Phonetic transcription of dotted Hebrew

Alon Itai

Alon Itai is Professor Emeritus of the Computer Science Department of the Technion, Haifa. His research interests include Machine Learning and Natural Language Processing. He is the director of the MILA center for processing Hebrew (http://mila.cs.technion.ac.il/eng/index.html), dedicated to developing tools for processing Hebrew.
itai@cs.technion.ac.il

Abstract
The pronunciation of Israeli Hebrew mostly follows the pronunciation rules of dotted Hebrew script, though there are several systematic deviations. As part of the development process of a new Hebrew lexicographic resource by K Dictionaries, we have constructed and implemented an algorithm to deduce the pronunciation from dotted texts and tested it on a large manually tagged database. The database contains 35,443 dotted Hebrew words and their IPA (International Phonetic Alphabet) transcription. The program succeeded in correctly predicting the pronunciation of over 89% of the words in the database. 78.5% of the errors occurred when predicting the stress of loan words. Most of the remaining errors occurred in predicting the pronunciation of *schwa*. We found out that the traditional phonological explanation that is based on sonority theory correctly predicts 88.3% of all pronunciations of *schwa*. We constructed an alternative algorithm that correctly predicts the pronunciation of *schwa* in 99% of the words of the database.

1. Historical background

Hebrew is among the first languages transcribed by a phonetic alphabet. The original script constitutes an abjad, i.e., a script where the vowels are not represented. However, some consonants served as matres lectionis, namely the consonants ה,ו,י in addition to their role as consonants are used to indicate the vowels, a,u/o and i . During the first centuries A.D. Aramaic replaced Hebrew as the main spoken language of Jewish people in Palestine. However, the Holy Scriptures and mainly the Bible were canonized using the original abjad. Because of the need to correctly read the Bible, a system of diacritics, called dots, was added during the 10th century C.E. These symbols were small, so as to not change the holy texts, and reflected the way the scriptures were pronounced at the time.

All through the Middle Ages, scholars continued to write in unvocalized Hebrew using the matres lectionis more extensively, and this is the standard script of Israeli Hebrew.

With the revival of Hebrew as a spoken language, at the end of the 19th century, the Sephardic pronunciation was adopted. This pronunciation fused several diacritics and Israeli Hebrew further modified the pronunciation. Thus dotted texts are only a rough guide to pronouncing Hebrew, and various systematic deviations exist.

2. Pronunciation Rules

2.1 Consonants. The consonants follow a regular pattern. Table 1 shows a many-to-one map to IPA.

2.2 Vowels. Most vowels follow a many-to-one transcription. The two problematic cases are the diacritic *qamatz*, which is most often pronounced /a/ but sometimes /o/, and the diacritic *schwa*, which is in most cases silent but sometimes pronounced /e/.

2.3 Stress. In most Hebrew words the stress is on the last syllable, though some are penultimate. Even though stress is phonemic, it is not transcribed explicitly in Hebrew dotted script. In nearly all cases one can deduce the stress from the vowel pattern of the word or from its morphological analysis.

Some noun patterns also have penultimate stress. For example, the segolite word pattern class can be easily recognized since the last vowel is the diacritic *segol* (e.g. כֶּלֶב /ˈkelev/ *dog*). There are several related patterns which are easy to identify. Verbs in the past tense end with an unstressed suffix (e.g. כָּתַבְתִּי /kaˈtavti/ *I wrote*). These inflections can be identified by a morphological analyzer.

Loan words pose a greater challenge as their stress does not follow the rules of native words. The stress is most often penultimate and, contrary to native Hebrew words, its position does not change even in the presence of a stressed suffix. Thus, a prerequisite

Heb	א,ע	ב	ב,ו	ג	ד	ה	ז	ח,כ	ט,ת	י	כ,ק	ל	מ	נ	ס,ש	פ	פ	צ	ר	ש
IPA	ʔ	b	v	g	d	h	z	x	t	j	k	l	m	n	s	p	f	t͡s	ʁ	ʃ

Table 1: Transcription of Hebrew consonants

for determining the stress position of a word is to determine whether it is a loan word, and in some cases a morphological or semantic analysis is necessary.

For example, the word *bira* with ultimate stress is a native word meaning *capital* (city) and with penultimate stress is a loan word meaning *beer*. The dotted script renders both words identically. Thus, to correctly determine the stress position one first needs to disambiguate the word, which requires examining the context and performing a sematic analysis.

2.4 Miscellaneous. Some combinations of letters/diacritics do not follow the above rules. For instance, חַ /χa/ at the end of a word is always pronounced /aχ/. כְּח /χχ/ is often, but not always, pronounced /kχ/.

3. The experiment

We constructed an algorithm to transform dotted Hebrew to IPA.

3.1 The database. We tested our algorithm on a database of 36,358 dotted words provided by K Dictionaries. The database was created from the Hebrew dictionary core edited by Orna Ben Natan. Then the project managers, Anat Merdler-Kravitz and Yifat Ben-Moshe reviewed the automatically-transcribed words and amended them as necessary. As a result, each database entry consists of a word in both its dotted transcription and IPA counterpart.

The database consisted of 21,126 lemmas, represented by their base form. In addition there were some plural forms of nouns, verb inflections, and 184 multiword expressions. Many Hebrew conjunctions and prepositions are represented as prefixes in the standard script. Except for the latter, the words did not contain such prefixes.

3.2 Evaluation. The program correctly transcribed 32,736 words which consist of 90% of the database.

The miscellaneous category consists of 9 occurrences of כְּח that were incorrectly transcribed as /kx/, and some occurrences of עָ and הָ that were transcribed as the null character and /h/ instead of /ʔa/ and /ha/.

The main source of errors is the misplacement of stress.

The program correctly identified the stress position of all but 3% of the original Hebrew words in the sample. Loan words are the main source of errors. The program identified some of these words using heuristics, such as the existence of non-native consonants (ʤ, ʒ, ʃ), suffixes (t͡sija, nik,…), words starting with /f/ and other patterns that defy Hebrew phonotactics (a cluster of four consecutive consonants or three consecutive consonants at the beginning of a word). There are some diacritics (*hataf-qamatz*, *hataf-patax* and *hataf-segol*) that appear only in native words.

Total number of errors	stress	*schwa*	*qamatz*	miscellaneous
3,622	3,295	172	122	33
	91.0%	5.2%	3.4%	0.9%

Table 2: Distribution of errors

Total sample	loan words	program error	slang	database error
500	476	15	8	1
100%	95.2%	3%	1.6%	0.2%

Table 3: Distribution of stress errors on a random sample of 500 words

Since in most loan words (though not in all) the stress is penultimate, the program placed the stress there, thus eliminating a potential error.

As described above, the diacritic *Schwa* in Hebrew is sometimes pronounced /e/ and sometimes omitted (in Hebrew the diacritic *schwa* is never pronounced as a phonologist *schwa*). Hebrew phonologists used sonority theory to predict this behavior. Phonologists define sonority as the audible energy omitted with each phoneme (Burquest and Payne 1998, O'Grady and Archibald 2013). In each syllable the sonority rises until reaching the syllable's nucleus (usually a vowel) and then it falls. In English, the sonority scale, from highest to lowest, is the following:

a > e o > i u > r > l > m n ŋ > z v ð > s f θ > b d g > p t k.

Rosen (1957, following Segal), and later Boletzky (2007), postulated that when the onset of the syllable defies this order, i.e., the first phoneme is more sonorous than the second, the syllable is split by inserting the phoneme /e/ between the first and second phonemes. The sonority of the phonemes of the onset of each of the two syllables increases. Thus, for example, since /l/ is more sonorous than /v/, /lvi.ˈva/ becomes /le.vi.ˈva/, i.e., the syllable /lvi/ becomes two syllables /le/ and /vi/, thus causing the sonority of each syllable to increase.

To accommodate for the observed behavior of Israeli Hebrew, Rosen (1957) postulated the following sonority hierarchy for Hebrew:

a > e o > i u > j > l > m n > z v > f x χ > b d g > p t k > ʔ h

Rosen did not place /ʁ/ (r) and the silibants (s and ʃ) in this hierarchy. To properly place phonemes one should check whether an /e/ is inserted before or after the occurrences of the phoneme. Thus, we found that it is best

LOTKS 2017

Workshop on Language, Ontology, Terminology and Knowledge Structures

On September 19th the second edition of the Language, Ontology, Terminology and Knowledge Structures (LOTKS) workshop will take place as a satellite workshop of the 12th International Conference on Computational Semantics (IWCS) in Montpellier, France. Following on from a successful first edition as a joint workshop at LREC 2016, the intention is once again to provide a forum for different research communities to interact and discuss issues within the intersection of computational linguistics, ontology engineering, knowledge modelling and terminologies.

LOTKS grew out of the need for a workshop that dealt, on the one hand, with enhancing knowledge bases or conceptual schemes with linguistic knowledge, as well as on the other, the growing use of ontologies and concept schemes to enrich linguistic or lexical datasets -- in particular computational lexicons.

The workshop also offers showcasing the use of conceptual/terminological/ ontological resources in NLP or computational linguistics in general. This year we have introduced new themes relating to the use of terminology schemes and ontologies in the digital humanities. The workshop welcomes contributions from both academics and industry professionals.

Fahad Khan
Istituto di Linguistica Computazionale (A. Zampolli) – CNR

https://langandonto.github.io/langonto-termiks-2017/

to place /r/ together with /l/. However, since such a split never occurs before or after s, ʃ it is not possible to place the silibants to conform to this rule.

We tested this sonority rule on our database (omitting syllables with silibants in the onset). The theory successfully predicted the omission/inclusion of /e/ in 88.3% of words.

We have developed an alternative algorithm with better performance: When *schwa* immediately follows the first letter it is pronounced /e/ if and only if at least one of the following occurs:

- The first phoneme is a word prefix, such as b (=in) v (=and).
- The first phoneme is a verb conjugation prefix, e.g., tsaˈper te.sa.ˈper, *you will tell* = t (future, 2nd person)+saˈper.
- The first phoneme is j,l,m,n,r.
- The second phoneme is ʔ,h,ʕ .
- If *schwa* occurs elsewhere it is pronounced /e/ if and only if it is:
  - The second *schwa* in the pattern C1 *schwa* C2 *schwa* C3 (Ci a consonant).
  - Between two identical or similar letters (e.g., between /d/ and /t/)

The first two rules require a morphological analyzer to identify the correct analysis of a word in context. Since we did not have at our disposal a morphological analyzer for dotted texts, we could not apply these rules, which could have prevented at least 49 errors. The verb conjugation prefixes with *schwa* are t,j,l,n,m. With the exception of /t/ the prefix has high sonority and should, in most cases, cause a syllable break. Thus the second rule is often subsumed by the third. (This explains the low number of errors when rules 1-2 are ignored.) Since the number of remaining errors was small, we were able to manually identify when rules 1-2 were applicable, thus obtaining an error rate of less than 1%.

Qamatz

The diacritic *qamatz* is most often pronounced /a/ (*big qamatz*). The database contained 199 occurrences where *qamatz* is pronounced /o/ (*small qamatz*). We used two heuristics to identify (some of) them: The *qamatz* was followed by a consonant with the diacritic *hataf-qamatz* (which is always pronounced /o/). Thus, the pattern was /oCo/.

The consonant after the *qamatz* had a *schwa* and the following consonant had a *dagesh* (that indicates germination or strong pronunciation). Thus, the pattern was *qamatz* C1 *schwa* C2 *dagesh*. Hebrew grammar dictates that the *dagesh* is a light *dagesh* and C2 is either ב,ג,ד,כ,פ,ת.

This allowed us to identify 74 cases of /o/ (37.2%). The *small qamatz* is relatively rare, appearing in only 0.6% of all words of the database and in only 3% of the errors.

Conclusions

We have constructed an algorithm to transcribe dotted Hebrew texts to IPA conforming to the observed Israeli Hebrew pronunciation. The algorithm was implemented as a Python 3 program and is available from the author. The program was tested on a large database and the error rate was 11.2%.

We used the database to test how well sonority theory explains the pronunciation of *schwa*, and have formulated a simple alternative algorithm that outperforms the sonority theory algorithm.

References

Bolotzky, S. 2007. The sonority in the phonology of Israeli Hebrew. In *Hebrew and her sisters*, Efrat, M. (ed.). Haifa University Press, 239-248.

Burquest, D. A., and Payne, D. L. 1998. *Phonological analysis: A functional approach*. Dallas, TX: Summer Institute of Linguistics.

O'Grady, W. D., and Archibald, J. 2013. *Contemporary linguistic analysis: An introduction*. (7th ed.). Toronto: Pearson Longman, 70.

Rosen, H. 1957. *Ha-Ivrit Shelanu*, (*Our Hebrew*). Tel Aviv: Am Oved.

	Sonority theory w/o silibants	The alternative algorithm		
		Rules 3-6 w/o silibants	Rules 3-6 with silibants	Rules 1-6 with silibants
Sample size	7449	7449	8612	8612
# errors	871	125	126	77
% error	11.69%	1.68%	1.46%	0.89%

Table 4: Sonority theory and the alternative algorithm for words with ***schwa***