springfield, Mass., 1963.

[16] M. Neff and B. Boguraev. Dictionaries, dictionary grammars and dictionary entry parsing. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, 1989. Association for Computational Linguistics.

[17] Olive Joe P. and Mark Y. Liberman. A set of concatenative units for speech synthesis. In In J. J. Wolf and D. H. Klatt, editors, *Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America*, pages 515–518, New York: American Institute of Physics, 1979.

[18] Paul Procter, editor. *Longman Dictionary of Contemporary English.* Longman Group, Burnt Hill, Harlow, Essex: Longman, 1978.

[19] Elisabeth O. Selkirk. *The Syntax of Words.* MIT Press, Cambridge, Mass., 1982.

[20] Evelyne Tzoukermann and Mark Y. Liberman. A finite-state morphological processor for spanish. In *Proceedings of Coling90*, Helsinki, Finland, 1990. International Conference on Computational Linguistics.