Categorical account of gradient acceptability of word-initial Polish onsets

Kalina Kostyszyn and Jeffrey Heinz
Department of Linguistics and Institute of Advanced Computational Science at Stony Brook University

1 Introduction

Polish is known for complex onsets that defy principles of sonority sequencing (Gussman, 2007; Jarosz, 2017; Zydorowicz & Orzechowska, 2017). This is a feature prominent among Slavic languages, for example including Slovak (Bárkányi, 2011). Here, we examine phonotactic acceptability of these onsets in word-initial position, as well as the ability of three different phonotactic learning models to predict speaker judgments. In particular, we compare how well categorical and gradient phonotactic models match the results of a survey of speaker judgments on word-initial onsets in Polish.

Hayes & Wilson (2008) argue that gradient acceptability judgments shuold be accounted for with grammars that equate probabilistic likelihood with well-formedness. However, this kind of account has not been without criticism. Gorman (2013) compared Hayes and Wilson's maximum-entropy-based phonotactic learner (HWPL) with a categorical baseline and found the baseline accounted for the gradient judgments just as well, if not better than, HWPL. This baseline marked onsets and rhymes as well-formed only if they occurred in a representative sample of words.[1] More recently, Durvasula (2020) also argued that the categorical baselines perform at least equally well as HWPL, even showing that HWPL's difference in performance is negligible on (1) real input data with actual frequencies and (2) synthesized data where the forms have equalized frequencies.

While Gorman and Durvasula considered other languages in their broader arguments against equating gradient acceptability judgments with non-categorical grammars, the only language directly compared against HWPL in their categorical baselines was English. Because of the complex phonotactics at play in Polish clusters – in particular, elaborate defiance of sonority sequencing – we considered them an interesting testing ground for studying the ability of different phonotactic models to predict phonotactic acceptability. This paper examines the hypothesis that categorical grammars perform as well as, if not better than, gradient grammars in making these predictions.

A group of native Polish speakers were surveyed for their judgments on nonce words made with Polish-like onset clusters. Using a dictionary of Polish as a basis, three phonotactic learning models were trained. The first two are categorical and only pay attention to whether the word-initial onset and its bigram components, respectively, are present in this dictionary. The third was HWPL, which returned a weighted list of constraints represented by sequences of feature matrices. Given these three models, we compared the resulting correlations between the averaged acceptability scores for each cluster and the grammatical scores the grammars output by the learning algorithms assigned to each cluster. The results show that the categorical models generally outperformed the gradient model.

2 Previous Work

2.1 *Phonotactic modeling* Phonotactic learning algorithms can be classified according to various parameters. These include the representations they use for human speech, how they generalize from the input data, and whether the grammar they output is gradient or categorical.

Representations are important. A learning model must choose whether to represent constraints with phones as symbols or as collections of their individual features. Constraints could also be stated in terms of

[1] Gorman's thesis also argues that grammatically real phonotactic constraints are those which are primarily supported by phonological alternations.

Proceedings of *AMP 2021*

larger prosodic units such as onsets and rhymes.

Learning models must also specify what kind of syntagmatic relations it will consider among the phones. If a model examines phones in adjacent proximity, up to some window of size k, it's considered *strictly local* (Rogers & Pullum, 2011), but if it is instead limited to a subset of those phones projected to a tier, it is instead *tier-based strictly local* (Heinz et al., 2011; Lambert, 2021). Another syntagmatic relation that has been studied is general precedence which yields *strictly piecewise* patterns (Heinz, 2010; Rogers et al., 2010).

Another parameter is whether the grammar the learning algorithm outputs is categorical or gradient. Examples of categorical learning models include the ones introduced by Gorman (2013) and Durvasula (2020). Chandlee et al. (2019) also present a general method for finding inviolable constraints using any representational scheme in a data sample. As mentioned, HWPL (Hayes & Wilson, 2008) uses the principle of Maximum Entropy to return weighted constraints over featural representations which are used to assign probabilities to forms. More recently, Wilson & Gallagher (2018) augment the HWPL with gain-based constraint selection.

2.2 *Gradient vs. Categorical Models* It has become a standard perspective that phonotactics models should account for gradient judgments. Hayes & Wilson (2008:p. 382) explain.

> "All areas of generative grammar that address well-formedness are faced with the problem of accounting for gradient intuitions... Wen the particular domain of phonotactics, gradient intuitions are pervasive: they have been found in every experiment that allowed participants to rate forms on a scale... Thus, we consider the ability to model gradient intuitions to be an important criterion for evaluating phonotactic models."

The crux of this approach is that "well-formedness can be interpreted as probability," and that the "probabilities assigned to forms by a maxent grammar will correspond to the well-formedness judgments of native speakers, with lower probabilities for forms judged less acceptable" (Hayes & Wilson, 2008:p. 383). Additionally, they add that, because gradience surfaces in native speaker judgments of rare forms, a phonotactic model *must* model gradient intuitions. Because of the success of this seminal work, this approach has become standard for modeling phonotactic judgments.

In contrast to gradient grammars, categorical grammars put forms into binary categories. For example, a potentially outdated vowel-raising rule in Polish persists in novel words nonetheless, not by its own productivity, but because novel words may express a similarity to words known by the speaker that exhibit this vowel-raising (Buckley, 2001). Here, the categories of interest were 'vowel-raising in inflection' and 'no vowel raising in inflection', and membership was decided based on similarity to pre-existing members of the group. Novel words that contained the appropriate environment to trigger raising, but did not match an existing word with raising, were in the 'no vowel-raising' category, and thus belonged to a different inflectional paradigm.

Gorman (2013) made several arguments against the position articulated by Hayes and Wilson. The first was that gradient observations are not clear-cut evidence that the underlying system is gradient or categorical, as it is known that human judgments in psycholinguistic tasks lead to gradient observations even if the underlying system is categorical (Armstrong et al., 1983). Second, he argued that evidence for gradient grammars would be obtained if HWPL (or any gradient grammar) could outperform categorical baselines. His analysis was that HWPL does not. Third, he argued that statistically under-represented structures in the lexicon alone are not sufficient evidence to establish a synchronically real phonotactic constraint. Gorman identifies two other sources of evidence of importance: psycholinguistic experimentation and effects of productive phonological processes (Gorman, 2013).

Heinz & Idsardi (2017) argue against equating well-formedness with probabilities on conceptual grounds. Their argument rests on the fact that longer well-formed strings are eventually less probable than shorter ill-formed strings.

Durvasula (2020), conducted two comparative studies of acceptability judgments on English onset clusters in previous work (Scholes, 1966; Albright, 2007). In one study, Durvasula compared HWPL with a simple, categorical phonotactic learner. The grammar for the categorical leaner contained all the observed single feature unigrams and bigrams, as well as all the observed *segment* unigrams and bigrams in the learning corpus (Durvasula, 2020:slide 36). The evaluations showed that this categorical model was at least as good as the UCLA Phonotactic Learner in accounting for the reported gradient judgments.

In a second study, Durvasula compared two grammars output by HWPL: one was trained with with actual frequencies, and the other trained with equalized frequencies (to neutralize the information frequencies provide). Durvasula showed that the grammar generated by the learner with frequencies equalized was able to make similar predictions to that of the model with the actual frequencies, concluding that distinctive frequency information was irrelevant. In fact, he argues that added gradience can harm the performance of the model by showing in some cases a model learning weighted constraints had weaker predictions than one learning categorical constraints.

In sum, the results by Gorman and Durvasula cast doubt on the claim that gradient phonotactic learning models are necessary to account for gradient acceptability judgments and are inherently better than categorical learning models.

3 Polish phonology

We investigate this question of categorical versus probabilistic grammars for gradient acceptability with data from Polish, which has a substantial number of word-initial consonant clusters.

3.1 *Polish* Polish, the primary language of Poland, is a West Slavic language and of significant interest to the study of phonotactics.

The vowel inventory of Polish is straightforward, with six oral vowels ⟨i, ɨ, ɛ, a, u, ɔ⟩, as well as two nasal counterparts to ⟨ɛ, ɔ⟩. These vowels are 'nasal' in the sense that they are followed by a nasalized glide, akin to a diphthong (Gussman, 2007). The exact quality of these nasal vowels is not important here.

Polish also has a significant number of consonants, including affricates and palatalized consonants (Gussman, 2007). The sonorants include ⟨m, mʲ, n, ŋ, ɲ, l, r, w, j⟩. The stops include ⟨p, b, t, d, k, g⟩, and their palatal counterparts ⟨pʲ, bʲ, tʲ, dʲ, c, ɟ⟩. The fricatives are ⟨f, v, fʲ, vʲ, s, z, ɕ, ʑ, ʃ, ʒ, c, x⟩. Finally, the affricates are ⟨t͡s, d͡z, t͡ɕ, d͡ʑ, t͡ʃ, d͡ʒ⟩.

There are some issues worth mentioning. The quality of some sibilants in Polish can be in question. Some authors, for example, use [ʃ] (Jarosz, 2017) and others use [s̪] (Gussman, 2007). For the sake of this project, we use [ʃ] but remain agnostic as to the specific quality. Palatalized ⟨ʃʲ, ʒʲ⟩, which are distinct from other palatalized sibilants, are also found, but typically only in loanwords (Gussman, 2007).

Polish allows for famously complex consonant clusters, adhering weakly – if at all – to the Sonority Sequencing Principle. These can include sonority plateaus (Jarosz, 2017) (as in [ktɔ], 'who', and [gdaɲsk], 'Gdańsk'), or sequences entirely defying sonority sequencing, which can occur word initially - as in [mgwa], 'fog' – or word-medially – as in [jabwkɔ], 'apple'.

Stress is predictably penultimate, regardless of syllable structure; occasionally, stress will occur in the antepenultimate position or in the final syllable, but as stress is not relevant to this present study, it will not be mentioned further.

With regards to the affricates, it bears mentioning as well that Polish speakers are able to differentiate between single segment stop-fricative affricates and multi-segment stop-fricative clusters. For example, the difference between [tʃɨstɨ] ('clean') and [t͡ʃɨstɨ] ('three hundred') is clear to native speakers, and relies wholly on affrication of the onset (Gussman, 2007).

3.2 *Yer-alternation* One of the reasons why word-initial consonant clusters in Polish are so complex has to do with the *yer*, and the processes surrounding it. A *yer* (sometimes written as *jer*, though both are acceptable) is a super-short vowel found in Slavic languages that alternates between some non-high vowels and an empty 'zero' (Gussman, 2007). In Polish, *yers* typically surface as [ɛ]. Sometimes the vowel alternation can be seen within a single morphological paradigm, as between [sfɛtɛr] and [sfɛtra] ('sweater', singular nominative and genitive respectively). The second [ɛ] is the *yer*.

When a *yer* does not surface, this can result in particularly complicated consonant clusters, such as in [mgwa] ('fog'), [pʃɛstɛmpstf] ('crimes, genitive'), or the [kt] and [gd] clusters in function words such as [kto] ('who') or [gdɪ] ('if') (Kraska-Szlenk, 2007). Such clusters clearly violate the Sonority Sequencing Principle (Jarosz, 2010). Many of these clusters 'plateau' in sonority, as in the [kt] and [gd] clusters mentioned. Because these clusters result from empty *yer*s, it is an interesting question whether new words with these clusters can be made. More generally, one can ask whether there are phonotactic constraints that govern these clusters.

Additionally, it is worth noting that Polish has geminate consonants; note the difference between [rɔd͡zʲiɲɨ] ('family', gen.) and [rɔd͡zʲinnɨ] ('familial'). While this generally occurs at morpheme boundaries (note that /nɨ/ in the second example is a suffix attached to the root /rɔd͡zʲin/), there are instance where this can occur within a root – and, more interestingly, word-initially. Examples such as [ssak] ('mammal', nom.) retain the length distinction in the initial cluster. Perhaps unsurprisingly, this cluster is the result of a deleted *yer*.

3.3 *Corpora and Extraction* To identify what sound sequences occur in word-onset clusters in Polish, three main corpora were used.

The most sizable corpus used was the Web 1T 5-gram corpus (Brants & Franz, 2009) (henceforth known as Web 1T), taken from a forum of Polish speakers on the internet. All word-initial consonant clusters were extracted from this corpus. However, due to this being an internet forum, a variety of irrelevant material occurred in this output: typos, foreign characters such as hiragana, and non-linguistic material such as hyperlinks. To differentiate between 'legitimate' and 'illegitimate' clusters, the clusters obtained from Web 1T were filtered through (the union of) two additional resources. One was the Polish data from the Universal Morphology project (Sylak-Glassman, 2016) (henceforth known as Unimorph). The other was a list of bigrams counted in a child elicitation study (Jarosz, 2017) (henceforth known as the Jarosz list).[2]. This resulted in a list of 329 unique clusters.

Because Polish orthography maps very closely to its phonology, extraction and transcription of the clusters from Web 1T was done at the same time by use of a finite state automaton built with Pynini (Gorman, 2016). Only words that began with a consonant were accepted and rewritten to their IPA counterpart; all vowels, any character following a vowel, and any invalid character (such as the aforementioned hiragana, but also including numbers) rewrote to the empty string.[3]

It is worth noting that two clusters that existed in the corpora were removed at this point because of the likelihood that speakers would misread the surface form. The written clusters ⟨dż⟩ and ⟨cz⟩ most commonly map to the affricates /d͡ʒ/ and /t͡ʃ/, respectively. However, the clusters generated also included the non-affricated /dʒ/ and /t͡sz/, which would be written identically in the orthography to the affricates. Because the affricates generally occur with a much higher frequency and are therefore more likely to be how these characters would be read by Polish speakers, the two clusters in question (/dʒ/ and /t͡sz/) were removed from consideration.

4 Acceptability survey

A survey was conducted with the 329 clusters described in §3.3. The list of clusters was written in IPA, appended to the nonce stems ⟨arek⟩, ⟨olek⟩, and ⟨ąc⟩. These stems were chosen to avoid conflict with ⟨i⟩ which can signal palatalization, as well as to allow interaction with nasal vowel ⟨ą⟩. This resulted in a list of 987 words. The nonce words were then reverse transcribed back into Polish orthography and presented to the speakers. Four native speakers were asked to give binary acceptability judgments for all nonce words, where '0' denotes unacceptable, and '1' denotes acceptable. Since there four speakers and three words for each cluster, each cluster received 12 ratings. These acceptability scores were then averaged across speakers for each initial cluster. Figure 1 shows the distribution of acceptability scores in the surveys. There are two observations that can be made about the acceptability scores. First, more than half of the clusters received a '1' score on all 12 ratings. Second, despite this, when clusters are ordered from least acceptable to most acceptable, there is a clear gradient cline among clusters that are not perfectly acceptable.

5 Models and Analysis

Given average acceptability scores from the speaker surveys, we examine how well three phonotactic learning models predict the average acceptability of the complex onsets. The first is a simple 'cluster model.' This learning model memorizes attested onset clusters in the training corpus. A word is only scored as

[2] The Jarosz list was written in IPA originally, but using slightly different conventions, namely the distinction between [ʃ] and [ʂ]. Corresponding symbols were adjusted for consistency.

[3] One exception to this was the case of palatalization, which is denoted in Polish orthography with the vowel ⟨i⟩, where the sequence ⟨si⟩ would map to /ɕ/ rather than /si/. In this case, the relevant substring is mapped to the palatalized consonant, rather than the empty string.

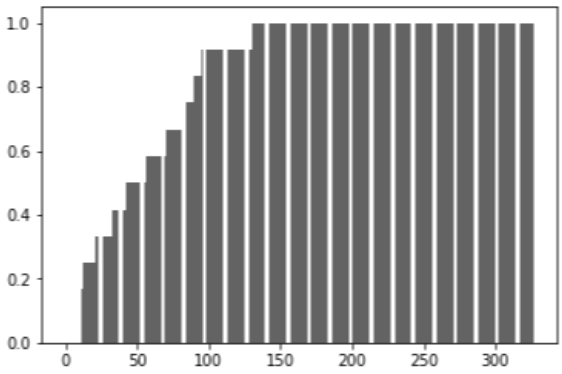

Figure 1: Distribution of results from acceptability survey.

acceptable (score '0') if its onset cluster was in the list of attested clusters. Otherwise it was unacceptable (score '1').

The second is a Strictly 2-Local model. This learner memorizes the attested bigrams in the onset clusters in the training corpus. The acceptability of a word is equal to the number of illicit bigrams in its onset clusters. So again a perfectly acceptable cluster would score '0'. Consequently, all words score '0' by the cluster model will be scored '0' by the bigram-type model as well because if a cluster is attested in the training data all of its internal bigrams are attested in the training data as well. Conversely, there are potentially clusters could be scored '0' in the bigram model because its bigrams appear individually in PWN, even though the full cluster does not (and so the cluster model would score such clusters '1').

The third model was HWPL (Hayes & Wilson, 2008). We used the feature system from Hayes (2009) for the Polish consonants. To facilitate comparison to the bigram model, we limited the width of constraints to 2. We also set HWPL to return 100 constraints.

All models were trained on the same data (see below), and each model then scored each cluster in the list of 329 clusters. Two forms of evaluation were used: Kendall's τ rank and Pearson's r correlation coefficient. Both Pearson's r and Kendall's τ measure the strength of a correlation between two lists of values (such as the speakers' acceptability judgments and a trained models' scores). They differ in exactly how that correlation is calculated, and some have argued that Kendall's τ is more appropriate (Albright, 2009; Gorman, 2013).

5.1 *Training data* For each model, input training data was created using the Słownik Języka Polskiego ('Dictionary of the Polish Language') hosted by Wielka Encyklopedia Powszechna PWN (henceforth known as PWN). PWN is a longstanding Polish institution. Training data constituted initial consonant clusters in PWN that occurred in PWN more than once. The general assumption was that these clusters would approximate the ones in the mental lexicon of Polish speakers. Clusters that occurred exactly once were excluded on the grounds that those words were rare, and that ones that occurred more than once were familiar.

It is worth mentioning that many of the 329 clusters were not attested in PWN. We hypothesize that this is due to the fact that Unimorph is built from crowdsourced entries on Wiktionary (Sylak-Glassman, 2016), the nature of which allows many more infrequent words, unfamiliar to native speakers, to be collected.

The clusters in the training data were padded with boundary symbols. For example, considering the single character ⟨p⟩ as an onset, padding would turn this cluster into ⟨#p#⟩, with # marking the beginning and end of the cluster. So, the bigrams for this 'cluster' would be ⟨#p⟩ and ⟨p#⟩.

Frequencies of each cluster were obtained from the Web 1T corpus. The frequency counted the number of type instances that began with this cluster. For example, for the cluster ⟨gd⟩, there were two words in the Web 1T corpus: 'gdy' ('when, the time at which') and 'Gdańsk' (a city in northern Poland). So the frequency count was 2. We observe that these frequency counts are only relevant for HWPL, and not for the cluster model or bigram model because they only are sensitive to presence versus absence and not frequency.

5.2 *Results for the Cluster Model* The correlation between the cluster model's scores and the average speaker judgments for the 329 clusters, as measured by the Pearson's correlation coefficient, was $r =$

-0.7550197 ($p < 2.2e-16$). Thus, we can see that there is a strong and significant correlation between familiarity with at least two lexical items beginning with a certain onset cluster, and acceptance of other words that have the same onset cluster, regardless of violations of sonority.

Using Kendall's τ coefficient, the correlation is $\tau = -0.7238829$ ($p < 2.2e-16$). Again, the correlation is high enough to justify stating that there is a strong relationship between onset familiarity and acceptability.

5.3 *Results for the Strictly 2-Local Model* The correlation between the Strictly 2-Local model's scores and the average speaker judgments for the 329 clusters, as measured by the Pearson's correlation coefficient, was $r = -0.7263231$ ($p < 2.2e-16$). Kendall's τ was $\tau = -0.6933915$ ($p < 2.2e-16$). Again, it can be determined that familiarity with an individual bigram is a strong predictor of a word's acceptability. It is interesting to note that these correlations are less than the ones obtained for the cluster model. However, it is unclear whether this difference is meaningful.

5.4 *Maximum Entropy* Of the 100 constraints returned by HWPL, 12 were scored with a weight above 6. The two most highly-weighted constraints referred to features within a single phone: *[-Anterior,-Continuant,+Strident] (referring to affricates such as [[t͡ʃ] and [d͡ʒ]) and *[-Anterior,+Strident,-DelayedRelease] (those same affricates, as well as palatalized sibilants). Only one of these top 12 constraints referred to a boundary symbol: *[-Anterior,-Continuant][-word_boundary]. This class largely referred to affricates or palatalized consonants. As noted, a palatal consonant is often written orthographically followed by ⟨i⟩in this position, and elsewhere in the cluster will be diacriticized.

The other highly-ranked constraints, in order of descending weight, were as follows:

- *[–Continuant, +Strident][–Round, –Voice]
- *[–Voice, –DelayedRelease][–Sonorant, +Anterior, +Voice]
- *[–Dorsal, –Voice, +Continuant][–Anterior, +Voice]
- *[+Labial, +Voice, –Nasal, –LabioDental][+Coronal, –Continuant]
- *[–Round, –Voice, +Continuant][–Sonorant, +Voice, –LabioDental]
- *[–Labial, +Continuant, +DelayedRelease][–Round, +Voice, +Continuant, –Lateral, –Trill, –LabioDental]
- *[-Continuant, +Strident][+Voice, –Strident, –Lateral, –Nasal]
- *[+Sonorant, –Round, –Lateral][+Anterior, +Strident]
- *[+Dorsal, +Voice, +Continuant][-Coronal, +Dorsal, +Continuant]

In three of these constraints, the Maximum Entropy learner does note the voicing assimilation within clusters in Polish. Other constraints are more informative if the combination '-Round, -Voice' is simplified to only '-Voice', since [w] was our only round consonant.

The correlation between the HWPL's scores and the average speaker judgments for the 329 clusters, as measured by the Pearson's correlation coefficient, was $r = -0.06955772$ ($p = 0.2083$). Using Kendall's τ, the results were better, with $\tau = -0.14244414$ ($p = 0.004941$). These correlations are not as strong as the ones obtained by the cluster and Strictly 2-Local models.

6 Discussion

The overall conclusion of the results reported above is that both the cluster model and the Strictly 2-Local model outperform HWPL in predicting the results of the acceptability judgments obtained from our survey.

One criticism of these results is that the categorical survey (speakers were given a categorical prompt "Is this word acceptable or not?") tipped the scales in favor of the categorical models. There are two responses to this. The first is that, as shown in Figure 1, the acceptability judgments do in fact show a gradient cline. The second is that probabilistic grammars subsume categorical ones since probabilities close to zero and one simulate categoricity. Hayes and Wilson themselves point this out when they write (p. 382) "it is an inherent

property of maximum entropy models that they can account for both categorical and gradient phonotactics in a natural way."

There are several directions for future research. On the phonotactic learning side, this paper investigated three learning models. However these learning models differed among multiple parameters which makes isolation of relevant factors difficult. For example, how would a probabilistic model that only paid attention to initial onset clusters compare to the categorical cluster model studied here? Similarly, it would be interesting to compare a classic bigram model to the strictly 2-local model studied here. Finally, we would also be interested in categorical model that can generalize constraints over featural representations as the HWPL does. The Bottom Up Factor Inference Algorithm (BUFIA) is a promising candidate in this regard (Chandlee et al., 2019; Rawski, 2021). Figure 2 shows these six possibilities with the learning model in bold the ones that we examined in this paper.

CATEGORICAL	GRADIENT
Cluster	Clusters with type frequency
2-local categorical	Probabilistic bigram model
BUFIA	**HWPL**

Figure 2: Possible categorical vs. gradient models for investigation.

It has been observed that different kinds of phonotactic models are better or worse at making gradient distinctions among attested structures as opposed to unattested ones (Albright, 2009). Since the 329 clusters used in the survey were clusters that were observed in Unimorph and Jarosz's list, this sample likely under-represents unattested (and potentially ill-formed) clusters in Polish. Developing a set of test clusters which addresses this could help clarify whether the HWPL is better at predicting acceptability judgments of unattested clusters as opposed to (mostly) attested ones.

Relatedly, the acceptability judgments themselves could be collected with different tasks. A Likert scale could be used. Also, the forms could be presented auditorily instead of orthographically. It would be interesting to see whether such changes in the tasks impact the results reported here.

Finally, a further more general question is how to reconcile the place of phonological unacceptable clusters that nonetheless exist in the lexicon (Algeo, 1978). Consider the following triples:

1. [fiɹ] 'fear', [siɹ] 'seer', [sfiɹ] 'sphere'
2. [fit] 'feet', [sit] 'seat', *[sfit] 'sfeet (hypothetical)'

Despite the cluster [sf] being an existing initial cluster in the lexicon of English, [sf] can only precede [-iɹ], not [-it]. This appears to be different from Polish, where clusters that exist in the lexicon, such as [ʐd̪ʐbw], are generally considered to be acceptable.

English speakers may have gradient reactions to these clusters, ranking some as 'more' English than others (Algeo, 1978). However, these reactions can still be attributed to the effects of categorical rules. Algeo (1978) notes specifically that [sf] and [sv] belong to this category of existing but illicit clusters, but also that [sv] violates the additional rule of obstruent voicing agreement. The voicing disagreement may very well make [sv] less acceptable across the sum of speaker judgments.

To our knowledge, there is no widely acceptable term for these occurring but ill-formed words, sometimes called 'structurally excluded' (Gorman, 2013). These phones can be held over from loan words or sounds that simply are no longer productive in English. Other sequences may have once occurred in the lexicon, such as [lg], but left it for one reason or another, such as vocalization (Algeo, 1978). In a sense, such sequences can be considered exceptional, and how they are treated in the lexicon and by the phonological is an outstanding question for phonological theory.

7 Conclusions

Gorman (2013) and Durvasula (2020) have questioned whether probabilistic phonotactic learning models, such as the one presented in Hayes & Wilson (2008), which return probabilistic grammars, are any better than categorical grammars for predicting gradient acceptability judgments. We have examined this

question by examining how well categorical and probabilistic phonotactic learning models extract grammars which predict Polish speakers' acceptability judgments on words with varied initial consonant clusters. Polish is an especially interesting language to look at because of its rich inventory of sonority-sequencing defying consonant clusters, often as a result of *yer*-deletion. We found that the categorical baselines considered here generally outperformed the Hayes and Wilson's maximum-entropy based phonotactic learner (Hayes & Wilson, 2008). We conclude that gradient acceptability judgments do not provide unambiguous evidence for gradient, probabilistic grammars.

8 Appendix: Materials

All materials are available on GitHub at https://github.com/kkostyszyn/PolishClustersData. This includes the acceptability scores from the native speaker survey, the clusters as noted for PWN, and the input files for HWPL.

References

Albright, Adam (2007). Natural classes are not enough: Biased generalization in novel onset clusters. *15th Manchester Phonology Meeting*, Manchester, UK.

Albright, Adam (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology* 26:1, 9–41.

Algeo, John (1978). What consonant clusters are possible? *Word* 29:3, 206–244.

Armstrong, Sharon L., Lila R. Gleitman & Henry Gleitman (1983). What some concepts might not be. *Cognition* 13, 263–308.

Bárkányi, Zsyzsanna (2011). Gradient phonotactic acceptability: a case study from Slovak. *Acta Linguistica Hungarica* 58(4), 353–391.

Brants, Thorsten & Alex Franz (2009). Web 1T 5-gram, 10 European Languages Version 1 lDC2009T25.

Buckley, Eugene (2001). Polish o-raising and phonological explanation.

Chandlee, Jane, Remi Eyraud, Jeffrey Heinz, Adam Jardine & Jonathan Rawski (2019). Learning with partially ordered representations. *Proceedings of the 16th Meeting on the Mathematics of Language*, Association for Computational Linguistics, Toronto, Canada, 91–101.

Durvasula, Karthik (2020). Oh gradience, whence do you come? Keynote presentation at the Annual Meeting of Phonology.

Gorman, Kyle (2013). *Generative Phonotactics*. Ph.D. thesis, University of Pennsylvania.

Gorman, Kyle (2016). Pynini: a Python library for weighted finite-state grammar compilation. *Proceedings of the ACL Workshop on Statistical NLP and Weighted Automata*, 75–80.

Gussman, Edmund (2007). *The Phonology of Polish (The Phonology of the World's Languages*. Oxford University Press.

Hayes, Bruce (2009). *Introductory Phonology*. Wiley-Blackwell, Malden, MA.

Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39, 379–440.

Heinz, Jeffrey (2010). Learning long-distance phonotactics. *Linguistic Inquiry* 41:4, 623–661.

Heinz, Jeffrey & William Idsardi (2017). Computational phonology today. *Phonology* 34:2, 211–219.

Heinz, Jeffrey, Chetan Rawal & Herbert G. Tanner (2011). Tier-based strictly local constraints for phonology. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Portland, Oregon, USA, 58–64.

Jarosz, Gaja (2010). Implicational markedness and frequency in constraint-based computational models of phonological learning. *Journal of Child Language* 37(3), 565–606.

Jarosz, Gaja (2017). Defying the stimulus: Acquisition of complex onsets in Polish. *Phonology* 34(2), 269–298.

Kraska-Szlenk, Iwona (2007). *Analogy: the Relation between Lexicon and Grammar*. Lincom Europa.

Lambert, Dakotah (2021). Grammar interpretations and learning tsl online. Chandlee, Jane, Rémi Eyraud, Jeff Heinz, Adam Jardine & Menno van Zaanen (eds.), *Proceedings of the Fifteenth International Conference on Grammatical Inference*, PMLR, vol. 153 of *Proceedings of Machine Learning Research*, 81–91.

Rawski, Jonathan (2021). *Structure and Learning in Natural Language*. Ph.D. thesis, Stony Brook University.

Rogers, James & Geoffrey Pullum (2011). Aural pattern recognition experiments and the subregular hierarchy. *Journal of Logic, Language and Information* 20, 329–342.

Rogers, James, Jeffrey Heinz, Gil Bailey, Matt Edlefsen, Molly Visscher, David Wellcome & Sean Wibel (2010). On languages piecewise testable in the strict sense. Ebert, Christian, Gerhard Jäger & Jens Michaelis (eds.), *The Mathematics of Language*, Springer, vol. 6149 of *Lecture Notes in Artifical Intelligence*, 255–265.

Scholes, Robert J. (1966). *Phonotactic grammaticality*. Mouton, Germany.

Sylak-Glassman, John (2016). The composition and use of the universal morphological feature schema (Unimorph Schema). Tech. rep., Johns Hopkins University, https://unimorph.github.io/.

Wilson, Colin & Gillian Gallagher (2018). Accidental gaps and surface-based phonotactic learning: A case study of South Bolivian Quechua. *Linguistic Inquiry* 49, 610–623.

Zydorowicz, Paulina & Paula Orzechowska (2017). The study of Polish phonotactics: Measures of phonotactic preferability. *Studies in Polish Linguistics* 12, 97–121.