Alien Phonotactics:
What can Science Fiction tell us about implicit knowledge?

Wendell Kimper :: University of Manchester

Introduction

The broad question: does implicit phonotactic knowledge contain information about *markedness*, beyond what can be induced from the lexicon?

- ☆ Surfeit of the stimulus: Becker et al. (2011); Hayes and White (2013) find evidence that phonotactic judgements rely more on 'natural' generalisations than unnatural ones (though cf. Hayes et al. 2009).
- ☆ But is naturalness directly encoded?

This project applies novel data from naming practices in science fiction to that question.

Markedness Subversion

Why use science fiction to answer this?

- ☆ Phonotactic novelty or oddity contributes to the 'alien-ness' of non-human characters.
- ☆ As in e.g. language games (Bagemihl, 1988), this is a linguistic context where subversion of markedness is expected.

In written media, orthotactic novelty is available — e.g. *R'lyeh* and *Cthulhu* (Lovecraft, 1928).

This may not be directly connected to phonotactics...

> MELLLVAR *Sign it to Mellllvar. Mellllvar has three L's*
> GEORGE TAKEI: *(sigh) I think I've done enough conventions to know how to spell Mellllvar.*
> *— Futurama 4x11, 'Where No Fan Has Gone Before'*

The Database

1,593 phonetically-transcribed proper names of alien characters from the Star Trek franchise (prime universe canon).

- ☆ Initial list names compiled from Memory Alpha (a fan wiki).
- ☆ Located (with the help of scripts and fan transcripts) and transcribed in IPA (also converted to ARPABET).
- ☆ Names that were missing from initial list and alternate pronunciations were added as they were encountered.

Excluded: descriptions using English words (e.g. The Crystalline Entity, The Caretaker), names with clear human provenance in the plot (e.g. Alexander Rozhenko).

What does this data represent?

Large data set from a (somewhat) consistent fictional universe.

- ☆ Diverse sources: 374 writers contributed to episodes and films.
- ☆ Only a handful of species had a deliberately constructed language (e.g. Klingon) — many were one-off 'alien of the week'.

These names are the result of a form of creative word-play by (relatively) linguistically naïve individuals; it's likely that they represent non-analytical intuitions.

Corpus Analysis

Individual segment frequencies were extracted using Phonological Corpus Tools (Hall et al., 2015).

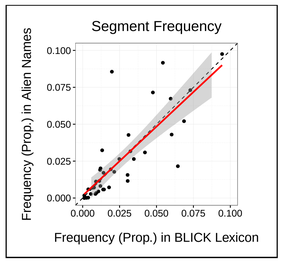
The English lexicon used for reference was the lexicon used to create the BLICK phonotactic probability calculator (Hayes, 2012), which consists of all entries in the CMU Pronouncing Dictionairy with a CELEX frequency of at least one.

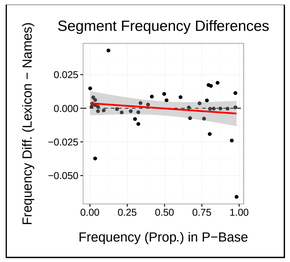
Typological frequencies were obtained using P-Base (Mielke, 2008).

Segment Frequencies

Individual segment frequencies in the alien names were highly coorrelated with those in the English lexicon (r = 0.80), with the 95% CI indicating that the slope is indistinguishable from 1.

Frequency differences between alien names and the English lexicon did not differ along the dimension of typological frequency, with a slope indistinguishable from 0.

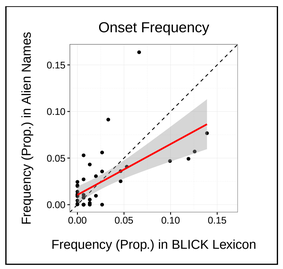


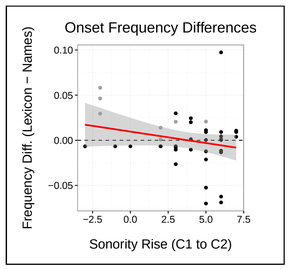


Onset Clusters

The frequency of C_1C_2 word-onset clusters was also fairly closely correlated with the English lexicon (r = 0.61), but with a noticeably shallower slope — suggesting that lower-frequency English onsets are over-represented in the alien names.

Differences between cluster frequency in the alien names and in the English lexicon were examined according to the rise (or fall) in sonority between C_1 and C_2. The magnitude of these differences ddi not differ according to sonority profile, with a slope indistinguishable from 0.

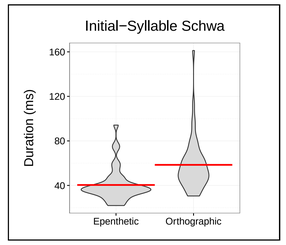




Phonotactic Repairs

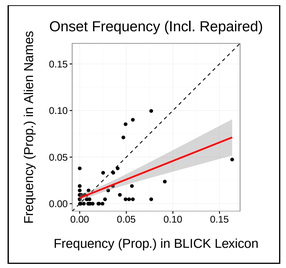
Names with complex onsets in the orthography were (nearly invariably) repaired via epenthesis by actors in production.

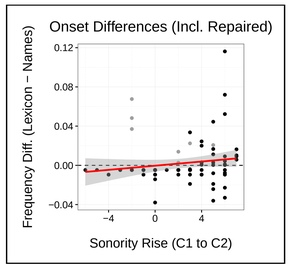
Durations of initial-syllable [ə] were measured in a sample of 86 names — 44 with corresponding orthographic vowels, and 42 which corresponded to complex onsets in the orthography. For both categories, the sonority profiles of the flanking consonants ranged from a fall of 5 to a rise of 6 steps along the sonority scale.



There was a statistically significant difference bewteen the orthogrphic and epenthetic vowels (t = 4.168, $p < 0.001$), with the former approximately 18ms longer than the latter — well above the Just Noticeable Difference of 5ms (Nooteboom and Doodeman, 1980).

With the 'underlying' clusters from the epenthetically-repaired (but likely recoverable) items added to the set of onset clusters for the alien names, the extent to which low-English-frequency onsets are over-represented in the alien names is increased. The slope of the relationship between sonority profile and this over-representation, however, remains indistinguishable from 0.

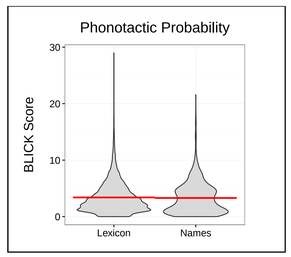


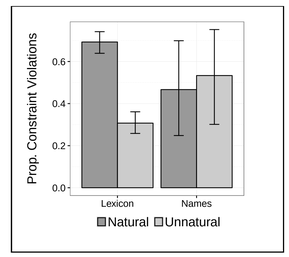


Phonotactic Novelty

The BLICK phonotactic probability calculator (using the default grammar) was used to assign violation scores to each item in both the alien names database and the English lexicon. There was no overall difference beween the BLICK scores.

Items violating the specific set of 'natural' and 'unnatural' constraints from Hayes and White (2013) were identified in both the English lexicon and the alien names database. The alien names showed no difference.

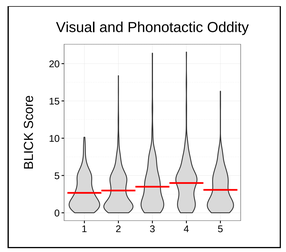




Degrees of Visual Novelty

Alien species were classified along a 10-point scale of visual 'alien-ness' — only the lower half of the scale had sufficient numbers of named characters for analysis.

1 2 3 4 5



Despite the limited range, there was a statistically significant relationship between visual and phonotactic novelty (t = 3.152, $p < 0.01$), with increasingly visually alien characters receiving slightly more phonotactically novel names.

Conclusion

An analysis of the phonotactics of alien names in the Star Trek franchise suggests that writers are relying on implicit phonotactic knowledge in conveying 'alien-ness'.

In particular, word onsets with low frequency in English are over-represented in alien names, and illicit onsets are repaired recoverably, consistent with markedness subversion.

However, there is no evidence to suggest that the markedness involved is distinct from the lexical statistics of English.

References

Bagemihl, Bruce. 1988. Alternate phonologies and morphologies. Doctoral Dissertation, University of British Columbia.

Becker, Michael, Nihan Ketrez, and Andrew Nevins. 2011. The surfeit of the stimulus: analytic biases filter lexical statistics in turkish laryngeal alternations. *Language* 87:84–125.

Hall, Kathleen Currie, Blake Allen, Michael Fry, Scott Mackie, and Michael McAuliffe. 2015. Phonological corpustools, version 1.1.1 [computer program].

Hayes, Bruce. 2012. *Blick: a phonotactic probability calculator*. UCLA.

Hayes, Bruce, and James White. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* 44.

Hayes, Bruce, Kie Zuraw, Péter Siptár, and Zsuzsa Cziráky Londe. 2009. Natural and unnatural constraints in Hungarian. *Language* 85:822–863.

Lovecraft, H. P. 1928. The call of cthulhu. *Weird Tales* 11:159–78.

Mielke, Jeff. 2008. *The emergence of distinctive features*. Oxford.

Nooteboom, SG, and GJ Doodeman. 1980. Production and perception of vowel length in spoken sentences. *Journal of the Acoustical Society of America* 67:276–87.

Acknowledgements

Thanks to Emily Driscoll for help compiling the intial list of characters' names, and to the NYU PEP Lab for insightful discussion on the very early stages of this research.

wendell.kimper@manchester.ac.uk