Probabilities improve stress-prediction in a CFG of Hawaiian phonology

'Ōiwi Parker Jones
Phonetics Laboratory
41 Wellington Square
Oxford, OX1 2JF
United Kingdom
`oiwi.parkerjones@ling-phil.ox.ac.uk`

1 Introduction

Should probability play a role in linguistics? Whereas Chomsky (1957: 16–17) influentially rejected probability in syntactic theory, "[i]n phonological theory, probability has not so much been rejected as disregarded" (Coleman, 2003: 89). Disregard has, however, given way to a growing literature on the use of probability across the various linguistic sub-disciplines (see, e.g., Bod et al., 2003; Coleman, 2003).

This paper is a case-study of probability in phonology, both as it applies to an improved description of Hawaiian stress-assignment, and as this description, in turn, reflects back on the probability question, above.

2 Grammars

By formalizing two strongly equivalent analyses, where one is a non-probabilistic Context-Free Grammar (CFG) and the other is a Stochastic Context-Free Grammar (SCFG) (Booth, 1969; Suppes, 1970), we can put the probability question to a test. For a given data set, if the SCFG does not outperform its strongly equivalent CFG, then parsimony should compel us to reject, rather than disregard, the added probabilities.

On the other hand, should the SCFG outperform its strongly equivalent, non-probabilistic CFG, then we ought, at least, to accept some role for probability in phonology; this would support the growing literature mentioned above.

Let our data set be Hawaiian. Schütz (1978, 1981) argues that Hawaiian stress-assignment is not 100% predictable, based on words like /ma.ˌku.a.ˈhi.ne/ 'mother' and /ˌʔe.le.ma.ˈku.le/ 'old man'. It might help to illustrate this argument by developing Schütz's analysis into a CFG.
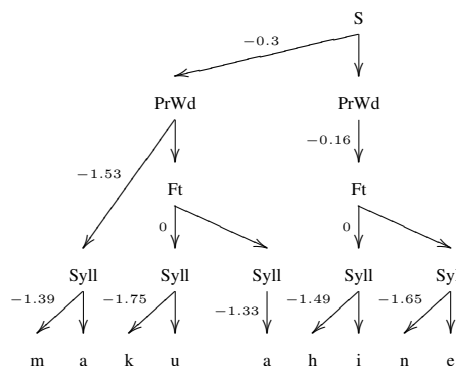

Figure 1: /ma.ˌku.a.ˈhi.ne/ parse-tree.

The crucial parse-trees for 'mother' and 'old man' are in Figures 1–4. Note that the terminal symbols are phonemes. The non-terminal symbols are syllables (Syll), metrical-feet (Ft), prosodic words (PrWd), and the start symbol (S). Also note that the leftmost syllable in each metrical-foot is stressed. The rightmost stress in a word is primary. Finally, let us ignore the labeled branches for the moment, as they do not apply to the non-probabilistic CFG.

The argument follows. The correct parse for 'mother' (Figure 1) is paralleled by an incorrect parse for 'old man' (Figure 2); except for their terminal expansions, these parse-trees have the same structure. Thus, the correct parse for 'mother' implies an incorrect parse for 'old man'. Moreover, the correct parse for 'old man' (Figure 3) implies an incorrect parse for 'mother' (Figure 4). Therefore, no matter how we order the CFG rules, a procedural interpretation will get either 'mother' or 'old man' wrong. Hence, Hawaiian stress-assignment is not 100% predictable.

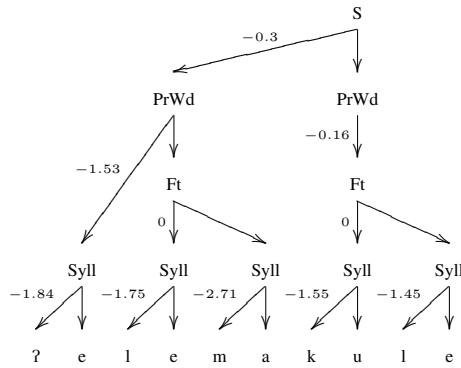Although this conclusion might be true (and

S
−0.3
PrWd   PrWd
−1.53   −0.16
Ft   Ft
0   0
Syll Syll Syll   Syll Syll
−1.84 −1.75 −2.71   −1.55 −1.45
ʔ e l e m a k u l e

Figure 2: */ʔe.ˌle.ma.ˈku.le/ parse-tree.

S
−0.3
PrWd   PrWd
−0.36   −1.31
Ft   Ft
0   0
Syll Syll   Syll Syll Syll
−1.5 −1.84   −1.6 −1.48
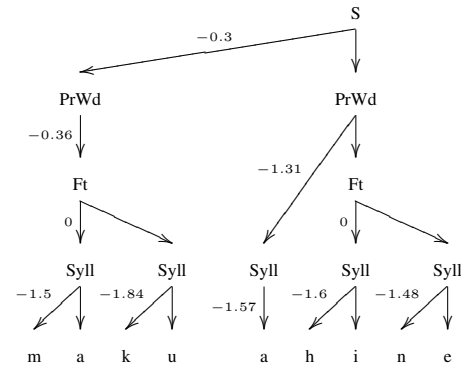−1.57
m a k u a h i n e

Figure 4: */ˌma.ku.a.ˈhi.ne/ parse-tree.

there is nothing here to disprove it), the SCFG turns out to be better than its strongly equivalent CFG at predicting stress-assignment in Hawaiian.
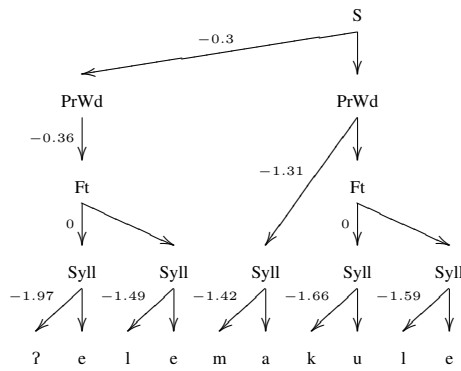
S
−0.3
PrWd   PrWd
−0.36   −1.31
Ft   Ft
0   0
Syll Syll   Syll Syll Syll
−1.97 −1.49   −1.42 −1.66 −1.59
ʔ e l e m a k u l e

Figure 3: /ˌʔe.le.ma.ˈku.le/ parse-tree.

In Figures 1–4, each labeled branch expresses the base-10 log probability for some SCFG rule, where the probabilities were obtained by training the grammar on data from a Hawaiian dictionary (Pūkuʻi and Elbert, 1986). The probability of a parse-tree is just the sum probability of its rules, so Figure 2's probability is −11.29. By contrast, Figure 3's probability is −10.09. The SCFG correctly picks /ˌʔe.le.ma.ˈku.le/ over */ʔe.ˌle.ma.ˈku.le/, since a log probability of −10.09 is higher than a log probability of −11.29. Moreover, the SCFG correctly picks /ma.ˌku.a.ˈhi.ne/ over */ˌma.ku.a.ˈhi.ne/, since the log probability of −9.59 is higher than that of −9.95. In both examples, the SCFG correctly disambiguates the parses.

3 Evaluation

In a computational evaluation of 16,900 Hawaiian words, the CFG correctly parsed 84.6%. However, the SCFG correctly parsed 97.38%. These results demonstrate that probabilities improve stress-prediction in a CFG of Hawaiian phonology; there is a role for probability in phonology.

Acknowledgments

References

R. Bod, J. Hay and S. Jannedy, eds. 2003. *Probabilistic Linguistics*. MIT Press, Cambridge, MA.

T. L. Booth. 1969. Probabilistic representation of formal languages. *Tenth Annual IEEE Symposium on Switching and Automata Theory*. Pp. 74–81.

N. Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

J. S. Coleman. 2003. Commentary: probability, detail and experience. In J. Local, R. Ogden, and R. Temple, eds. *Phonetic Interpretation: Papers in Laboratory Phonology VI*. Cambridge University Press, Cambridge. Pp. 88–100.

M. K. Pūkuʻi and S. H. Elbert. 1986. *Hawaiian Dictionary*. University of Hawaiʻi Press, Honolulu.

A. J. Schütz. 1978. Accent in two Oceanic languages. *Anthropological Linguistics*, 20(4): 141–149.

A. J. Schütz. 1981. A reanalysis of the Hawaiian vowel system. *Oceanic Linguistics*, 20(1): 1–43.

P. Suppes. 1970. Probabilistic grammars for natural languages. *Synthese*, 22: 95–116.