Alien Phonotactics: What can science fiction tell us about implicit knowledge?

Document Version
Final published version

Link to publication record in Manchester Research Explorer

Citation for published version (APA):

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
Alien phonotactics: What can Science Fiction tell us about implicit knowledge?

Wendell Kimper
University of Manchester

1. Introduction

A long-standing question in phonological theory is the extent to which markedness principles are encoded in synchronic grammars or in the minds of speakers. Evolutionary phonology (Blevins 2004, and others) argues that phonetic substance plays a role in sound change, but not in synchronic grammar. On the other hand, Grounded Phonology (Archangeli & Pulleyblank 1994) argues that principles of phonetic grounding play a crucial role in constraining the set of possible synchronic grammars. Synchronic representation of grounding principles is also an important aspect of Optimality Theory (Prince & Smolensky 1993/2004), expressed as a set of universal markedness constraints in CON.

In this paper, I apply novel data from naming practices in Science Fiction (SF) to the question of synchronic representation of markedness. Because alien characters’ names must be consistent with the sense of otherness that is crucial to their narrative function, they represent a situation where violating phonotactic expectations is desirable. This provides an opportunity to examine whether these expectations are based on solely language-specific statistical regularities or involve more general principles of phonetic grounding. A corpus analysis of proper names of non-human characters in the Star Trek franchise provides evidence that some phonotactically rare or ill-formed structures are over-represented in alien names, but finds no evidence that universal markedness principles are involved.

2. Background

2.1 Implicit Knowledge and Markedness

Does language users’ implicit knowledge of phonotactics include information about markedness or phonetic grounding?

*Many thanks to Emily Driscoll for help compiling the initial list of names, and to the NYU PEP Lab for insightful discussion on an early stage of this research.*
Lab-based Artificial Grammar (AG) studies approach the question by training subjects on novel phonological patterns. For example, Wilson (2006) found asymmetrical generalisation of velar palatalisation that mirrors an implicational triggering relationship found in typology — subjects trained on a \([ki] \rightarrow [t\text{s}]\) pattern did not generalise to \([ke]\), but subjects trained on a \([ke] \rightarrow [t\text{s}]\) pattern palatalised \([ki]\) and \([ke]\) equally. Wilson shows that a model of learning which includes biases towards phonetically-grounded patterns matches the data from human learners better than an unbiased model. Other AG studies have shown similar effects — however, Moreton & Pater’s (2012) review of the literature on substantive bias suggests that overall results have been mixed.

Other studies have looked for evidence of markedness subjects’ native-language phonotactic knowledge. Becker et al. (2011) find evidence of surfeit of the stimulus effects — native speakers of Turkish extended phonetically ‘natural’ patterns to novel words, while equally statistically robust (but phonetically unnatural) generalisations in the Turkish lexicon were not extended. Hayes et al. (2009) and Hayes & White (2013) find similar (though less extreme) effects — statistically regular but unnatural patterns were extended, but to a lesser extent than natural patterns.

Martin (2007) finds evidence for markedness in instances of lexical competition, arguing that less marked words will edge out more marked alternatives. Particularly relevant here is his finding that a bias against multiple identical liquids influences American infant naming trends over time — names like Leila or Rory are (a) less likely to become popular, and (b) more likely to fall out of popularity than names with non-identical liquids (such as Laura or Riley). He finds that identical liquids are less frequent than expected by chance in unusual or ‘made up’ baby names, novel names for prescription medications, and character names for a fantasy role-playing game, suggesting that situations where speakers are faced with a degree of lexical freedom are influenced by markedness-based biases.

2.2 Markedness Subversion

Why are alien names useful for approaching this issue?

Bagemihl (1988) points out that language games often make use of phonological operations that are not found in natural languages, such as string reversal or long-distance metathesis. They are, in effect, a form of encryption — one social role that language games fulfill is to restrict access by rendering the message uninterpretable to the uninitiated. Because the aim is obscuration rather than transmission, natural language markedness principles favouring ease of transmission are a liability rather than an asset, and we would expect to find patterns which contravene them. I will refer to this phenomenon as markedness subversion.

Naming practices for non-human characters in SF represents another context where markedness subversion might be expected. As Sardar & Cubitt (2002, p.6) observe, “difference and otherness are the essence of aliens” — their narrative roles often rely heavily on that sense of difference, and the names assigned to them have potential to aid in constructing these characters as ‘other’ by violating phonotactic expectations.

In written media, orthotactic expectations are available as a tool in constructing otherness — for example, the use of the apostrophe in R’lyeh and the orthotactically illegal
Alien phonotactics

strings in *Cthulhu* (Lovecraft 1928). Furthermore, there is some reason to believe that this use of orthography is a trope that SF media consumers are consciously aware of, since it features in genre parodies — as in the exchange in (1).

(1) MELLVAR: Sign it to Melllvar. Melllvar has three L’s
    GEORGE TAKEI: (sigh) I think I’ve done enough conventions to know how to spell Melllvar.
    (Goodman et al. 2002)

In audiovisual media, however, orthography is largely unavailable,¹ and the use of naming to create a sense of otherness is dependent on phonotactics. As in (1), deviant orthographic forms do not necessarily translate straightforwardly into identifiably deviant phonological forms. To the extent possible, orthography is set aside in this paper — the aim is to investigate how the construction of otherness in alien characters’ names makes use of implicit phonotactic knowledge.

The use of markedness subversion in alien names provides a novel opportunity to ask questions about the nature of this implicit phonotactic knowledge. Do these names derive their otherness solely from departures from native language-specific phonotactic expectations? Or do they also make use of expectations based in non-language-specific principles of markedness?

3. Methods

3.1 The Database

The database used in this paper consists of 1,593 phonetically-transcribed proper names of alien characters from the Star Trek franchise.² To assemble this database, an initial list of character names for non-human aliens (in their orthographic form) was compiled using the fan wiki Memory Alpha.³ These names were located within each episode and a broad IPA transcription was made by ear by the author. Where names appeared more than once, the first appearance within the series was used; subsequent instances with variant pronunciations were also included as separate entries and tagged as alternates. For analysis, IPA transcriptions in the database were converted to ARPABET.

Forenames and surnames (where both were present) were entered separately and tagged accordingly; surnames associated with multiple characters were only counted once. Names were excluded from the database if they consisted of descriptions using English words (e.g. The Crystalline Entity, The Caretaker) or if clear human provenance for the name was

---

¹Though see Section 4.3 for further discussion on orthographic recoverability.
³http://memory-alpha.wikia.com
indicated in the narrative (e.g. Alexander Rozhenko). Names were also excluded from the present analysis if they only appeared in written form (i.e. as a character name in a script, or shown in on-screen visual displays).

3.2 What does this data represent?

How should we understand the cognitive processes that resulted in these observations? While these names are all drawn from the same fictional universe, they represent contributions from a fairly wide range of individuals — a total of 374 writers contributed to the television episodes and films used in creating the database. In addition, more than 250 fictional alien species are represented in the data set\textsuperscript{4} — with the exception of Klingon (Okrand 1985), the vast majority of these do not have deliberately constructed languages, and many species are only encountered in a single episode.

Because of this, it’s likely that the orthographic forms of the names in the database are the result of a form of creative word-play by (relatively) linguistically naïve individuals, representing non-analytical intuitions. In addition, these are then uttered by actors (mostly native speakers of English) — this results in another layer of naïve intuition, with native-language phonotactic restrictions constraining the set of possible phonological forms. Each name thus represents a combination of the writer’s word-play and the actor’s performance.

Determining precisely which cognitive processes are reflected in this data is not entirely straightforward. However, if language users have implicit knowledge of markedness that shapes the extent to which phonotactic sequences elicit intuitive judgements of oddity, this is a context were we might expect that to be evident.

4. Analysis

The corpus analysis in this paper examines two areas of potential interplay between language-specific statistical learning and universal markedness: (a) individual segment frequency, and (b) syllable structure — specifically, the role of the Sonority Sequencing Principle (SSP) in word-initial onset clusters.

4.1 Segment Frequencies

If otherness in alien names is signalled by departure from expectations based in the English lexicon, we would expect to see that segments which are uncommon in English should be over-represented in alien names. If otherness is signalled by departure from principles of universal markedness, we would expect to see segments that are typologically rare over-represented.

The English lexicon used for comparison consisted of all entries in the CMU Pronouncing Dictionary with a CELEX frequency of at least one — the same lexicon that the BLICK phonotactic probability calculator (Hayes 2012) uses. Phonological Corpus Tools (PCT; Hall et al. 2015) was used to extract individual segment counts from both the alien

\textsuperscript{4}Not all characters’ species are identified or named.
names database and the English lexicon (with frequencies expressed as proportions of the total count), and P-Base (Mielke 2008) was used to determine typological frequencies.

The segment frequencies found in alien names closely resemble those found in English ($r = 0.80$). To obtain a measure of divergence from English, each segment’s frequency in the English lexicon was subtracted from its frequency in the alien names database; positive values on this measure indicate that a segment is over-represented in alien names, while negative values indicate under-representation. The figure in (2) shows this measure plotted against frequency in the English lexicon and frequency in P-Base.

(2) Segment frequency divergence from English as a function of English lexical frequency and typological frequency. Dashed lines show a slope and intercept of 0 for reference, and shaded areas around (bold) regression lines show 95% CIs.

If segments which are rare in English are over-represented in alien names, a negative slope is expected for lexical frequency. However, a linear regression shows that there is no statistically detectable effect ($t = -0.559$, $p = 0.68$). Similarly, if segments which are typologically rare are over-represented in alien names, a negative slope is expected for typological frequency, but this is also not found ($t = 1.028$, $p = 0.41$).

4.2 Onset Clusters

If otherness is alien names is accomplished via language-specific phonotactics, we would expect to see syllable structures that are rare (or prohibited) in English over-represented. If more general principles of markedness are involved, we would expect to see typologically marked syllable structures (for example, violations of the SSP) over-represented.

PCT was used to extract counts for word-initial $C_1C_2$ (\_V) sequences from both the alien names database and the BLICK lexicon. Frequencies for these sequences showed a reasonably large correlation ($r = 0.61$). As with segment frequencies, the frequency for $C_1C_2$ sequence in the English lexicon was subtracted from its frequency in the alien names to obtain a measure of divergence from English. Each consonant was assigned a position on
the sonority hierarchy in (3), and the degree of sonority rise was calculated by subtracting $C_1$ from $C_2$.

(3)  \textit{Sonority Hierarchy (Selkirk 1984)}

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>voiceless plosive</td>
</tr>
<tr>
<td>2</td>
<td>voiced plosive</td>
</tr>
<tr>
<td>3</td>
<td>voiceless fricative</td>
</tr>
<tr>
<td>4</td>
<td>voiced fricative</td>
</tr>
<tr>
<td>5</td>
<td>nasal</td>
</tr>
<tr>
<td>6</td>
<td>liquid</td>
</tr>
<tr>
<td>7</td>
<td>rhotic</td>
</tr>
<tr>
<td>8</td>
<td>glide</td>
</tr>
</tbody>
</table>

The figure in (4) shows the difference in frequency between English and alien names plotted against frequency in the English lexicon and degree of sonority rise.

(4)  $C_1C_2$ frequency divergence from English as a function of English lexical frequency and sonority rise. Dashed lines show a slope and intercept of 0 for reference, and shaded areas around (bold) regression lines show 95% CIs.

If initial $C_1C_2$ sequences which are rare in English are over-represented in alien names, we would expect to find a negative slope for lexical frequency. Indeed, a linear regression finds a statistically significant effect in the expected direction ($t = -2.296, p < 0.05$). This includes a number of initial clusters that are not attested in the BLICK lexicon. One of these can plausibly be attributed to an accidental gap in the lexicon used: [dj] is absent from the BLICK lexicon, but attested in (esp. British English) pronunciations of words such as \textit{durable} or \textit{duel}. Others have a marginal status in English: [jm] is present in Yiddish substrate words such as \textit{schmuck} and \textit{schmaltz}, [vl] is found in borrowed items such as \textit{Vladimir}, and [vr] is attested in the ideophone \textit{vroom}. However, [kf], [tp], [zl], [ng],
Alien phonotactics

[lw], and [gj] (each of which had a count of 1 in the alien names database) are all clearly phonotactically ill-formed in English.

If initial $C_1C_2$ sequences which violate the SSP are over-represented, a negative slope for sonority rise is expected — however, this is not found ($t = 1.509$, $p = 0.21$). It would seem, then, that while infrequent (or even unattested) word onsets in English are over-represented in alien names, this is not necessarily related to sonority-based markedness principles.

4.3 Phonotactic Repairs

As described above, a small number of alien names with phonotactically-illicit word onsets in their phonological realisations were present in the database. However, a greater number of items with orthographic forms which might suggest illicit word onsets were realised with initial $C_1\sigma C_2$ sequences.

Instances of word-initial $C_1\sigma C_2$ with no orthographic vowel corresponding to the [ɔ] were identified, and considered examples of a vowel-epenthesis repair. The ‘faithful’ $C_1C_2$ clusters and their counts were reconstructed and added to the data set, and the analysis from Section 4.2 was repeated; the result is plotted in (5).

(5) $C_1C_2$ frequency divergence from English as a function of English lexical frequency and sonority rise. Dashed lines show a slope and intercept of 0 for reference, and shaded areas around (bold) regression lines show 95% CIs. Reconstructed clusters are shown in grey.

With the reconstructed clusters included, there is a larger (and more statistically robust) effect of English lexical frequency ($t = -9.380$, $p < 0.001$). However, there is still no detectable effect of sonority rise ($t = -1.346$, $p = 0.18$).

---

5 Most, but not all, contained an apostrophe in the orthographic form. Three out of the six phonotactically illicit onsets that were uttered faithfully also contained apostrophes.

6 The affricates tʃ and dʒ were treated as plosives for the purpose of sonority classification.
Kimper

Is this reconstruction of ‘underlying’ or orthographic clusters something that could plausibly be attributed to listeners, particularly in light of their likely awareness of orthographic conventions in SF? To answer this question, durations [ə] in word-initial C₁əC₂ sequences were measured in a sample of 86 names — 42 which corresponded to complex onsets (or apostrophes) in the orthography (epenthetic vowels), and 44 with corresponding orthographic vowels (where [ə] is the result of unstressed-vowel reduction). The sample of orthographic vowels was approximately matched to the epenthetic items with respect to the sonority of the flanking consonants, and 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.

(6) Duration of initial-syllable schwa in epenthetic (corresponding to no orthographic vowel) and orthographic contexts. Horizontal bars indicate group means.

There was a statistically significant difference between the orthographic and epenthetic vowels (t = 4.168, p < 0.001), with the former approximately 18ms longer than the latter — well above the Just Noticeable Difference (JND) of 5ms (Nooteboom & Doodeman 1980). While this does not indicate that listeners are able to (or do) reconstruct the orthographic forms, it does demonstrate that this is a possibility.

5. Conclusion

This paper has presented the results of a corpus analysis of alien names in the Star Trek franchise. Individual segments showed no effects of either frequency in English or typological frequency. Syllable onsets showed an effect of native-language frequency — C₁C₂ clusters which are rare or unattested in English are over-represented in alien names, consistent with the use of phonotactics to establish a sense of otherness in constructing alien characters. However, there was no detectable effect of SSP well-formedness, suggesting that broader markedness factors are not involved.

While a small number of phonotactically illicit syllable onsets were produced, it was broadly the case that native English-speaking actors produced names which had ortho-
Alien phonotactics

graphically indicated illicit clusters with a vowel epenthesis repair. Interestingly, these repairs may be recoverable — epenthetic vowels were statistically significantly shorter than their counterparts with corresponding orthographic vowels, and the mean difference was above the JND threshold. If the repaired clusters are considered recoverable and counted in their ‘underlying’ forms, the effect of English phonotactic frequency becomes more robust, but there is still no detectable effect of sonority.

This paper has covered only a small portion of the empirical ground opened up by the use of this type of data, and further research is warranted. It would be unwise to conclude prematurely that there can be no use of principles of universal markedness or phonetic grounding in cases of markedness subversion. However, the results presented here are cause for scepticism.

References

Hall, Kathleen Currie, Blake Allen, Michael Fry, Scott Mackie, & Michael McAuliffe. 2015. Phonological corpus tools, version 1.1.1 [computer program].
Hayes, Bruce. 2012. *BLICK: A phonotactic probability calculator*. UCLA.


Wendell Kimper
wendell.kimper@manchester.ac.uk