Determinants of phonotactic acceptability: sonority or lexical statistics?

Fergus O'Dowd
12232815
University of Amsterdam

MA Thesis, General Linguistics
Supervisors: Paul Boersma & Silke Hamann

25th August 2019

Abstract

Conceptions of phonotactics differ as to whether phonotactic knowledge is based on statistical generalisation across the lexicon (the 'lexicalist' view), or whether it instead involves prior analytic biases (the 'universalist' view). The conclusions of previous research have not converged on either a lexicalist or universalist explanation of 'sonority projection' effects, in which novel sequences which conform to the Sonority Sequencing Principle are judged more acceptable than those that do not. To empirically test these two alternative views of sonority projection, a predictive difference between universalist and lexicalist hypotheses was formulated and then tested experimentally on speakers of English, in a reading and a listening task. The results of this experiment were run through a linear mixed-effects model. The outcome of this model gave effects that did not differ significantly from the null hypothesis. This neither proved nor disproved either the lexicalist or universalist hypotheses. Nor did any individual participant behave exactly in line with the predictions of either hypothesis.

However, a fairly robust preference was found for /kn/ over /fn/. This may have been due to orthographic effects which persisted in both the listening and reading tasks, suggesting a link between orthography and phonology. Reasons for the overall null result (including differing conceptions of sonority and lexical statistics) are discussed, and ways to mitigate possible flaws in the experimental paradigm are then proposed.

Acknowledgements

The utmost thanks are owed to my supervisors, Silke Hamann and Paul Boersma, both for their teaching and for their help throughout the process of producing this thesis. I am extremely grateful to Dirk Jan Vet for helping with the recording of stimuli and providing scripts with which these stimuli could be easily integrated into an experiment. The continued support of my family and friends, both in the UK and in the Netherlands, was also invaluable.

Contents

1 Introduction

Nonwords, as syntactic and semantic 'blank slates', can serve as good test cases for putative phonological processes. One such case is phonotactics; speakers' judgements on nonwords can reveal which phonotactic patterns are internalised and which are not learnt.

Conceptions of phonotactics differ as to whether phonotactic knowledge is simply statistical generalisation across the lexicon, or whether it instead involves a number of prior 'analytic biases' (Moreton, 2008). One such postulated analytic bias is the Sonority Sequencing Principle ('SSP'; Clements, 1990: 285). This states that "between any member of a syllable and the syllable peak, only sounds of higher sonority rank are permitted". Though the exact nature and order of the sonority cline is hotly debated (Clements, 1990; Ohala, 1992; Henke et al., 2012), it can be roughly outlined as follows (following Prince and Smolensky, 2004; Berent et al., 2007):

(1) *Least sonorous* *Most sonorous*
stops < fricatives < nasals < liquids < glides < vowels

In practice, the SSP means that onsets may only progress rightwards along the cline while codas may only progress leftwards.

Whether speakers have internalised something akin to the SSP is far from agreed upon. Experiments on 'sonority projection', the effect by which SSP-like generalisations are extended to clusters unattested in the target language, have failed to converge on one point of view. Berent et al. (2007, 2009) and Albright (2007) claim that an effect of sonority persists in English even after controlling for lexical statistics via a number of models. Jarosz and Rysling (2017) find the same for Polish, a language in which there are a greater number of SSP-violating clusters. Ren et al. (2010) show an SSP effect in Mandarin Chinese despite little in the way of lexical statistics in Mandarin Chinese to confirm such a case. However, Hayes (2011) suggests that a rudimentary feature system which distinguishes the major classes shown above can account for SSP effects without the need for an independent SSP, even in languages without onset clusters like Mandarin Chinese. Similarly, Daland et al. (2011) suggest SSP effects can be reduced to statistics if syllabification is also specified. Meanwhile, the status of the SSP as a coherent principle has been frequently called into question (Ohala, 1992; Wright, 2004; Henke et al., 2012 and others), casting doubt on whether it is even possible for speakers to internalise the SSP.

Despite this debate over the psychological reality of the SSP, no authors have explicitly tested whether speakers can make judgements that go against the SSP. This is despite the fact that some proponents of the SSP suggest that such judgements are impossible. Berent et al. (2009: 77) argue that "the learner must end up formulating just those generalisations that coincide with sonority-sequencing principles and not others that contradict those principles". Meanwhile, Jarosz and Rysling (2017: 11) see some phonotactic models as flawed because "nothing prevents such models from inducing

constraints with contradictory effects to the SSP, given appropriate evidence". Such a viewpoint – that sonority is inviolable – is called 'universalist' by Daland et al. (2011); they contrast it with a 'lexicalist' viewpoint, which posits that sonority is generalised from the lexicon. The present study aims to formulate a test case for a predictive difference between universalist and lexicalist hypotheses, and then test this case on a number of native speakers to see which view fits better with the data. If speakers' generalisations align with the SSP despite statistical evidence to the contrary, this provides evidence that these speakers have internalised something akin to the SSP.

Though a number of phonotactic studies mentioned above claim to have found an effect of sonority on acceptability judgements independent of lexical statistics, there are some potential flaws in these conclusions (§2.2.1). Chief among these is their inadequate compensation for simple statistical factors. The definition of sonority is also rather confused (§2.2.3), and it is unclear whether many models of grammar that allow for a universal SSP even have a mechanism for learners to learn lexical generalisations (§2.2.2). Furthermore, there is some evidence from modelling studies (Hayes, 2011; Daland et al., 2011) that sonority projection can be achieved without recourse to a universalist SSP. In a case where lexical statistics and the SSP make different predictions, my expectation is therefore that speakers' preferences will align with the prediction of lexical statistics, in accordance with the lexicalist viewpoint of Daland and colleagues.

2 Conceptions of phonotactics

Phonotactics, as defined by Algeo (1978: 206), is "the study of the positions occupied by phonological units relative to one another". A definition this broad seems necessary in order to accommodate the radically different approaches to conceptualising phonotactics, of which the more popular approaches are discussed below.

Before discussing conceptions of phonotactics, it is worth clarifying one terminological issue, following Albright (2009: 9). Describing a string as 'grammatical' (e.g. Scholes, 1966) implies some kind of grammatical formalisation of phonotactics. Describing a string as 'wordlike' (e.g. Bailey and Hahn, 2001) implies analogy. Describing a string as 'probable' or 'likely' implies statistical generalisation[1]. Albright instead uses 'acceptable' in an attempt to be more theory-neutral; I follow this henceforth.

2.1 Phonotactics in generative grammar

One of the first formalisms of phonotactics was the 'morpheme structure constraint' (e.g. Halle, 1959: 56), an early generative theoretical mechanism for encoding phonotactic generalisations. Under Halle's formalism, there is a limited set of constraints which make broad generalisations demarcating what is possible and what is impossible as a word of a given language. However, a large body of research has

[1] Daland et al. (2011) make a distinction between 'likely' and 'probable', but this is mathematical rather than linguistic.

shown, both experimentally (e.g. Greenberg and Jenkins, 1964; Frisch et al., 2004; Albright and Hayes, 2003) and logically (Algeo, 1978; Coetzee, 2008), that the acceptability of a word is gradient. The existence of degrees of acceptability has long been recognised (Chomsky and Halle, 1968: 416) as a problem for strictly categorical morpheme structure constraints.

Chomsky and Halle (1968: 417) keep the general architecture of morpheme structure constraints but add a fix for gradience, whereby the more features that differ between a nonword and an extant lexical item, the worse the word is judged. That is to say, the more complex a rule that differentiates an item from its closest lexical neighbour, the more significant its violation. The nonword /bɹɛk/ is judged better than /bzɪk/ as /bɹɛk/ only differs in [±high] from /bɹɪk/, while /bzɪk/ differs in [±vocalic], [±strident] and [±anterior]. However, this view of phonotactics as the licit-ness of the worst part of the word lacks empirical support; language users instead judge acceptability based on the whole word's phonology (Ohala and Ohala, 1986; Coleman and Pierrehumbert, 1997).

Most subsequent work in phonotactics abandons any conception that phonotactics is part of a structure-building generative grammar. A recent exception, however, is Futrell et al. (2017), who uniquely argue for a model that builds acceptable words generatively, based on an algorithm trained on the lexicon of English. Their model assigns high (but not necessarily accurate) acceptability scores to real words with many productive morphemes (e.g. *mistrustful*). However, there is one omission in Futrell and colleagues' empirical support for this model: it is not tested on nonwords. Though Futrell et al. (p. 73) argue that "phonotactic restrictions mean that each language uses only a subset of the logically, or even articulatorily, possible strings of phonemes", when presented with new strings, language users are happy to differentiate between them. It therefore remains to be seen whether their model can account for sonority projection.

It is worth noting that generative models of phonotactics have no theoretical mechanism which explicitly derives sonority projection effects. A metric which distinguishes a nonword from the closest extant lexical item along the lines of Chomsky and Halle (1968, as discussed above) would have the dual effect of under-penalising SSP-violating clusters that are close to extant lexical items (e.g. /zkɪll/, which differs from *skill* in only the voicing of the first segment) and over-penalising SSP-conforming clusters that have no nearby lexical neighbours (e.g. /pwɒst/, which has no nearby lexical items: */wɒst/, */pɒst/, */plɒst/ etc. are not words of English). No other studies in generative phonotactics have specifically examined sonority projection.

2.2 Phonotactics as markedness constraints

In contrast, Optimality Theory (OT) sees phonotactic generalisations as arising from the interaction between faithfulness constraints, which specify that some aspect of the output must be the same as the underlying form, and markedness constraints, which disprefer certain structures. In this way, phonotactic generalisations are seen as the result of the structure of the phonological grammar rather than as a separate

phonotactic system (Prince and Smolensky, 2004: 223). Thus the phonotactic generalisation in German and Dutch that all word-final obstruents must be voiceless is encoded formally as a constraint interaction: the markedness constraint which prohibits final voiced obstruents outranks the faithfulness constraint for voicing.

2.2.1 Sonority as a universal analytic bias

Following Prince and Smolensky (2004), Berent et al. (2007) see the SSP as a universal analytic bias caused by IS THIS REALLY WHAT I MEAN? the interaction of universal markedness constraints; this is a hallmark of the universalist approach to sonority outlined above. In their conception of sonority, it has three concrete manifestations in speakers' acceptability judgements:

a) a preference for sonority rises over sonority falls
b) a preference for greater sonority rises over smaller sonority rises
c) a preference for smaller sonority falls over greater sonority falls

Berent and colleagues examined the prevalence of misperception in unattested onset clusters (including sonorant-stop and stop-sonorant clusters) as a proxy for phonological markedness. Their results showed a greater presence of consonant epenthesis in sonority falls than in rises, suggesting listeners find sonority rises more acceptable. However, there are numerous flaws with the stimuli and experiment design. These flaws are both phonetic – with sonorants being acoustically closer to vowels and therefore less phonetically distinct from sonorant-vowel sequences (Peperkamp, 2007) – and phonologically – with their results quite easily explicable with a total absence of sonorant-stop clusters in the English lexicon.

Berent et al. (2009) attempt to rectify these flaws by testing nasal-initial clusters, which are unattested in English. Specifically, they predict that sonority falls /md/ and /nb/ will be less acceptable than sonority rises /ml/ and /nw/, which is indeed what they find. Berent and colleagues claim that this is inexplicable with lexical statistics, partly by claiming that there is no statistically significant effect of word-level position-sensitive phoneme logarithmic frequency (i.e. log frequency of each phoneme in either first or second position in a cluster). However, this *cannot* be true; the position-specific frequency of /d/ and /b/ as the second consonant in a cluster is zero, meaning that their log frequency, log(0), is undefined. Berent and colleagues' premise therefore rests upon interpreting a mathematical impossibility[2]. Perhaps their conclusions should be taken with more than a pinch of salt.

Albright (2007) also finds an SSP effect independent of natural class-based lexical statistics in unattested clusters. Albright formalises this as an analytic bias towards sonority rises, but notes that his formalism of this analytic bias is "hopelessly hand-crafted" (p. 24). Albright's original conclusion, similarly to that of Berent et al. (2009), views phonotactics as being based on a combination of statistical lexical information

[2] Another possibility is that Berent and colleagues calculated log-frequency based on position as second segment in the word, i.e. taking into account words like *about* where /b/ occurs in the onset of the second syllable, or words like *absent* where /b/ is in the coda of the first syllable. This is an equally flawed methodology as it violates structure-dependence.

and analytic bias. This bias is formalised with OT markedness constraints in Berent and colleagues' case; Albright does not elaborate a precise formalism. This contrasts with the traditional OT view, in which phonotactic patterns are caused by constraint interaction alone[3].

An analytic bias for sonority has been posited for a number of languages apart from English. Jarosz and Rysling (2017) find an independent SSP effect for Polish, a language in which there are a greater variety of sonority falls than in English. However, they note that all sonority falls they tested in Polish are very infrequent compared to their sonority rises and plateaus, yet they draw a trendline between the acceptability of attested sonority plateaus/rises (their examples of which come from *all* levels of cluster frequency in Polish) and the acceptability of attested sonority falls (their examples of which come only from *low*-frequency clusters). This represents a failure to account for lexical statistics for the authors' attested clusters. Also casting doubt upon their conclusions is the fact that Jarosz and Rysling do not control for position-specific phoneme frequency, e.g. that /j/ never occurs initially in an onset cluster in Polish[4].

A more robust argument for sonority as an analytic bias is given by Ren et al. (2010), who show sonority projection for Mandarin Chinese speakers there being despite little in that language's lexicon to support the SSP. Hayes (2011) claims to prove that sonority projection can be achieved without the SSP, as long as the phonology of the language in question has a rudimentary feature system. Given features that allow users to induce a sonority-like cline, Hayes argues that sonority projection falls out naturally. More interestingly, Hayes shows that these effects even occur in a language without onset clusters; all that is needed is a feature system in which more 'sonorous' consonants share more features with vowels.

2.2.2 The source of constraints: the input or the lexicon?

There is one notable theoretical problem with phonotactics as the ranking of general constraints. The OT grammar has no way of directly encoding statistical generalisations over the lexicon; these must all be formalised into discrete constraints before they can be active in the phonology. This requires the grammar to have some mechanism by which the lexicon can influence constraint rankings. However, most OT learning algorithms (e.g. Boersma, 1997; Tesar, 1998) see constraint rankings as learnt by generalisations over the *input* rather than the *lexicon*. This implies that a word or pattern's effect on phonotactic judgements should positively correlate with the word or pattern's token frequency: the more frequent it is in the input, the more effect it has on constraint rankings, and the more effect it therefore has on phonotactic judgements. Yet this relationship is *not* what is found; type frequency of a given phonotactic pattern has repeatedly been shown to correlate better with phonotactic acceptability than token frequency (Hay et al., 2004; Hayes and Wilson, 2008; Albright, 2009), suggesting generalisation from the lexicon rather than from the input.

[3] See §2.2.2 for further discussion of the representation of phonotactics under Optimality Theory.
[4] Note how the failure to adequately account for position-specific phoneme frequency was also a critical flaw of Berent et al. (2009).

There is also no clear positive correlation at the word level between token frequency and effects on phonotactic judgements. Both very infrequent and very frequent words have smaller effects on phonotactic judgements, while words of middling frequency have the greatest effects (Bailey and Hahn, 2001: 578). This relationship requires language users to independently access token frequency by lexical item. This need for language users to access lexical items to formulate constraints contradicts theories in which constraints are formulated from the input. To achieve adequacy in explaining phonotactics, OT learning algorithms therefore need a mechanism to access the lexicon directly, rather than only accessing the input. I am aware of no OT theorists who see constraints as learnt from the lexicon.

While it is technically possible to have morpheme structure constraints which learn from the lexicon, this violates the concept of Richness of the Base (Smolensky, 1996), which holds that there is no restriction on the input into the phonological derivation, i.e. the underlying form stored in the lexicon. As such, prominent OT theorists (e.g. McCarthy, 1998) reason that morpheme structure constraints do not exist[5].

2.2.3 Flaws in the notion of sonority

Despite frequent reference to sonority in a number of works outlined above, the coherence of 'sonority' as a concept has frequently been criticised as circular, variable and subject to systematic exceptions (Ohala and Kawasaki-Fukumori, 1997; Wright, 2004; Henke et al., 2012).

Ohala and Kawasaki-Fukumori (1997) see the main problem with sonority as its circularity. Sonority is defined as restrictions on what can occur in syllable margins, but syllable margins are also defined by sonority. Ohala and Kawasaki-Fukumori give the example of the medial cluster in *scoundrel*. The syllable boundary is usually placed as in [scoun][drel], with the reason being that /n/ is more sonorous than /d/, so must occur in a separate syllable. But the reason for /n/ being more sonorous is – at least partly – that nasals like /n/ do not occur before stops like /d/ in syllable onsets. This logic is entirely circular; the definition of and the motivation for sonority are the same.

The precise scale and order of the sonority hierarchy is also unclear. The exact reasoning for the various proposed orders is beyond the scope of this paper, but the various proposals will be mentioned below and discussed later (§5.3) when they interact with the experiment design. Minimally, the sonority hierarchy consists of the following (Zec, 1995: 87):

(2) vowels > sonorants > obstruents

Clements (1990) construes the SSP as:

(3) vowels > approximants > nasals > obstruents

[5] However, Booij (1999) argues for the necessity of constraints which act over the lexicon.

Levin (1985: 63), formalising Steriade (1982), sees it as:

(4) vowels > approximants > nasals > fricatives > stops

Prince and Smolensky (2004: 12) expand this scale further:

(5) low vowels > high vowels > liquids > nasals > voiced fricatives > voiceless fricatives > voiced stops > voiceless stops

Basbøll (2005: 197), in attempting to formulate a sonority generalisation that is universally unviolated, takes a radically different approach:

(6) vowels > voiced consonants > unaspirated consonants > aspirated consonants

Zec (1995: 88) also notes that there may be a need for a distinction between /l/ and /r/ (the precise realisations of which she does not elaborate on). Such a multiplicity of definitions of the sonority hierarchy (and therefore the SSP) makes it harder to empirically verify the effect of sonority, and easier to cite whichever model comes closest to fitting the data. An effect of sonority in a study that uses the sonority hierarchy in (5) does not imply an effect of sonority in the hierarchy in (2); it could be the case that an effect of sonority based on (5) is due to the ordering 'liquids > nasals', which is not present in (2). Similarly, an effect of sonority based on the hierarchy in (2) does not imply an effect of sonority in the hierarchy in (5); an effect confirming the ordering 'sonorants > obstruents' in (2) does not preclude an effect that violates the ordering 'nasals > liquids' in (5).

The SSP also appears violable, and, more damagingly, there are typological tendencies for these violations to be of certain types, suggesting a systematicity of violation. Wright (2004) elaborates on some of these counterexamples, including the cross-linguistically common prevalence of nasal-stop and sibilant-stop clusters. In arguing that the SSP can be based on phonetic experience of cue salience, he cites the relative cue reliability of sibilants and nasals as motivation for their disobedience of the SSP. Wright therefore unifies the formal SSP and its exceptions with one functional motivation, rather than introducing additional theoretical mechanisms to explain away counterexamples.

Incorporating all these criticisms into a coherent definition of sonority proves challenging. In a monograph on sonority, Parker (2011: 1160) describes it as "a unique type of relative, n-ary feature-like phonological element that potentially categorises all speech sounds into a hierarchical scale". This definition is perhaps so vague as to be meaningless.

Instead, Wright (2004) and Henke et al. (2012) see the SSP not as universally-endowed grammar, but induction from phonetic experience; both argue that the SSP is due to cue robustness. Wright implies that this knowledge of cue robustness can feed into a psychologically real constraint, under the theoretical apparatus of Phonetically Based Phonology (Hayes and Steriade, same volume). Henke and colleagues, meanwhile, see

the SSP as an epiphenomenon of perceptually-motivated sound change, arguing that lexical statistics can explain both the SSP *and* crosslinguistic variation in its formulation and exceptions.

Berent et al. (2007) acknowledge these criticisms of sonority, but maintain that it has psychological reality. They claim (p. 625) that "the possibility that the sonority markedness hierarchy might be induced from phonetic experience is perfectly compatible with the existence of innate constraints on the organization of the grammar". But Henke et al. (2012: 67) explicitly "dispute… whether the SSP is a universal principle of synchronic grammars". Indeed, given different input in different languages, it is only logical for sonority effects to vary cross-linguistically just as 'phonetic experience' varies. But cross-linguistic variation in the SSP (cf. Steriade, 1982) raises the problem of circularity; if we can define the SSP differently between languages, then its definition is circular and its predictive power is weakened.

There seems to be an element of cognitive dissonance in many researchers' work on sonority, in which said researchers make claims about the influence of sonority on linguistic processes and patterns, but then have significant trouble defining sonority in any logically consistent way.

2.3 Phonotactics as generalisations across the lexicon

Others do away with the idea of innate constraints on phonological patterns, like the SSP. In the view of Frisch et al. (2000: 494), "phonotactic knowledge is best viewed as an emergent property of the encoding and processing of lexical information". Speakers learn such lexical generalisations both probabilistically and analogically (Bailey and Hahn, 2001). The usual method postulated for finding such generalisations is the use of lexical statistics: statistical facts, patterns and trends about the words in the lexicon. The immediate conceptual motivation for this is clear: the gradient nature of phonotactic acceptability fits well with the gradient nature of statistical patterns. And indeed, both experimental (Frisch et al., 2000; Bailey and Hahn, 2001; Hay et al., 2004) and modelling (Bailey and Hahn, 2001; Hayes and Wilson, 2008; Albright, 2009; Daland et al., 2011) studies have shown a significant correlation between lexical statistics and phonotactic acceptability.

Those who argue for lexical generalisations as the source of phonotactic acceptability may[6] accept the possibility that such generalisations can be used to build more abstract grammatical constraints. Yet crucially, such a viewpoint entails that learners may come to abstract only those generalisations which are supported in the lexicon; in other words, phonotactic constraints may be "abstract, but not too abstract" (Frisch and Zawaydeh, 2001: 104-5). This is in contrast to hypotheses that see phonotactics as partly determined by innate factors (e.g. Berent et al., 2007). It also contrasts with hypotheses that derive phonotactics from an interaction between markedness constraints which are neither phonotactics-specific nor lexically-derived.

[6] Though also may *not*; Daland et al. (2011), among others, do not follow such a line of argument.

In evaluating lexical statistics' effect on phonotactic acceptability, there are two commonly-used (Albright, 2009: 10) measures: phonotactic probability and neighbourhood density. Phonotactic probability measures evaluate the transitional probability between combinations of phonemes or features. There are multiple ways of calculating phonotactic probability; it can be purely linear and segmental (e.g. Vitevitch and Luce, 2004), or take into account the similarity between phonemes (e.g. by encoding features; Albright, 2009), or include syllabic and metrical structure (Coleman and Pierrehumbert, 1997; Bailey and Hahn, 2001; Daland et al., 2011).

Neighbourhood density measures the number of nearby attested words; it thus measures the propensity for analogy at the lexical level. It is also possible to analogise from levels below the word. Davidson (2006) suggests analogies may be made on the featural level, while measures of feature- and natural class-based similarity (Frisch, 1996; Frisch et al., 2004; Albright, 2009) are also a form of feature-based analogy.

2.3.1 Statistical probability

The transitional probability of combinations of segments, features and syllabic constituents in real words has been repeatedly shown to be a strong predictor of phonotactic acceptability in nonwords. Jusczyk et al. (1994) were the first to empirically find an effect of lexical transitional probability on phonotactics. In this case, a nonword's average biphone probability – the transitional probability between two segments – correlated with listening preference in infants (a proxy for phonotactic acceptability). The authors also found that (token) frequency of phonemes in a given (linear) position in words in the lexicon predicted acceptability. Vitevitch et al. (1997) replicated this finding for adults. However, Hayes (2012, citing McCarthy and Prince, 1996: 1) notes that Vitevitch and Luce's model engages in limitless segment-counting, thereby challenging commonly-held assumptions about possible phonological processes[7]. Vitevitch and Luce's model also counts segments in linear order with no reference to their syllabic or prosodic structure, violating the linguistic principle of structure-dependence (Crain and Nakayama, 1987).

Hay et al. (2004), examining specific medial nasal-obstruent clusters in nonwords (e.g. /nt/ in /klɛntɪk/), found a correlation between their frequency in attested words and their acceptability in nonwords. When adding an effect of morphonological parsing (e.g. the fact that /klɛntɪk/ could be parsed with an attested morpheme /tɪk/), this correlation becomes particularly strong.

Albright (2009) creates a model of phonotactics as biphone probability, which he then shows to be quite predictive of phonotactic acceptability judgements. To achieve this, Albright gives the model an ability to analogise between segments in the same natural class, adding a level of linguistic knowledge to raw statistical calculation. Coleman and Pierrehumbert (1997) take a similar approach, creating a model of transitional probability based on (hierarchical) syllabic constituents rather than segments, avoiding the problems with raw biphone probability mentioned above. This penalises unlikely combinations more *within* onsets (or rimes) than *between* onsets and rimes. Crucially,

[7] For further detail on these problems, see Hayes (2012) and McCarthy and Prince (1996).

they also prove that such a statistical model does significantly better than a model which penalises nonwords based on the acceptability of their least acceptable part, which is what traditional generative or violation grammars (e.g. Optimality Theory) would predict[8].

As well as making low-level statistical generalisations over the lexicon, learners may be able to use lexical statistics to build abstract linguistic constraints. These abstract constraints may no longer perfectly correlate with raw statistics (Frisch and Zawaydeh, 2001; Coetzee, 2008; Hayes and Wilson, 2008), but they will crucially not contradict the statistical tendencies from which they were built. In other words, statistical tendencies can be warped by abstraction into higher-order constraints. Frisch et al. (2004) claim that speakers' knowledge of the ratio of observed frequency to expected frequency[9] allows them to build abstract constraints such as the Obligatory Contour Principle, a general cross-linguistic restriction on the co-occurrence of two similar segments. Crucially, this allows such constraints to vary between languages, based on the statistical tendencies of a given language's lexicon (Frisch et al., 2004: 182). This variation is in the constraints' "degree of gradience", ranging from linearly gradient to categorical. Indeed, Henke et al. (2012) argue that cross-linguistic variation in sonority can be explained by cross-linguistic variation in the lexical evidence for sonority.

Daland et al. (2011) argue that lexical statistics alone can account for sonority, given a model which incorporates syllabification and generalises over features. They compare a number of prior models, including those of Coleman and Pierrehumbert (1997), Bailey and Hahn (2001), Vitevitch and Luce (2004), Hayes and Wilson (2008), and Albright (2009). The feature-based models tend to do best, and their accuracy is enhanced when combined with a syllabification mechanism. However, there is a limit to which Daland and colleagues' conclusions are proof that sonority need not be innate; the models they test are approximations, not empirical proofs, of how speakers judge phonotactic acceptability.

Most researchers who posit some kind of sonority hierarchy do accept that lexical statistics have at least some effect on phonotactic acceptability. Jarosz and Rysling (2017) elaborate on this, arguing that language learners have an initial state including innate primitives like sonority that is then subjected to, and "warped" (p. 11) by, experience. However, their hypothesised SSP is nevertheless persistent; it cannot be overridden, and indeed their results show a persistent SSP bias (though note the criticisms of their methodology in §2.2.1). This is of course contrary to an approach in which sonority projection is *purely* driven by lexical statistics.

A more troubling critique of an approach to phonotactics based purely on lexical statistics comes from Becker et al. (2011). Becker and colleagues find that some statistically significant phonotactic correlations are undergeneralised. Turkish speakers, when tested on novel items, fail to generalise the correlation between vowel

[8] Coetzee (2008), however, argues for an Optimality Theory violation grammar in which cumulative violations create gradient acceptability. Such a perspective is also found in Linear Optimality Theory (Keller, 2000) and Harmonic Grammar (Legendre et al., 1990).
[9] Where expected frequency is calculated using position-specific phoneme frequency

backness and voicing alternation of the following consonant. Becker and colleagues argue that this is an effect of a 'surfeit of the stimulus': the idea that there are too many possible statistical generalisations for speakers to compute. Their solution is to add universal analytic biases to the grammar which constrain which statistical generalisations can be made; in their case, this is some kind of restriction on consonant-vowel dependencies (cf. Moreton, 2008). Though the authors do not touch on this, sonority could conceivably be another such analytic bias.

Perhaps, however, something else causes speakers to under-generalise this relationship. Though vowel backness statistically significantly *correlates* with consonant voicing alternation, it is not highly *predictive* of voicing alternation; front vowels are roughly evenly split between alternating and non-alternating consonants, while back vowels are slightly biased to precede alternating consonants. It is possible that Turkish speakers fail to generalise this pattern because an even split is not a useful predictor, even though the variable's overall correlation may be significant. Perhaps it is worth disentangling *predictiveness* from sheer correlation.

Hayes and Wilson's (2008) maximum entropy model does just this; it accounts for phonotactic acceptability based on the predictiveness of generalisations across the lexicon, which are then formalised into constraints. The model evaluates predictiveness by valuing constraints that combine high lexical regularity (i.e. lack of violations in attested words) with high lexical generality (i.e. ability to account for large numbers of words). This model, with no innate biases beyond a standard feature system, does well in accounting for sonority effects in acceptability judgements. However, in a follow-up study, Hayes and White (2013) suggest that some of the model's individual constraints are under-learnt by real speakers, while others are robustly used – and that the robust constraints include constraints that encode sonority. Hayes and White therefore suggest that a set of analytic biases (assumedly including sonority) limit which constraints can be learnt. However, they acknowledge that the under-learning effect could instead be due to the nature of their under-learnt constraints, which tend to be both more formally complex and consonant-vowel dependencies, which Moreton (2008) shows are harder to learn. Hayes and White thus fail to adequately prove that sonority is an analytic bias.

2.3.2 Analogical generalisations and neighbourhood density

The ability of attested words and clusters to analogically affect phonotactic acceptability was recognised by Greenberg and Jenkins (1964), who asked participants presented with nonwords to rate their acceptability and to give word associations. More acceptable nonwords prompted more word associations, suggesting that a nonword's number of possible real-word analogies is correlated with its acceptability. This effect of 'neighbourhood density' (following Luce, 1986) has been repeatedly shown (e.g. Bailey and Hahn, 2001; Frisch and Zawaydeh, 2001) to correlate with nonwords' phonotactic acceptability.

Bailey and Hahn (2001) formalise a model of phonotactic judgement, the Generalized Neighborhood Model, which incorporates neighbourhood density alongside bigram and trigram phoneme frequency (i.e. phonotactic probability). This performs relatively

well at modelling the "wordlikeness" of nonwords, and Bailey and Hahn show that the effect of neighbourhood density is independent of any of the other effects. This suggests an effect of lexical analogy – distinct from that of phonotactic probability – determines phonotactic acceptability.

Analogy and statistical probability (§2.3.1), though both involve abstracting from the lexicon, are not equivalent. Frisch (1996: 163) uses this to account for the non-uniformity between acceptable English /stVt/[10] (given *stout*, *stat*, *stoat* etc.) on one hand and unacceptable /spVp/ and /skVk/ (no analogous forms) on the other. These three forms differ only marginally in featural similarity and the transitional probabilities of their segments, but there are no instances of /spVp/ and /skVk/ in the lexicon from which analogies can be made. Hence, according to Frisch, words of these types are disproportionately penalised in new word formation[11]. For Frisch, analogy compounds phonotactic probability; both are active in determining phonotactic acceptability.

Davidson (2006) also argues for a distinction between discrete analogy and lexical statistics. In a test of word-initial fricative-obstruent cluster production, she notes that there is no effect of frequency (type or token) of those clusters in other word positions. For example, the relative frequency of medial /zb/ (in e.g. *husband* or *frisbee*) has no effect on production accuracy compared to totally unattested clusters (e.g. /fm/). However, different fricative-obstruent clusters do have significant differences in relative acceptability for English speakers, which tend to follow the cline[12]:

(7) sC > fC > zC > vC

Davidson uses this as evidence for discrete featural analogy as opposed to frequency-based lexical statistics. However, the frequency statistics that Davidson uses may not be expected to strongly correlate with acceptability in this case. The effect of cluster frequency elsewhere in the word on phonotactic acceptability is supported by prior work (e.g. Jusczyk et al., 1994), but this is by no means the only measure of frequency. Models of phonotactic probability which make generalisations over features (e.g. Bailey and Hahn, 2001; Hayes and Wilson, 2008) can likely get Davidson's effects using statistics alone without resorting to discrete analogies. This is due to the fact that (for example) /s/ shares more features with /f/ than it does with /v/. Davidson (2010) also critiques her previous conclusions, after finding a similar cline of production accuracy in Catalan, a language in which the sC clusters are *less* clearly analogisable than the fC clusters. Thus, whether analogy is the motivation behind this pattern is questionable. Indeed, positing inspecific 'analogy' with no positive evidence to support such a claim could act as a last resort for when alternative explanations do not fit the data.

Daland et al. (2011: 221) show that neighbourhood density alone cannot wholly account for phonotactic acceptability. They give the example of *guzu* and *bzoker*, both

[10] Where V = any vowel

[11] This contrasts with Coetzee's (2008) use of formal markedness constraint interaction to achieve the same outcome.

[12] Where C = any other given consonant

of which are one phoneme away from one attested word (*guru* and *broker* respectively, /r/>/z/ in both cases). This results in both having the same acceptability under a simple neighbourhood density model. Yet *bzoker* is unambiguously a less acceptable word of English. Daland and colleagues use this as a reason for the necessity of using contextual information (e.g. syllable structure or biphone probability) in determining phonotactic acceptability – though this does not mean that neighbourhood density is of no use at all.

3 Empirically testing the effect of sonority

3.1 Predictive differences

Past studies on sonority projection have almost entirely failed to identify solid predictive differences between views of sonority projection as lexical statistics and views of sonority projection as (universal) markedness. Berent et al. (2009) is one of the few studies to do this, though their methodology, as outlined above, is open to criticism. Testing predictive differences has proven fruitful in other studies of phonotactic judgements. Frisch and Zawaydeh (2001) and Coetzee (2008) both examine consonant co-occurrence restrictions, in Arabic and English respectively. Both come to the conclusion that such restrictions cannot be explained by lexical statistics alone, therefore requiring speakers to have knowledge of an abstract co-occurrence restriction. It is an open question as to whether the same level of abstraction is necessary for sonority projection.

In the view of Berent et al. (2009: 77), a universal sonority hierarchy entails that "the learner must end up formulating just those generalisations that coincide with sonority-sequencing principles and not others that contradict those principles". This is an empirically strong and testable claim, one that can be disproven by showing that speakers have preferences that violate the SSP. Jarosz and Rysling (2017) take a similar view, arguing that phonotactic models without an analytic SSP-like bias, like that of Hayes and Wilson (2008), are "not sufficient for deriving the sonority sequencing preferences of Polish speakers". They argue that this is due to the fact that "nothing prevents such models from inducing constraints with contradictory effects to the SSP, given appropriate evidence". Like Berent and colleagues, Jarosz and Rysling see the SSP as inviolable in the case of sonority projection.

Yet the results of Berent et al. (2009) can be explained away with lexical statistics. As outlined in §2.2.1, the authors failed to control for position-specific phoneme frequency. This leaves open the possibility that this (rather than an analytic bias towards sonority) could be the explanation for the SSP-like effect in their results. Similarly, the results of Jarosz and Rysling (2017: 11) may also be explicable without resorting to an analytic bias. The attested sonority falls examined by Jarosz and Rysling are all very infrequent (in both token and type frequency). If speakers statistically generalise from these infrequent attested clusters, the unattested sonority falls should also be very improbable, and thus less acceptable. Therefore, a model of sonority projection which works purely off lexical statistics would also predict that the sonority

falls would be less acceptable. Jarosz and Rysling thus fail to prove that lexical statistics cannot account for their data. They also fail to control adequately for position-specific frequency (as detailed above).

The results of Berent et al. (2009) and Jarosz and Rysling (2017) therefore may not be conclusive in disproving a purely lexicalist approach to sonority projection. Nor is the model comparison approach of Daland et al. (2011) or Hayes (2011) conclusive proof against a universalist approach. While modelling is a useful approach to theory comparison, these models are at best an approximation of how speakers judge phonotactic acceptability. Any effects that are unexplained by a particular model may be explicable in a better model – hence the need to test predictive differences between theories. To properly examine whether sonority projection relies on more than lexical statistics, we need to find a case where the universalist and lexicalist hypotheses make contrastive predictions. Such a test case should be able to solve the question of whether the SSP is an analytic bias. First, however, it is necessary to define what is meant by both sonority and lexical statistics in order to create a watertight test case.

3.2 Defining sonority and lexical statistics

3.2.1 Sonority

For the purposes of comparison with much other research done in the field (especially Berent et al., 2007; Berent et al., 2009; Davidson, 2006), the present study will examine the sonority hierarchy in (4). This, crucially for this experiment, ranks stops as less sonorous than fricatives and thus predicts that onsets with stop-fricative orders should be preferred to fricative-stop orders in cases of sonority projection. The sonority hierarchy in (5) makes this same prediction.

3.2.2 Phonotactic probability

In evaluating lexical statistics' effect on phonotactic acceptability, it is worth remembering the distinction between phonotactic probability and neighbourhood density (similar to, but not necessarily the same as, analogy) outlined in §2.3.

The standard measure of phonotactic probability, biphone transition probability, is obviously invalid for totally unattested clusters, which by definition have a probability of zero. As such, estimates of their phonotactic probability have to be based on statistics about similar attested clusters; this is the approach taken by Hayes and Wilson (2008) and Albright (2009). Therefore, for unattested clusters, phonotactic probability can be defined as (8), where a = the set of attested two-consonant onset clusters, n = the number of attested two-consonant onset clusters (in English), c = a given unattested two-consonant cluster:

(8)

$$(\textit{frequency of } a_1 \times \textit{similarity of } a_1 \textit{ to } c) + \ldots (\textit{frequency of } a_n \times \textit{similarity of } a_n \textit{ to } c)$$

This is very similar to the feature-based biphone probability measure shown by Albright (2009) to correlate well with acceptability, and broadly similar to the phonotactic probability metric in the Phonotactic Probability Calculator outlined by Vitevitch and

Luce (2004), with an added similarity metric to account for the unattested nature of the clusters at hand. The online calculator was used to compute the frequency of a given attested cluster word-initially (i.e. as the first two consonants). This calculation was done for each attested two-consonant cluster in English, and then each frequency-adjusted attested cluster was compared for similarity to each of the unattested clusters in the stimuli. The results were then summed to give a score for each unattested cluster which represents its frequency-weighted similarity to all attested clusters in English. The frequency-weighted similarity to all attested clusters is the aggregate measure of a given cluster's phonotactic probability. Appendix B details the full calculation.

One issue with Vitevitch and Luce's Phonotactic Probability Calculator is that it measures token frequency, not type frequency. This is contrary to numerous findings that type frequency is a better predictor of a given pattern's phonotactic acceptability than token frequency (Hay et al., 2004; Hayes and Wilson, 2008; Albright, 2009). The online CELEX corpus (Van Gerven, 2001, based on data from Baayen et al., 1995), which has been used in other literature on phonotactic acceptability (e.g. Frisch, 1996), includes both segmental transcription and the ability to find type frequency. However, it was unfortunately not possible to make use of this corpus as its web interface is both byzantine and unsupported as of July 2019. The token frequency-weighted data therefore have to serve as an approximation of frequency as it applies to phonotactic acceptability.

The metric of cluster similarity was adapted from Frisch's (1996) consonant-pair similarities[13]. To find the similarity of a pair of clusters, the equation used was (C1 = first consonant in the cluster; C2 = second consonant in the cluster):

(9)

$$\textit{similarity of } C1_a \textit{ to } C1_c \times \textit{similarity of } C2_a \textit{ to } C2_c$$

Cluster similarity is similar to feature-based generalisation of the kind outlined in Albright (2009).

3.2.3 Neighbourhood density

Neighbourhood density (see §2.3.2) is the second major aspect of lexical generalisation which has been shown to correlate with phonotactic acceptability (Greenberg and Jenkins, 1964; Charles-Luce and Luce, 1990; Bailey and Hahn, 2001). Single-phoneme edit distance is the "standard measure" of neighbourhood density (Bailey and Hahn, 2001: 571). Bailey and Hahn describe a single-phoneme edit distance neighbour as "any word that can be derived by substituting, deleting, or inserting a single phoneme". Some authors have also examined neighbourhood density at the segmental level (e.g. Frisch, 1996); this is essentially encoded in the cluster similarity metric above.

[13] These similarities are based on shared natural classes, and as such rely on some theoretical assumptions as to what is featurally encoded. However, Frisch (1996) shows good correlations with OCP effects in Arabic and English as well as with speech error likelihood. While probably not perfect, Frisch's figures present a reasonable approximation of similarity, and are used by a number of other phonotactic studies and models (e.g. Bailey and Hahn, 2001; Hayes and Wilson, 2008).

Variation in neighbourhood density (and resulting lexical analogy) can be minimised by the experiment design; this is done by using the same rime in both stimuli in each pair. Nevertheless, there are still small effects of neighbourhood density in this design. Consider the pair *tnot-fnot*. While *tnot* neighbours *tot*, *fnot* has no equivalent (**fot*)[14]. The nonword *tnot* therefore has a slightly denser neighbourhood than *fnot* and thus would be expected to be more acceptable, all else being equal. Designing experimental stimuli which achieved equal neighbourhood density alongside controlling for all other factors proved near-impossible. Instead, a predictor encoding the small neighbourhood density differences between a few of the stimuli will be added to a mixed-effects model to examine whether these differences had any effect.

3.3 Isolating two predictive differences

Two specific predictive differences were tested. The first relates to the relative order of stops and fricatives; henceforth, this is the 'stop-fricative condition'. The universalist hypothesis straightforwardly predicts (where 'T' = any stop, 'F' = any fricative, 'N' = any nasal, 'X > Y' = X is more acceptable than Y):

(10) TF > FT

However, the lexicalist hypothesis predicts the *reverse* (for calculations, see Appendix B). Based on the formula for phonotactic probability in (8), the lexicalist view predicts that:

(11) FT > TF

The second predictive difference (henceforth the 'nasal condition') relates to stops and fricatives preceding nasals. The universalist hypothesis predicts that:

(12) TN > FN

The lexicalist hypothesis predicts the *same* as the universalist hypothesis for the nasal condition (see Appendix B for calculations). The nasal condition's expected outcome is thus slightly different to that of the stop-fricative condition; the nasal condition acts as a control for the lexicalist hypothesis. If we see a preference for SSP-violating clusters, it should manifest itself only in the stop-fricative condition. A preference for SSP-violating clusters in both conditions would disprove both hypotheses.

The predictive differences are summarised in Table 1:

Table 1: Predictive differences of the lexicalist, universalist and null hypotheses

	Stop-fricative condition	**Nasal condition**
Lexicalist:	TF < FT	TN > FN
Universalist:	TF > FT	TN > FN
Null hypothesis:	TF = FT	TN = FN

[14] At least in the varieties of participants in this study, none of whom had the *cot-caught* merger

For this study, the target fricative was chosen to be as featurally and perceptually close to /s/ as possible, as /s/-stop clusters (/sp/, /st/, /sk/, /sm/, /sn/) are the main source of the SSP-violating lexical generalisations for English. The fricative /f/ was chosen over /z/ after a short pilot study in which /z/ was frequently misperceived as /s/. Davidson (2006) also suggests that English speakers find /f/ to be the most easily analogisable fricative from /s/.

The clusters tested were therefore as follows. Note that the cluster pairs in Table 2 are those which test the hypothesis; this is the condition for which there is a predictive difference between the universalist and lexicalist hypotheses. The cluster pairs in table 3 serve as the control.

Table 2: The cluster pairs for which there is a predictive difference

SSP-violating cluster	**SSP-conforming cluster**
/fp/	/pf/
/ft/	/tf/
/fk/	/kf/

Table 3: The cluster pairs for the control condition

SSP-violating cluster	**SSP-conforming cluster**
/fm/	/pm/
/fm/	/tm/
/fm/	/km/
/fn/	/pn/
/fn/	/tn/
/fn/	/kn/

Bailey and Hahn (2001) noted a significant predictiveness of orthographic bigram and trigram frequency on nonword acceptability judgements, when the nonwords are presented orthographically. This effect was not present for auditorily-presented stimuli. As such, two tasks were conducted: one with spoken stimuli and another with written stimuli. If there is a significant difference in results between the two conditions, further examination of orthographic factors may be necessary.

4 Experiment

Eleven participants' results were collected over two experimental tasks. Both tasks were presented via computer, and were created using Praat's (Boersma and Weenink, 2019) ExperimentMFC interface.

The first task was a listening task, in which participants heard two stimuli. Each stimulus was associated with a button, and participants were asked to press the button corresponding to the more acceptable word of English. Participants then heard the

sound again, and were asked to write down each word as best they could, on the grid provided on a sheet of paper. The experimenter's instructions were as follows:

Listening task

> "You will hear two words, and you should choose which of the two you think is a more possible word of English. To choose a word, press one of the two buttons with the mouse, or press the '1' or '2' keys.
>
> When you have heard each word, click 'write down'. Then, please write down both words. You can hear them again by pressing the 'repeat' button onscreen or the spacebar. You will hear each word again once, and you should write both words down as best you can.
>
> Once you have written both words down, click 'next'. You should then repeat the process.
>
> You may also stop the experiment at any point if you wish.
>
> Feel free to ask me any questions."

Reading task

> "You will see two words and you should choose which is a more possible word of English. To choose a word, click the yellow button below where that word is written.
>
> Once you have chosen a word, click 'next'. You should then repeat the process.
>
> You do not have to write any words down.
>
> You may also stop the experiment at any point if you wish."

The second task was a reading task, in which participants were presented with two written stimuli onscreen. Each stimulus was associated with a button, and participants were asked to press the button corresponding to the more acceptable word of English (as above). Participants were not asked to write anything.

Participants were asked for 'more possible' words of English, rather than 'more acceptable' words (as discussed up to this point). When asked informally, a number of potential participants[15] suggested that 'more/less acceptable' implied metalinguistic value judgement (for example, on how rude or polite the nonword sounded). It was thus decided to ask for 'more possible' words instead. For consistency and comparability to other research in the field (e.g. Albright, 2009), words that participants judged more 'possible' will continue to be referred to as more 'acceptable'. This also has the advantage of centring the participants' judgements rather than implying that there is some abstract notion of what is or is not possible as a word of English (cf. Algeo, 1976).

[15] None of these potential participants were then tested.

4.1 Participants

Fifteen participants, all native speakers of English from the UK and Ireland, were tested in total. The results of only eleven of these participants, however, were included in the final results. One was rejected after testing for speaking German with a parent while growing up, another for having spent one year in a Spanish-speaking environment, and another for revealing a medical diagnosis that may have impaired his ability to concentrate on the tasks. Four of the eleven remaining participants were female, and seven were male. Participants' ages varied from 19 to 25 (mean 21.8, median 22, sd 1.58). Listeners had a range of language backgrounds, from two participants with three years' secondary education in one foreign language to one participant with knowledge of French, German and Mandarin Chinese. None of the thirteen participants indicated that they had ever lived in an environment in which most of their daily interactions were not conducted in English, and none considered themselves to have any native language other than English.

4.2 Generating the stimuli

The target stimuli were all CCVC monosyllables, given in pairs. The clusters tested were all in onset position for three reasons. The first was to control for effects of word position; /ts/ is a valid coda but not a valid onset. The second was because the most salient part of word disproportionately affects its acceptability (Sendlmeier, 1987 [in Frisch et al., 2000]; Daland et al., 2011); thus we should expect the most visible effects if testing onset clusters. The third is that onsets in English are not vulnerable to morphological decomposition, which has been shown (Hay et al., 2004; Needle et al., in press) to affect nonword acceptability; a word such as *feps* could be parsed monomorphemic or as (plural, bimorphemic) *fep+s*, while *spef* could only be parsed monomorphemically.

All stimuli were recorded by a phonetically-trained male native speaker of British English (i.e. the author) and checked manually to ensure all stops (including final stops) were audibly released, no fricatives were voiced and no consonant clusters had evidence for an epenthetic vowel between their first and second consonants. All these errors are noted by Wilson and Davidson (2013) as common in the production of similar clusters. All stimuli were equalised in loudness after recording using a Praat script kindly provided to me by University of Amsterdam speech lab manager Dirk Jan Vet.

Onset pairs were assigned to one of a set of eight rimes: /ɪt ɪd æt æd ɛt ɛd ɒt ɒd/ (transcribed as in RP). All vowels chosen were relatively frequent and hypothesised to be present in all speakers' varieties of English (unlike /ʌ/ and /ʊ/, which are not contrastive in a number of varieties in England (Wells, 1982: 351)). Only short lax vowels were included to avoid effects of vowel length, tenseness or diphthongisation on acceptability.

The final consonant of a stimulus could be /t/ or /d/. These two consonants were chosen to provide some variation in codas so as to distract speakers from guessing the

dependent variable. Both are relatively similar phonologically, differing only in voicing. Only the coronal stops were chosen to avoid long-distance OCP effects which penalise sequences of the types CpVp, CpVb, CkVk and CkVg (where C is a consonant and V is a vowel; Coetzee, 2008). Both /t/ and /d/ are frequent in syllable codas (Treiman and Kessler, 1997).

The target stimuli were presented in pairs, with participants forced to deem one stimulus more wordlike. As all target stimuli were thought to be unlikely words, head-to-head comparison was deemed more suitable than rating on a scale, averting floor effects whereby all stimuli are given low ratings (Daland et al., 2011: 12). Daland et al. found little difference when comparing head-to-head and scalar judgements, but did notice a floor effect for the former.

Participants were also presented with filler stimulus pairs, of which there were 39 in the listening task and 16 in the reading task. These were CCVC and CVC monosyllables, with rimes balanced as in the target stimuli and onsets selected from a set including those used in filler tasks. Filler stimuli were chosen to represent a spectrum of acceptability, from acceptable stimuli such as *sot* (/sɒt/) to unacceptable stimuli such as *pket* (/pkɛt/). The initial clusters consisted of the consonants {s f p t k n m} (i.e. the same as those for the target stimuli, with the addition of /s/).

There was a small neighbourhood density difference (see §3.2.3) between the items in a few (n = 5) stimulus pairs as measured in the WebCELEX corpus (Van Gerven, 2001). In each case, the SSP-violating stimulus had one fewer lexical neighbour (at single-phoneme edit distance) than the SSP-conforming stimulus. There were thus only two categories of stimulus pair: one category for stimuli with equal neighbourhood density and another category in which the SSP-violating stimulus had one fewer lexical neighbour.

All participants were presented with the same stimuli, but with their orders randomised within each block. The full set of stimuli (target and filler) is given in Appendix A.

4.3 Task 1: listening task

4.3.1 Experiment design

The task consisted of six blocks, of which the first was for training purposes and for which results were not recorded. The training block consisted of nine stimulus pairs, selected for two purposes: firstly, ensuring participants accurately perceived the acoustic difference between the speaker's /s/ and /f/, and secondly, familiarising the participants with the various types of cluster to occur thenceforth. The five following blocks each consisted of twelve pairs. Each of these five blocks contained a mixture of target pairs (three or four per block) and filler pairs (eight or nine per block).

The target stimuli consisted of 21 stimulus pairs (42 stimuli in total). Both items in each pair shared the same rime; thus the only difference between each member of a target pair was the initial cluster. Each stimulus pair contained both members of one of the cluster pairs listed in §3.3, with two exceptions: the homorganic pairs /pm/-/fm/

and /tn/-/fn/ were excluded from the target stimuli for the listening task. This is due to their frequent misperception in the pilot study; this misperception likely occurred due to decreased cue salience or coarticulation. Thus the seven target stimulus pairs for the listening task were as follows:

/pf/-/fp/		/pn/-/fn/
/tf/-/ft/	/tm/-/fm/	
/kf/-/fk/	/km/-/fm/	/kn/-/fn/

Ten stimulus pairs had the SSP-violating cluster as the first stimulus and the SSP-conforming as the second; the other eleven had the order reversed. All instances of final /t/ and /d/ in the target and filler stimuli contained clear and loud releases. Target stimuli were matched such that ten ended with /d/ and eleven with /t/.

The rimes in the target stimuli originally contained equal numbers of each vowel, but this balance was sacrificed to ensure perceptually clearer onset clusters. Pairs involving a stimulus misperceived by two listeners in a short pilot study were swapped with clearer stimulus pairs containing the same initial cluster. This unfortunately resulted in a slight imbalance in rime frequencies (e.g. rimes with /æ/ were more common in the target pairs than those with /ɪ/).

Immediately after making a judgement, participants were asked to write both stimuli 'as best you can' to avoid the possibility that the stimuli had been misperceived. This controlled for the well-attested effects of misperception in novel consonant clusters (Dupoux et al., 1999; Wilson and Davidson, 2013). Participants had the option to hear both stimuli once more before writing.

4.3.2 Filtering results

Stimuli which were perceived differently from intended were analysed as tokens of their percept rather than of their target – but only if the percept cluster was also tested elsewhere in the experiment. For example, if a listener heard /pnɪt/ as /knɪt/, it was analysed as the percept; this correction was made for ten stimulus pairs. Where such misperceptions resulted in a difference between the two items of a stimulus pair in *both* consonants (e.g. perceiving the pair /pnɪt/-/fnɪt/ as /pnɪt/-/fmɪt/), that stimulus pair was discarded. All other stimulus pairs in which one or both stimuli were misheard (n = 83) were discarded.

Responses to target stimuli were discarded if the stimuli were written as anything but an obvious transcription of the intended target. Some variance in transcription was tolerated, including writing /f/ as <f> or <ph> and /k/ as <c> or <k>. Any transcription that suggested misperception (e.g. participant DB's <ferbid> for [fpɪd]) resulted in that pair being discarded. Common mistranscriptions included vowel epenthesis (cf. Berent et al., 2007) and stop voicing in /f/-stop clusters, perhaps paralleling repairs in the production of similar clusters shown by Wilson and Davidson (2013). Stimuli were also discarded if the vowel was not written as intended, as a difference in vowels between the two stimuli in a pair would contribute to different neighbourhood effects (see §3.2.3).

4.4 Task 2: reading task

Various effects of production can affect which stimulus is judged 'more possible as a word of English'. Wilson and Davidson (2013) showed a number of effects of minor phonetic details on categorical perception of unattested onset clusters. For example, small variations in the burst amplitude of stops created significant differences in the rate of misperception. Though Wilson and Davidson measured misperception (and the present study's reading task *excluded* misperceived stimuli), it follows from their results that minor phonetic cue variation can be used by listeners of unattested onset clusters, and thus could have a significant effect on acceptability. Though the stimuli for the listening task were recorded and screened by a trained phonetician and equalised in loudness, it was impossible to eliminate phonetic variation between phonologically similar stimuli (e.g. the amplitude of the final release burst in /pnɪt/ versus /fnɪt/). For this reason, the second experimental task asked participants to read two words from a screen. This reading task consisted of two blocks of eighteen pairs, with ten target pairs and eight filler pairs per block. The pairs were randomised within each block. Each pair in the first block also occurred in the second, meaning each pair was presented twice.

Unlike in the listening clusters, the target pairs also included homorganic nasal clusters <tn> and <pm>. These were excluded from the listening task after a short pilot study revealed that they were frequently misperceived by listeners. However, when written, the cue and coarticulation effects that might cause such misperception are eliminated. Post-hoc analysis revealed no significant difference in judgement between the homorganic and heterorganic nasal clusters in the reading task. Thus the nine target stimulus pairs for the reading task were as follows:

/pf/-/fp/	/pm/-/fm/	/pn/-/fn/
/tf/-/ft/	/tm/-/fm/	/tn/-/fn/
/kf/-/fk/	/km/-/fm/	/kn/-/fn/

The target consonant cluster /kn/ was represented as <cn> (rather than <kn>) in the reading task, to avoid such clusters being interpreted with 'silent' <k> (as in *knot*, *knead*, *know* etc.). Four stimulus pairs had the SSP-violating cluster as the first stimulus and the SSP-conforming as the second; the other five had the order reversed. The spelling for all stimuli can be found alongside their phonemic transcriptions in Appendix A.

The stimulus pairs were assigned the same set of rimes as in the listening task. As no unclear stimuli had to be substituted, the rimes were more balanced, with two to three occurrences of each vowel and five of each /d/ and /t/.

The design of this task parallels that used by Jarosz and Rysling (2017), who presented Polish nonwords to their participants solely orthographically, finding an SSP effect which they claim is independent of lexical statistics. It diverges from their design by using head-to-head stimulus presentation rather than asking participants to rate the stimuli on a scale.

4.5 Modelling the data

A linear mixed-effects regression analysis was carried out using the *glmer* function from the *lme4* package (Bates et al., 2015) in R (R Core Team, 2019). The predictors were task type (a binary variable, with levels for the reading task and the listening task), neighbourhood density (a binary variable, with levels for 'equal neighbourhood density' and 'one fewer neighbour'; see §4.2) and cluster type. The cluster type was a 9-way variable orthogonally coded to include:

a) the contrast between the nasal condition and the stop-fricative condition
b) the contrast between /pf/-/fp/ one the one hand and /kf/-/fk/, /tf/-/ft/ on the other (to explicitly search for whether the homorganicity of /pf/ had an effect)
c) the contrast between /kf/-/fk/ and /tf/-/ft/
d) the contrast between clusters containing /m/ and /n/
e) the contrast between /kn/-/fn/ on the one hand and /pn/-/fn/, /tn/-/fn/ on the other (to explicitly search for whether the attestation of orthographic <kn> had an effect)
f) the contrast between /pn/-/fn/ and /tn/-/fn/
g) the contrast between /pm/-/fm/ and /km/-/fm/, /tm/-/fm/ (to explicitly search for whether the homorganicity of /pm/ had an effect)
h) the contrast between /km/-/fm/ and /tm/-/fm/.

The task type contrast, the neighbourhood density contrast and each of the 9-way cluster type contrasts were manually orthogonally coded in R, in order to allow the model to arrive at interpretable estimates of their effects. Item (i.e. stimulus *pair*) and participant were included as random effects[16].

4.6 Applying predictive differences to the model

Recall that the (lexicalist) hypothesis outlined in §3.3 – in which the SSP is based off lexical generalisations – predicts (11) and (12):

(11) FT > TF

In practice, (11) implies that participants will choose SSP-violating /f/-stop clusters at a significantly greater degree than chance.

(12) TN > FN

In practice, (12) implies that participants will choose SSP-conforming stop-nasal clusters at a significantly greater degree than chance.

[16] The full model in R was therefore (where *result* = either SSP-violating or SSP-conforming, *mode* = reading or listening, *cluster* = cluster type, *nd* = neighbourhood density):
*glmer (result ~ cluster * mode * nd +(cluster * mode * nd|participant)+(1|item), family=binomial)*

Therefore, the lexicalist hypothesis predicts a significant preference *for* SSP-violating clusters in the stop-fricative condition and a significant preference *against* SSP-violating clusters in the nasal condition. The combination of these two significant preferences necessarily entails a significant *difference* between the probability of participants choosing the SSP-violating cluster in the stop-fricative condition and the probability of participants choosing the SSP-violating clusters in the nasal condition. If the lexicalist hypothesis is true, we should therefore find this difference to be significant; this will be encoded in the model as a significant effect of cluster type. This is a necessary, but not sufficient, condition of the lexicalist hypothesis[17] – if the hypothesis is true, we should also find that the preferences in *both* the stop-fricative and the nasal conditions are significant.

Meanwhile, the universalist hypothesis predicts (10) and (12):

(10) TF > FT

In practice, (10) implies that participants will choose SSP-conforming stop-/f/ clusters at a significantly greater degree than chance.

(12) TN > FN

In practice, (12) implies that participants will choose SSP-conforming stop-nasal clusters at a significantly greater degree than chance.

Therefore, the universalist hypothesis predicts a significant preference *against* SSP-violating clusters in the stop-fricative condition and a significant preference *against* SSP-violating clusters in the nasal condition. The combination of these two significant preferences necessarily entails a significant preference *against* SSP-violating clusters overall. If the universalist hypothesis is true, we should therefore find this overall preference to be significant. This is a necessary, but not sufficient, condition of the universalist hypothesis – if the hypothesis is true, we should also find that the preferences in *both* the stop-fricative and the nasal conditions are significant.

4.7 Results

4.7.1 Aggregated results

No significant effect of cluster type (the nasal versus stop-fricative conditions) was found, therefore meaning that there was no significant difference between preferences in the nasal condition and preferences in the stop-fricative condition. This is despite the prediction of my lexicalist hypothesis elaborated above. Nor was a significant main effect across all cluster types found, despite the prediction of the universalist hypothesis. Thus, the results neither prove nor disprove either hypothesis. Nor was a

[17] It would be possible, given enough measurements, to have a significant difference between the probability in the nasal condition and the probability in the stop-fricative condition, even though both conditions show a significant preference in the same direction (i.e. for either sonority-violating or sonority-conforming clusters). Hence the difference is not a sufficient condition.

significant effect of task type (reading versus listening) found. Figure 1 summarises these results, clearly illustrating the similarity of participants' aggregate preferences over all conditions.

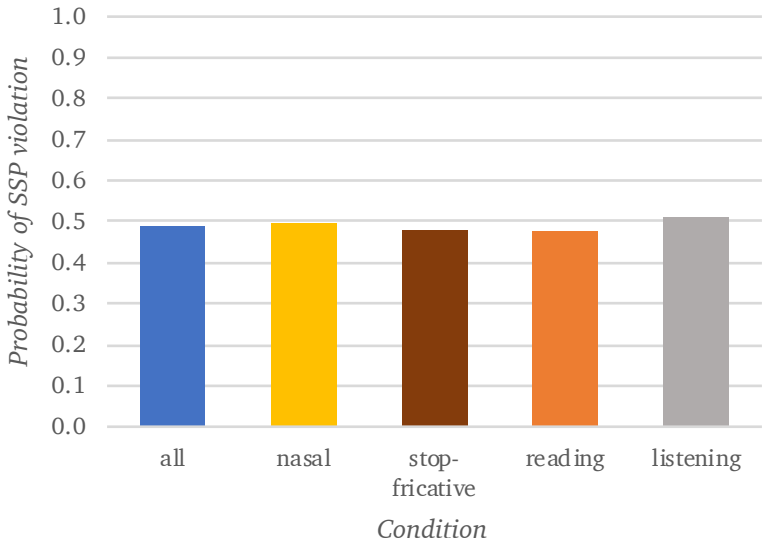

Figure 1: Preference for SSP-violating clusters by condition. The stop-fricative condition is the condition for which there is a predictive difference between the universalist and lexicalist hypotheses.

There was also no significant effect of neighbourhood density.

There was only one significant effect found. This was a preference for SSP-conforming clusters in the /kn/-/fn/ condition, when compared to the average of the /pn/-/fn/ and /tn/-/fn/ conditions. The odds that /kn/ is chosen over /fn/ are 95.85 times higher than the odds that /pn/ and /tn/ are chosen over /fn/ (95% confidence interval = $4.11 \leq x \leq 593.46$, $p = 0.0021$). There was no significant effect of the interaction between the /kn/-/fn/ condition and task type. Nor were there any other significant effects of the interactions between other cluster types and task type or the interactions between cluster type and neighbourhood density.

However, the output of the model contained eleven main effects and eight interactions, giving nineteen estimated effects and nineteen p-values for those effects. Given a significance level of $p < 0.05$, in a random distribution we would expect one p-value in twenty to reach significance. To have one p-value reach significance (in this case, the p-value for the effect of the /kn/-/fn/ condition) is not significantly different from what is expected in 19 effects[18]. However, the effect remained significant when a Bonferroni correction to the significance threshold (for 19 effects) was applied. This created a new significance threshold at $p < 0.0026$, which still makes this effect ($p = 0.0021$) statistically significant. Implications of this result will be discussed below.

[18] To prove that two significant p-values in 19 is not significantly different from expected, a χ^2 test was run ($\chi^2(1) = 0.0000919$, $p = 0.975$) with Yates's continuity correction applied. This did not show a significant difference from the null hypothesis that 5% of trials should return significant p-values.

4.7.2 Results by participant

It is possible that the crude aggregate preferences in the model above mask variation between participants. The model in R encodes participant as a random factor, and thus the results by participant are not statistically interpretable in the model. As such, it is worth examining individuals' preferences further to see if there is any effect in one or more participants that is masked by aggregating the data across participants. Of course, such examination should be treated as strictly exploratory; effects found henceforth do not negate the lack of a significant effect across participants.

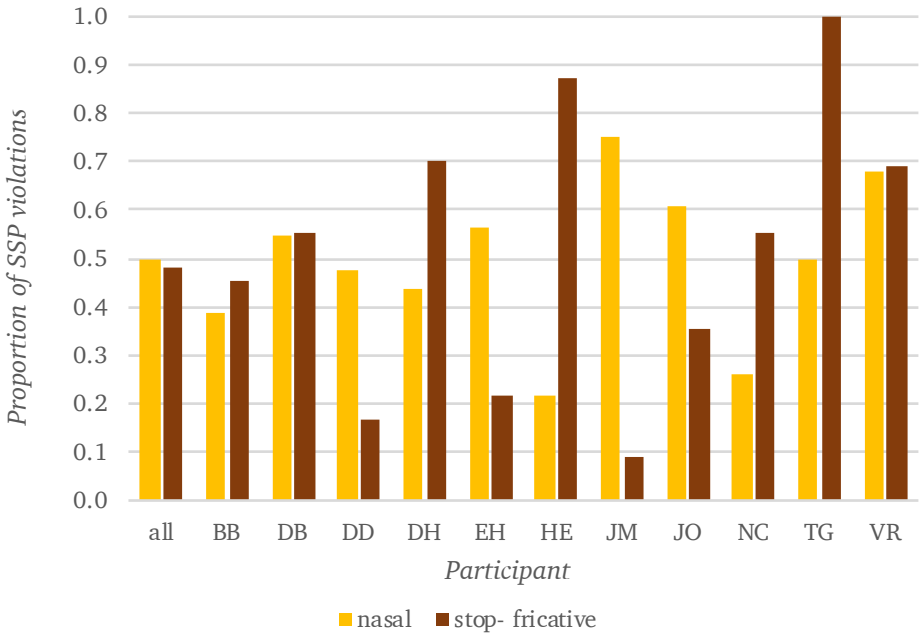

Figure 2: Preferences for SSP-violating clusters by participant and cluster type (nasal/stop)

Figure 2 shows that at first glance, there seems to be considerable variation between participants' preferences in the nasal condition and the stop-fricative condition. Some participants quite reliably prefer SSP-violating or SSP-conforming clusters for either stops or nasals. Compare JM's 9.1% SSP-violating preference on stop clusters with the same participant's 75% SSP-violating preference on nasal clusters.

It is possible to statistically test whether individual participants exhibited preferences (dis)proving either the lexicalist or the universalist hypotheses. Recall from §4.6 that the lexicalist hypothesis predicts a significant preference *for* SSP-violating clusters in the stop-fricative condition and a significant preference *against* SSP-violating clusters in the nasal condition. Similarly, recall that the universalist hypothesis predicts a preference *against* SSP-violating clusters in both the stop-fricative condition and the nasal condition. These preferences can be tested by participant.

To test the possibility that some individual participants may have followed the predictions of either hypothesis, two-tailed binomial tests were carried out on each participant's responses using the *binom.test* function in R. The null hypothesis for these tests was that participants were choosing at random, i.e. with 50% probability of choosing the SSP-violating or SSP-conforming choice. The results of these tests are

displayed in Tables 4 and 5. Results that meet statistical significance ($p < 0.05$) are highlighted. It should be noted that these results are not quite as reliable as the mixed-effects model, as they fail to account for the fact that (due to misperceptions) different participants judged a slightly different subset of the stimuli.

Table 4: Preference for/against SSP violation in the stop-fricative condition by participant

Participant	*SSP-violating judgements*	*Total number of judgements*	*Observed probability of SSP violation*	*p-value*	*Lower probability bound (2.5%)*	*Upper probability bound (97.5%)*
BB	5	11	0.455	1	0.167	0.766
DB	5	9	0.556	1	0.212	0.863
DD	2	12	0.167	0.0386	0.021	0.484
DH	7	10	0.7	0.3438	0.348	0.933
EH	3	14	0.214	0.0574	0.047	0.508
HE	7	8	0.875	0.0703	0.473	0.997
JM	1	11	0.091	0.0117	0.002	0.413
JO	5	14	0.357	0.4240	0.128	0.649
NC	5	9	0.556	1	0.212	0.863
TG	8	8	1	0.0078	0.631	1
VR	9	13	0.692	0.2668	0.091	0.614

Table 5: Preference for/against SSP violation in the nasal condition by participant

Participant	*SSP-violating judgements*	*Total number of judgements*	*Observed probability of SSP violation*	*p-value*	*Lower probability bound (2.5%)*	*Upper probability bound (97.5%)*
BB	7	18	0.389	0.4807	0.173	0.643
DB	11	20	0.55	0.8238	0.315	0.769
DD	10	21	0.476	1	0.257	0.702
DH	10	23	0.435	0.6776	0.232	0.655
EH	13	23	0.565	0.6776	0.345	0.768
HE	5	23	0.217	0.0162	0.074	0.437
JM	18	24	0.75	0.0227	0.533	0.902
JO	14	23	0.609	0.4049	0.385	0.803
NC	6	23	0.261	0.0347	0.102	0.484
TG	12	24	0.5	1	0.291	0.709
VR	15	22	0.682	0.1338	0.138	0.549

There are therefore six significant preferences ($p < 0.05$) in these results by participant. These are:

1) DD's preference for SSP-conforming clusters in the stop-fricative condition
2) JM's preference for SSP-conforming clusters in the stop-fricative condition
3) TG's preference for SSP-violating clusters in the stop-fricative condition

4) HE's preference for SSP-conforming clusters in the nasal condition
5) JM's preference for SSP-violating clusters in the nasal condition
6) NC's preference for SSP-conforming clusters in the nasal condition

There is no clear pattern in preferences across speakers; although, given the lack of significant effects in the aggregate model, this is not unexpected. Only one participant, JM, has a significant result in both the nasal and the stop-fricative conditions. However, this result – a preference for SSP-conforming clusters (i.e. stop-/f/) in the stop-fricative condition but a preference for SSP-violating clusters (i.e. /f/-nasal) in the nasal condition – is not predicted by either hypothesis.

One possible objection to such findings is that the examination of eleven participants over two conditions allows for 22 possible statistically significant results, each of which can then be interpreted. To correct for this, the Bonferroni correction was applied; with 22 results this gives a Bonferroni-corrected p-value of $p < 0.0023$, thereby pushing all effects below the significance threshold. However, this is probably because of the small number of measurements for each condition per participant; note that TG's fully consistent preference for SSP-violating clusters in the stop-fricative condition still does not meet this corrected p-value. Examination of the binomial distribution reveals that, where there are fewer than 10 observations, it is impossible for a fully consistent result like TG's to meet this new significance level.

5 General discussion

5.1 Main results

The results of this experiment are, despite the predictive differences identified, inconclusive. There was no difference found in the effect of sonority between the stop-fricative condition and the nasal condition. Such a difference would have been predicted by the lexicalist hypothesis. The study also found no main effect of sonority across all cluster types; a positive effect here (i.e. a universal bias towards SSP-conforming clusters) would have been predicted by the universalist hypothesis. Therefore, despite identifying a predictive difference between the two hypotheses of sonority projection, these results neither disprove neither a universalist nor a lexicalist hypothesis. Potential reasons for this null result will be discussed below.

The significant preference for /kn/ over /tn/ and /pn/ could be explained by a persistent effect of the attested orthographic bigram <kn>, even when the stimuli were presented auditorily (in the listening task) or with orthographic <cn> (in the reading task). Such an explanation is necessarily speculative, but could prove to be a worthwhile avenue for future research.

The token and type frequency of orthographic <kn> in English is worth noting with regards to this effect. The bigram <kn> is significantly more frequent than any of the other orthographic bigrams that could reasonably represent the consonant clusters in this experiment. All English words beginning with <kn> had a combined 43391

occurrences in the WebCELEX corpus (Van Gerven, 2001, based on data from Baayen et al., 1995). This compares to 88 occurrences of <pn> and 4 of <pf>. The type frequency of <kn> is also much higher; 97 lemmas in CELEX begin with <kn> but only 4 with <pn> and one with <pf>; this is the German loanword *pfennig* which, considering the pfennig's abolition in 1999, is unlikely to have been in the lexicons of many of my participants[19]. None of the other clusters are orthographically represented in CELEX; there are no words beginning with <cn>, <fm>, <kf> or any other reasonable orthographic transcription of the clusters in my stimuli[20]. Thus, the only cluster with robustly attested corresponding orthographic bigrams (/kn/ with <kn>) was significantly preferred by participants when compared to /tn/ (which has no attested orthographic correspondence) and /pn/ (the orthographic correspondence for which is very weak).

Further supporting such an effect is a large amount of psycholinguistic evidence (e.g. Tanenhaus et al., 1980; Barron, 1994; Perre and Ziegler, 2008) for listeners activating orthography even in cases where they only hear an auditory stimulus. We could therefore expect the hearers of /k/ and /n/ to activate orthographic <k> and <n>, and therefore in turn activate English words beginning with <kn>. There is some indirect evidence from the transcription element of the listening task that suggests this could be the case; many participants wrote /kn/ as <kn>. Indeed, the need under the task design to write words down could have primed participants to activate phonology-orthography links. This would contribute to a greater neighbourhood density effect for stimuli beginning with /kn/ – but, in this case, the neighbours would be orthographic rather than phonological.

In the reading task, a link from <cn> to <kn> (and resulting activation of words beginning with orthographic <kn>) is slightly more conceptually complex, but equally plausible. Such an effect requires links from the orthographic form to the phonological form (i.e. in the opposite direction to those in the reading task), activating /k/ and /n/ when reading <cn>. There is also psycholinguistic support for such a link (Lange, 2002). This activation of phonological /k/ and /n/ would then activate orthographic <kn> by the route outlined above.

The null result found overall in this study has a number of possible causes and explanations; the remainder of this section discusses why such a result may have come about. Section 5.2 discusses variation between participants, section 5.3 notes the different predictions made by different models of sonority, section 5.4 considers some shortcomings of the experiment design, while section 5.5 considers how a deeper model of lexical statistics may have yielded statistically significant effects. Section 5.6 outlines how the problems discussed could be remedied in a further study.

[19] Its inclusion in the CELEX corpus is likely due to the fact that the pfennig was still in circulation at the time this corpus was built. The initial cluster /pf/ has a probability of zero in Vitevitch and Luce's Phonotactic Probability Calculator, and so was deemed unattested for the purposes of this experiment.
[20] Such words may be in my participants' lexicons, however. For example, the word *tmesis*, with <tm>, is not in the corpus.

5.2 Variation between participants

Only one participant, JM, was found to have an effect in both conditions. JM had a preference for SSP-conforming clusters (i.e. stop-/f/) in the stop-fricative condition but a preference for SSP-violating clusters (i.e. /f/-nasal) in the nasal condition. However, such a preference is predicted neither the lexicalist nor the universalist hypotheses. The source of JM's preferences is therefore unclear. It could be due to JM making discrete analogies in each condition (§5.4.3), or due to the operation of other constraints which outrank the SSP. It is also entirely possible that this result was due to chance.

Also interesting to ask is if any participant's results disprove either hypothesis for that participant. DD's preference for SSP-conforming clusters in the stop-fricative condition contradicts the lexicalist hypothesis of sonority projection, and therefore seems to disprove a lexicalist explanation of DD's preferences. Meanwhile, TG's preference for SSP-violating clusters in the stop-fricative condition contradicts the universalist hypothesis, and therefore seems to disprove a universalist explanation of TG's preferences.

The significance of these findings could well all be down to chance. To prove that 6 significant p-values in 22 is not significantly different from expected, a χ^2 test was run ($\chi^2(1) = 2.554$, $p = 0.11$) with Yates's continuity correction applied. This did not show a significant difference from the null hypothesis that 5% of trials should return significant p-values.

5.3 Differing conceptions of sonority

In the SSP as outlined by Selkirk (1982) and Clements (1990) (henceforth 'obstruent-SSP'), fricatives and stops are subsumed under 'obstruents', and are unordered on the sonority scale with respect to one another. Subsequent authors (e.g. Prince and Smolensky, 2004) have expanded the scale such that stops are defined as less sonorous than fricatives (henceforth 'extended SSP'), a conception of sonority frequently cited before the formalisation of the SSP (e.g. Levin, 1985: 63). This was the scale used in Berent et al. (2007), from which my formulation of the experiments was based.

However, these two conceptions of the SSP make crucially different predictions about the experimental stimuli. As stated above (§3.3), the extended SSP predicts (where 'T' = any stop, 'F' = any fricative, 'N' = any nasal):

(13) TN > FN
TF > FT

When the obstruent-SSP is considered, however, stops and fricatives are subsumed under the category of obstruent ('O'), such that the generalisation in (13) would be reformulated as follows. Note the equivalence on both items in each pair:

(14) TN > FN becomes ON > ON
TF > FT becomes OO > OO

Such a generalisation obviously makes no sense, as both sides of the new equations are by definition equivalent. The obstruent-SSP therefore does not predict a sonority-based ordering between TN and FN, or between TF and FT. This is a significant predictive difference between the two conceptions of the SSP. My results found no clear preference for either TN over FN or TF over FT, a result which at first seems to support the predictions of the obstruent-SSP. However, arguing that these results support the obstruent-SSP would involve interpreting a null result; in fact, no significant effect *of any kind* was found. Nevertheless, this result would be expected given a view that fricatives and stops are equally sonorous (i.e. unordered with respect to one another in the sonority hierarchy). Further research would be necessary to specifically test the possibility that fricatives and stops are equally sonorous – but given the results presented here, this may be a fruitful avenue for investigation.

The different conceptions of sonority are not so much a flaw with the experiment as a flaw with sonority in general (see §2.2.3). With a principle as ill-defined as the SSP, it is easy to pick whichever formulation best conforms to the data. Steriade (1982) sees sonority as parameterised across languages – but this kind of data-fitting proposal weakens the predictive power of the SSP (and fails to account for the statistical tendencies towards certain parameter settings). It also rests upon the circular logic that if a language is found to violate the sonority hierarchy, the hierarchy can simply be redefined in that language. Any empirical testing of the sonority hierarchy requires a well-defined sonority hierarchy to test against, and the hierarchy chosen for the present study (based on comparability to existing research; §3.2) was not confirmed by the results.

It would be theoretically possible to use a different set of clusters to examine whether the obstruent-SSP is inviolable. For example, one could contrast an SSP-preferred obstruent-nasal cluster with a statistically-preferred nasal-obstruent cluster. However, none of the latter kind of cluster occurs in English, so testing the obstruent-SSP would require examining another language. Jarosz and Rysling (2017) looked at Polish, a language in which onset clusters such as /mzˌ/ are attested. However, they note that such falling-sonority clusters are extremely rare; this suggests a case where lexical statistics prefer such clusters in sonority projection may be hard to find. This points to a flaw in verifiability of the more general formulations of the SSP such as those proposed by Zec (1995) and Clements (1990): they essentially predict the same as lexical statistics, making it hard to identify predictive differences between the two.

5.4 Shortcomings of the experiment design

5.4.1 Weakness of lexical statistics

Perhaps the differences in phonotactic probability between the SSP-conforming and SSP-violating clusters were too weak to have a significant effect on participants' judgements. Seen as this experimental paradigm is novel[21], it is impossible to know

[21] With the exception of Berent et al. (2007), who also failed to find an effect of lexical statistics – though there are potential flaws in this study, as discussed above

what threshold the effect of lexical statistics should reach before it begins influencing participants' judgements.

There also remains the possibility that this study examined the wrong *kind* of lexical statistics. It may be that the formula in (8) bears little resemblance to the statistics that participants actually use in judging phonotactic acceptability. Although Albright (2009) used a very similar measure, his study incorporated this into a model rather than an experiment; he thus did not empirically prove that speakers make use of feature-based bigram probability in judging acceptability. It is also possible that lexical statistics are not used directly in judging phonotactic acceptability, but that the statistics build constraints which are then independently active (Frisch and Zawaydeh, 2001; Hayes and Wilson, 2008). If participants used only this kind of more abstract constraint, it is possible that they would fail to pick up on the raw lexical statistics outlined in §3.2.2 (provided they had not learnt a constraint based on these statistics).

Computing clusters' type, rather than token, frequencies may also have produced clearer results, as type frequency has been shown to be more predictive of phonotactic acceptability than token frequency (Hay et al., 2004; Hayes and Wilson, 2008; Albright, 2009). Computing type frequency was not feasible due to the difficulty of use of the online CELEX corpus (see §3.2.2). However, type and token frequencies are relatively strongly correlated (Albright, 2000; Berg, 2014), so a significant result is unlikely to be found by switching frequency from token to type. In addition, other studies based on Vitevitch and Luce's token frequency-weighted data (e.g. Jusczyk et al., 1994; Vitevitch et al., 1997) do show significant correlation with phonotactic acceptability.

5.4.2 Misperception as unacceptability

Listeners were primed to perceive all stimuli as CCVC, in contrary to the perception experiments in, for example, Berent et al. (2007) and Wilson and Davidson (2013)[22]. Despite this priming, listeners frequently perceived intended CCVC stimuli as bisyllabic CVCVC (with epenthesis) or monosyllabic CVC (with deletion). Despite all stimuli being (intended) CCVC or CVC, five participants wrote at least one stimulus in a way that unambiguously indicated perception as two syllables, for example by writing a vowel between the first and second consonants. These stimulus pairs were excluded from the results, although such misperceptions may in fact have been illustrative of phonotactic unacceptability causing repair during the perception process, along the lines of Dupoux et al.'s (1999) perceptual epenthesis. An explanation of misperception does not hold for the reading task, as stimuli were not presented auditorily.

5.4.3 Many possible analogies

It is possible to speculate that the pattern of large variation between participants could be expected under a model of phonotactic acceptability based on discrete analogies. As there are a number of different analogies which participants could make, it could be possible for different participants to make different analogies in the same context. Consider the example of JM and TG, who showed essentially opposite patterns of

[22] Considering the dependent variable was relative judgement of different clusters, it does not matter that listeners were primed to perceive clusters.

preference in the stop-fricative condition: JM preferred SSP violations while TG preferred SSP-conforming clusters (§4.7.2). Perhaps JM analogised from attested /sp st sk/ to unattested /fp ft fk/, while TG analogised from attested /pɹ tɹ kɹ/ to unattested /pf tf kf/. However, any such line of reasoning is purely speculative; there is no way to empirically prove that the two participants were making different analogies. The need for caution in positing an analogical account is shown by the fact that Davidson's (2006) analogical account of fricative-obstruent cluster production accuracy (§2.3.2) was disproved in a subsequent study (Davidson, 2010).

5.4.4 Difficulty of the tasks

The fundamental experiment design of comparing two similar stimuli with different unattested onset clusters may have been too difficult for participants; a number of participants informally noted this (without prompt) to the experimenter after completing the tasks. However, corroborating evidence suggests that the design is not likely to be at fault. Indeed, the experiment design was very similar to that of Albright (2009) and Moreton (2002), both of whom obtained statistically significant patterns in their results.

5.4.5 Priming effects

One other potential issue relates to the nature of the filler items (detailed in Appendix A). Many of these contained sC- initial clusters (where C = any consonant). This may have primed participants to analogise from this SSP violation to SSP violations in the target stimuli (though see §5.4.3 on the unverifiability of analogy-based explanations). This perhaps negates an analytic bias towards sonority which would otherwise have emerged.

5.4.6 Too many variables

It was not possible to evaluate all possible variables which varied between stimuli, particularly with regards to vowels, codas and vowel-coda interactions, even though these do vary in frequency. This is a flaw in the experiment design: there were too few trials, and too many variables (and *far* too many interactions) to arrive at a computable statistical model; there were simply more measurements than possible parameters. An improved experiment would include fewer rime permutations, at the expense of participants being more likely to guess the dependent variable. This would also remove any possible effects of rimes on acceptability, which could have affected the results in aggregate (though it would not have affected judgements on each pair, as both items in each pair had a common rime).

5.4.7 Too few participants

Considering the individual variation between participants and the relatively small (n = 11) number of participants overall, a study with a greater number of participants could perhaps find more reliable results. Much of the lack of participants was due to the time limits of the thesis and location of the author (in the Netherlands), which made it much harder to access large numbers of (relatively) monolingual native English speakers in the UK and Ireland. This also affects the issue raised in §5.4.6 above; the relatively small number of participants resulted in a relatively small number of measurements, which meant that a complete model controlling for all effects of coda combinations and

frequencies was impossible. However, the number of participants was broadly comparable with a number of other similar studies: Hay et al. (2004) also tested 11 subjects, Berent et al. (2007) had 16 participants for their first two experiments, while Coleman and Pierrehumbert (1997) relied on just six.

5.4.8 Orthographic influence

Despite the efforts to avoid participants in the reading task reading letters as 'silent' (i.e. representing /kn/ as <cn>), participants still could have interpreted <pn>, for example, as representing /n/ (as in *pneumonia* and *pneumatic*). However, this seems unlikely, as the distribution of onsets in the previous listening task should hopefully have primed readers to read these onsets as if both consonants were pronounced: there was only one filler stimulus in the listening task with an onset consisting only of /n/ but nine of /kn/ (and a far larger number of clusters than singleton onsets in general). Nevertheless, readers interpreting letters as silent cannot be ruled out. Perhaps asking readers to speak each word of the pair could have eliminated this possibility, in parallel to how participants in the listening task were asked to write the words they heard.

It also remains possible that participants' acceptability judgements in the reading task were influenced by orthotactic probability. One possible such influence (albeit one for which there was no support in the model) could be that the relative frequency of <ft> in medial (e.g. *after*, *often*) and final (e.g. *soft*, *craft*) positions biased participants to disproportionately prefer this orthographic bigram when at the beginning of a written word. Bailey and Hahn (2001) note that a purely orthographic model could account for some variance in the data for their reading task[23], even once the effect of phonotactics was removed. A more comprehensive experiment would include some kind of control for the frequencies of various letters and letter sequences.

5.4.9 Incomplete neighbourhood density and phonotactic probability

A comprehensive model of neighbourhood density and frequency of each of the stimuli (including the rimes) wasn't considered; this is due to the 'too many variables' issue outlined in §5.4.6. Neighbourhood density was instead measured simply in lexical neighbours of single-phoneme edit distance; there was no direct control for neighbourhood in terms of syllabic constituents (the effect of which was noted by Coleman and Pierrehumbert, 1997). Phonotactic probability was only measured for the initial cluster, and did not include any control for position-specific phoneme frequency. More comprehensive models of neighbourhood density and phonotactic probability (incorporating all the factors that were shown to be predictive in, for example, Bailey and Hahn's (2001) Generalized Neighborhood Model) could perhaps have added some explanatory power to the model constructed here, and potentially improved the significance of the effects.

5.5 Outlining a refined experiment

It is worth briefly considering how the shortcomings outlined above could be minimised in a refined experiment design. First and foremost, a refined experimental methodology

[23] However, they found no independent effect of orthography in their listening task.

would test more participants and would elicit more values for each condition per participant; this would solve the problems discussed in §5.4.6 and §5.4.7. Furthermore, the filler stimuli would be balanced in a way that discouraged priming for anti-SSP or pro-SSP analogies (§5.4.5).

An improved experiment would include just one rime across all stimuli, to control for possible effects of the different rimes' different phonotactic probabilities and neighbourhood densities (cf. §5.4.9). This would sacrifice blinding the participants to the task's exact purpose; participants would know that the acceptability of the initial cluster was the dependent variable. However, the extent to which they were adequately blinded in my experiment was questionable, considering the large variety of initial clusters and small variety of rimes.

Adding predictors for orthographic neighbourhood density (§5.4.8), and a more complete model of lexical statistics including type frequency (§5.4.1) may improve the explanatory power of a mixed-effects model, while asking readers to speak the words they had read would control for the possibility of readers reading letters as silent.

It is unclear how to solve the effect of misperception (§5.4.2) in a task that asks speakers for explicit judgements. Participants could perhaps be told that all stimuli consist of only one syllable, but this would not solve cases where listeners perceived a cluster as a single consonant.

6 Conclusion

In short, the present study found no evidence for an analytic bias in line with the SSP or for lexical statistics in a test case in which the two hypotheses have predictive differences. This lent no support to either the lexicalist or universalist hypotheses of sonority projection. There were a number of possible reasons for this result; these include the possible weakness of lexical effects, the use of token rather than type frequency, misperception, alternative analogies, the difficulty of the tasks, possible priming, possible orthographic influence, and experiment design and execution flaws like the low number of participants or measurements.

Such unclear results might also be expected under a universalist conception of sonority in which stops and fricatives are unordered relative to one another, as posited by Clements (1990) and Zec (1995). However, it is hard to test the universality of such a conception of sonority, as it rarely predicts different outcomes to lexical statistics.

One finding was a disproportionate preference for /kn/ over other clusters, even when the stimuli were presented auditorily. This may suggest a persistent effect of orthographic neighbourhood density, due to the existence of orthographic initial <kn> in English. Such an effect is expected under models of lexical processing that posit links between phonological representations and orthography (e.g. Perre and Ziegler, 2008). However, conclusions on this effect should be taken with some caution, as this was not

the main effect examined and only revealed itself in a mixed-effects model with many parameters; it is possible that this effect was solely down to chance.

The reasoning for the original hypothesis – that lexical statistics would not be overridden by a universal SSP – was still valid. This reasoning highlighted flaws in the notion of sonority, centring on its circularity, violability and vagueness. Another conceptual reason for positing a lexicalist hypothesis rather than sonority as constraint interaction was the inability of established Optimality Theoretic learning models to learn generalisations from the lexicon, despite significant evidence of phonotactic acceptability correlating with the lexicon rather than the input. It is furthermore not clear that experiments claiming to disprove the lexicalist hypothesis (e.g. Berent et al., 2007, 2009; Jarosz and Rysling, 2017) sufficiently controlled for all possible statistical generalisations. The idea that language users' lexical generalisations are based on predictiveness, as opposed to correlation, was also advanced.

The null main result obtained here acts as an impetus to refine the experimental methodology in the hope of achieving an informative result. Briefly, the methodology could be improved by including less variable stimuli, more comprehensive models of neighbourhood density and lexical statistics, and more participants. Further experimentation on the salience of orthographic <kn> in determining the phonotactic acceptability of /kn/ may well also be fruitful.

Bibliography

Albright, A. (2000) The lexical bases of morphological well-formedness. In Bendjaballah, S., Dressler, W.U., Pfeiffer, O.E. & Voeikova, M.D. (eds.) *Morphology 2000: Selected papers from the 9th Morphology Meeting, Vienna, 24–28 February 2000*, p. 5-15. Amsterdam: John Benjamins

Albright, A. (2007) Natural classes are not enough: Biased generalization in novel onset clusters. Paper presented at the 15th Manchester Phonology Meeting, Manchester, 24-26 May 2007

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

Albright, A. & Hayes, B. (2003) Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition* 90: 2, p. 119-161

Algeo, J. (1978) What Consonant Clusters Are Possible? *Word* 29:3, p. 206-224

Bailey, T. & Hahn, U. (2001) Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods? *Journal of Memory and Language* 44, p. 568-591

Basbøll, H. (2005) *The Phonology of Danish*. Oxford: Oxford University Press

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48

Barron, R.W. (1994) The Sound-To-Spelling Connection: Orthographic Activation in Auditory Word Recognition and Its Implications for the Acquisition of Phonological Awareness and Literacy Skills. In V.W. Berninger (ed.), *The Varieties of Orthographic Knowledge I: Theoretical and Developmental Issues*, p. 219-242. Dordrecht: Springer

Becker, M., Ketrez, N. & Nevins, A. (2011) The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 87, p. 84-125

Berent, I., Lennertz, T., Smolensky, P. & Vaknin-Nusbaum, V. (2009) Listeners' knowledge of phonological universals: evidence from nasal clusters. *Phonology* 26, p. 75-108
Berent, I., Steriade, D., Lennertz, T. & Vaknin, V. (2007) What we know about what we have never heard: Evidence from perceptual illusions. *Cognition* 104, p. 591-630
Berg, T. (2014) On the Relationship between Type and Token Frequency. *Journal of Quantitative Linguistics* 21:3, p. 199-222
Boersma, P. (1997) How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam* 21, p. 43-58
Boersma, P. & Weenink, D. (2019) Praat: doing phonetics by computer [Computer programme]. Version 6.0.56, retrieved 2 July 2019 from http://www.praat.org/
Booij, G. (1999) Morpheme structure constraints and the phonotactics of Dutch. In H. van der Hulst & N.A. Ritter (eds.) *The Syllable: Views and Facts*. Berlin: Mouton de Gruyter
Charles-Luce, A. & Luce, P.A. (1990) Similarity neighbourhoods of words in young children's lexicons. *Journal of Child Language* 17:1, p. 205-215
Chomsky, N. & Halle, M. (1968) *The Sound Pattern of English*. New York: Harper & Roe
Clements, G.N. (1990) The role of the sonority cycle in core syllabification. In J. Kingston and M.E. Beckman (eds.), *Papers in Laboratory Phonology I: Between the grammar and physics of speech*, p. 283-333. Cambridge: Cambridge University Press
Coetzee, A. (2008) Grammaticality and Ungrammaticality in Phonology. *Language* 84:2, p. 218-257
Coleman, J.S. & Pierrehumbert, J. (1997) Stochastic phonological grammars and acceptability. In J. Coleman (ed.), *Third Meeting of the ACL Special Interest Group in Computational Phonology: Proceedings of the Workshop*, p. 49-56. East Stroudsburg: Association for Computational Linguistics
Crain, S. & M. Nakayama (1987) Structure Dependence in Grammar Formation. *Language* 63:3, p/ 522-543
Daland, R., Hayes, B., White, J., Garellek, M., Davis, A. & Norrmann, I. (2011) Explaining sonority projection effects. *Phonology* 28, p. 197-234
Davidson, L. (2006) Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics* 34:1, p. 104-137
Davidson, L. (2010) Phonetic bases of similarities in cross-language production: Evidence from English and Catalan. *Journal of Phonetics* 38:2, p. 272-288
Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999) Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance* 25:6, p. 1568-1578
Frisch, S.A. (1996) Similarity and Frequency in Phonology. PhD thesis, Northwestern University
Frisch, S.A., Large, N.R. & Pisoni, D.B. (2000) Perception of Wordlikeness: Effects of Segment Probability and Length on the Processing of Nonwords. *Journal of Memory and Language* 42:4, p. 481-496
Frisch, S.A., Pierrehumbert, J.B. & Broe, M.B. (2004) Similarity Avoidance and the OCP. *Natural Language and Linguistic Theory* 22:1, p. 179-228
Frisch, S.A. & Zawaydeh, B.A. (2001) The Psychological Reality of OCP-Place in Arabic. *Language* 77:1, p. 91-106

Futrell, R., Albright, A., Graff, P. & O'Donnell, T.J. (2017) A Generative Model of Phonotactics. *Transactions of the Association for Computational Linguistics* 5, p. 73-86

Greenberg, J.H. & Jenkins, J.J. (1964) Studies in the psychological correlates of the sound system of American English. *Word* 20, p. 157-177

Halle, M. (1959) *The Sound Pattern of Russian*. The Hague: Mouton

Hay, J. Pierrehumbert, J. & Beckman, M. (2004) Speech Perception, Well-Formedness, and the Statistics of the Lexicon. *Papers in Laboratory Phonology VI*, p. 58-74. Cambridge: Cambridge University Press

Hayes, B. (2011) Interpreting sonority-projection experiments: the role of phonotactic modeling. *Proceedings of the 17th International Congress of Phonetic Sciences*, p. 835-838

Hayes, B. (2012) The role of computational modeling in the study of sound structure. Paper presented at the Conference on Laboratory Phonology, Stuttgart, 27 July 2012

Hayes, B. & Steriade, D. (2004) Introduction: the phonetic bases of phonological Markedness. In B. Hayes, R. Kirchner & D. Steriade (eds.), *Phonetically Based Phonology*, p. 34-57. Cambridge: Cambridge University Press

Hayes, B. & White, J. (2013) Phonological naturalness and phonotactic learning. *Linguistic Inquiry* 44:1, p. 45-75

Hayes, B. & Wilson, C. (2008) A Maximum Entropy Model of Phonotactics and Phonotactic Learning. *Linguistic Inquiry* 39, p. 379-440

Henke, E., Kaisse, E.M. & Wright, R. (2012) Is the Sonority Sequencing Principle an epiphenomenon? In S. Parker (ed.), *The Sonority Controversy*, p. 65-100. Berlin: De Gruyter Mouton

Jarosz, G. & Rysling, A. (2017) Sonority Sequencing in Polish: the Combined Roles of Prior Bias and Experience. *Proceedings of the 2016 Annual Meetings on Phonology*, University of Southern California

Jusczyk, P.W., Luce, P.A. & Charles-Luce, J. (1994) Infants' Sensitivity to Phonotactic Patterns in the Native Language. *Journal of Memory and Language* 33, p. 630-645

Keller, F. (2000) *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. PhD thesis, University of Edinburgh

Legendre, G., Miyata, Y., & Smolensky, P. (1990) Harmonic Grammar: A Formal Multi-Level Connectionist Theory of Linguistic Well-Formedness: Theoretical Foundations: CU-CS-465-90. *Computer Science Technical Reports* 447

Levin, J. (1985) *A Metrical Theory of Syllabicity*. PhD thesis, Massachusetts Institute of Technology

Luce, P.A. (1986) *Neighborhoods of Words in the Mental Lexicon*. PhD thesis: University of Indiana, Bloomington

McCarthy, J.J. and Prince, A. (1996) Prosodic Morphology 1986. Manuscript. http://works.bepress.com/john_j_mccarthy/54/

McCarthy, J.J. (1998) Morpheme structure constraints and paradigm occultation. In M.C. Gruber, D. Higgins, K. Olson & T. Wysocki (eds.), *CLS 32, Part 2: The Panels*, p. 123-150. Chicago: Chicago Linguistic Society

Moreton, E. (2008) Analytic bias and phonological typology. *Phonology* 25, p. 83-127

Needle, J.M., Pierrehumbert, J.B. and Hay, J.B. (in press) Phonological and Morphological Effects in the Acceptability of Pseudowords. In A. Sims and A.

Ussishkin (eds.), *Proceedings of the 2017 Morphological Typology and Linguistic Cognition Workshop*. Cambridge: Cambridge University Press
Ohala, J.J. (1992) Alternatives to the Sonority Hierarchy for Explaining Segmental Sequential Constraints. Papers from the Parasession on the Syllable, p. 319-338. Chicago: Chicago Linguistic Society.
Ohala, J.J. & Kawasaki-Fukumori, H. (1997) Alternatives to the sonority hierarchy for explaining segmental sequential constraints. In S. Eliasson & E.H. Jahr (eds.), *Language And Its Ecology: Essays In Memory Of Einar Haugen. Trends in Linguistics. Studies and Monographs, Vol. 100*, p. 343-365. Berlin: Mouton de Gruyter
Ohala, J.J. & Ohala, M. (1986) Testing Hypotheses Regarding the Psychological Manifestation of Morpheme Structure Constraints. In J. J. Ohala & J. J. Jaeger (eds.), *Experimental phonology*. Orlando: Academic Press, p. 239-252
Parker, S. (2011) Sonority. In M. van Oostendorp, C.J. Ewen, E. Hume and K. Rice (eds.), *The Blackwell Companion to Phonology*, p. 1160-1184. Oxford: Wiley-Blackwell
Peperkamp, S. (2007) Do we have innate knowledge about phonological markedness? Comments on Berent, Steriade, Lennertz, and Vaknin. *Cognition* 104:3, p. 631-7
Perre, L. & Ziegler, J.C. (2008) On-line activation of orthography in spoken word recognition. *Brain Research* 1188, p. 132-138
Pierrehumbert, J.B. (1993) Dissimilarity in the Arabic Verbal Roots. in: *Proceedings of the North East Linguistics Society* 23, p. 367-381
Prince, A. & Smolensky, P. (2004) *Optimality Theory: Constraint Interaction in Generative Grammar*. Oxford: Blackwell
R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
Ren, J., Gao, L. & Morgan, J.L. (2010) Mandarin speakers' knowledge of the sonority sequencing principle. Paper presented at the 20th Colloquium on Generative Grammar, University of Pompeu Fabra, Barcelona
Scholes, R.J. (1966) *Phonotactic Grammaticality*. The Hague: Mouton
Selkirk, E. (1982) The syllable. In H. van der Hulst & N. Smith (eds.), *The structure of phonological representations* (part 2), p. 337-383. Dordrecht: Foris
Smolensky, P. (1996) The Initial State and 'Richness of the Base' in Optimality Theory. Technical Report, Johns Hopkins University
Steriade, D. (1982) *Greek Prosodies and the Nature of Syllabification*. PhD thesis, Massachusetts Institute of Technology
Tanenhaus, M.K., Flanigan, H.P & Seidenberg, M.S. (1980) Orthographic and phonological activation in auditory and visual word recognition. *Memory & Cognition* 8:6, p. 513-520
Tesar, B.B. (1998) An iterative strategy for language learning. *Lingua* 104, p. 131-145
Van Gerven, M. (2001) *WebCELEX*. Retreived from http://celex.mpi.nl/
Vitevitch, M.S. & Luce, P.A. (2004) A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers* 36:3, p. 481-487
Vitevitch, M.S., Luce, P.A., Charles-Luce, J. & Kemmerer, D. (1997) Phonotactics and Syllable Stress: Implications for the Processing of Spoken Nonsense Words. *Language and Speech* 40, p. 47-62
Wells, J.C. (1982) *Accents of English*. Cambridge: Cambridge University Press

Wilson, C. & Davidson, L. (2013) Bayesian analysis of non-native cluster production. In: *Proceedings of NELS 40*, Massachusetts Institute of Technology, Cambridge, Massachusetts
Wright, R. (2004) A review of perceptual cues and cue robustness. In B. Hayes, R. Kirchner & D. Steriade (eds.), *Phonetically Based Phonology*, p. 34-57. Cambridge: Cambridge University Press
Zec, D. (1995) Sonority constraints on syllable structure. *Phonology* 12:1, p. 85-129

Appendix A: list of stimuli

Task 1: listening task

All stimuli are written in IPA notation. The order of the stimulus pairs was randomised within each block.

Target stimuli

Block 1
fkæd kfæd
tfɛd ftɛd
fpɪd pfɪd
knæt fnæt

Block 2
kmæt fmæt
fnɪt knɪt
fmæd tmæd
pfɒd fpɒd

Block 3
kfɒd fkɒd
tfæd ftæd
fnɪt pnɪt
pfæt fpæt
fmɛt tmɛt

Block 4
tfæd ftæd
fmɒd kmɒd
fmɒt tmɒt
fnɪt pnɪt

Block 5
kmæt fmæt
fnɪt pnɪt
knɒt fnɒt

kfɛd fkɛd

Filler stimuli

Block 1

psɛt spɛt
psɛt tmæt
smɒt kmɒt
kɛd slæd
stɒt tsɒt
smɪd fɛt
knɪd snɪd
kɛd slæd

Block 2

pnɛd snɛd
ftɪt slɒd
pnɪd mɪt
ksɛd skɛd
knɛt tmɪt
smɛt tmɛt
spɒd fɒt
psɪd spɪd

Block 3

sɒt næd
spɛt psɛt
smɪd kmɪd
kɛd smɪt
pnɒd snɒd
snæt knæt
pkɛt stɪt

Block 4

snɪd pnɪd
spɛt fɒt
mɒt spɛt
stɛt tsɛt
smæd tmæd
kmæt smæt
psɒd fmɒt
knæt tkɒt

Block 5

stɪd tsɪd
slɒd kfɒt
ksɒt snɛt
snɒd knɒd

smɛt tmɛt
ktæt skæd
fkɛt næd
ksæd skæd

Task 2: reading task

Each stimulus pair was presented twice, once in block 1 and again in block 2. The order of the stimulus pairs was randomised within each block.

Target stimuli

pmid fmid
fmod tmod
kfed fked
pfad fpad
kmat fmat
fkit kfit
fnet cnet
fnad pnad
tfed fted
tnot fnot

Filler stimuli

fnit fmit
pnot psot
sot smot
stit tsid
ksod skod
kpat pkat
snid smid
nat nad

Appendix B: clusters by phonotactic probability

The calculation for phonotactic probability is given in (8) above and restated here:

$$(\text{frequency of } a_1 \times \text{similarity of } a_1 \text{ to } c) + \ldots (\text{frequency of } a_n \times \text{similarity of } a_n \text{ to } c)$$

a = the set of attested two-consonant onset clusters
n = the number of attested two-consonant onset clusters (in English)
c = a given unattested two-consonant cluster

The set of attested two-consonant onset clusters a in English is presented in Table 6, with their relative frequencies in word-initial position[24] compared to all other two-phoneme sequences. There are 25 such clusters[25] as determined by Vitevitch and Luce's (2004) Phonotactic Probability Calculator:

Table 6: Attested clusters by word-initial frequency

Cluster	Frequency	Cluster	Frequency
/fɹ/	0.0056	/gɹ/	0.0081
/fl/	0.0063	/gl/	0.0031
/pɹ/	0.0239	/gw/	0.0001
/pl/	0.0060	/θɹ/	0.0018
/bɹ/	0.0075	/θw/	0.0001
/bl/	0.0050	/sp/	0.0091
/tɹ/	0.0124	/st/	0.0177
/tw/	0.0013	/sk/	0.0078
/dɹ/	0.0048	/sf/	0.0003
/dw/	0.0003	/sm/	0.0017
/kɹ/	0.0094	/sn/	0.0015
/kl/	0.0067	/ʃɹ/	0.0010
/kw/	0.0048		

The similarity of each unattested (target) cluster to each attested cluster is given in (9) above and restated here:

$$\text{similarity of } C1_a \text{ to } C1_c \times \text{similarity of } C2_a \text{ to } C2_c$$

C1 = first consonant in the cluster
C2 = second consonant in the cluster
a, n, c = as above

[24] This includes the first two consonants in triconsonantal clusters /spɹ spl stɹ skɹ skl skw/. The inclusion of these clusters was deemed acceptable, as the possibility of basing analogies off the /sp/ in /spɹ/ (for example) does not seem too far-fetched.
[25] The /j/ in words like *few* /fjuː/ and *new* /njuː/ was deemed to be part of the syllabic nucleus and thus not included in the set of attested clusters.

The results of this calculation are presented in Table 7 below. For reasons of page space, this table is divided in two.

Table 7: Similarity of each target cluster to each attested cluster

Unattested cluster	Attested cluster												
	fɹ	fl	pr	pl	br	bl	tr	tw	dr	dw	kr	kl	kw
pf	0.018	0.010	0.070	0.040	0.028	0.016	0.021	0.027	0.010	0.013	0.031	0.018	0.040
fp	0.100	0.040	0.026	0.010	0.013	0.005	0.010	0.014	0.005	0.007	0.014	0.006	0.020
tf	0.007	0.004	0.021	0.012	0.010	0.006	0.070	0.090	0.027	0.035	0.025	0.014	0.032
ft	0.090	0.110	0.023	0.029	0.012	0.014	0.009	0.003	0.005	0.002	0.013	0.015	0.004
kf	0.010	0.006	0.031	0.018	0.013	0.008	0.025	0.032	0.011	0.014	0.070	0.040	0.090
fk	0.040	0.050	0.010	0.010	0.005	0.007	0.004	0.005	0.002	0.003	0.006	0.007	0.007
fm	0.440	0.170	0.114	0.044	0.057	0.022	0.044	0.044	0.022	0.022	0.062	0.024	0.062
pm	0.114	0.044	0.440	0.170	0.176	0.068	0.132	0.132	0.062	0.062	0.194	0.075	0.194
tm	0.044	0.017	0.132	0.051	0.062	0.024	0.440	0.440	0.172	0.172	0.154	0.060	0.154
km	0.062	0.062	0.194	0.075	0.084	0.032	0.154	0.154	0.070	0.070	0.440	0.170	0.440
fn	0.400	0.530	0.104	0.138	0.052	0.069	0.040	0.012	0.020	0.006	0.056	0.074	0.017
pn	0.104	0.138	0.400	0.530	0.160	0.212	0.120	0.036	0.056	0.017	0.176	0.233	0.053
tn	0.040	0.053	0.120	0.159	0.056	0.074	0.400	0.120	0.156	0.047	0.140	0.186	0.042
kn	0.056	0.056	0.176	0.233	0.076	0.101	0.140	0.042	0.064	0.019	0.400	0.530	0.120

Unattested cluster	Attested cluster											
	gɹ	gl	gw	θɹ	θw	sp	st	sk	sf	sm	sn	ʃɹ
pf	0.015	0.008	0.019	0.010	0.010	0.014	0.003	0.007	0.007	0.010	0.026	0.010
fp	0.008	0.003	0.011	0.054	0.054	0.079	0.011	0.034	0.018	0.054	0.180	0.054
tf	0.012	0.007	0.015	0.030	0.030	0.025	0.009	0.021	0.013	0.030	0.047	0.030
ft	0.007	0.009	0.002	0.180	0.180	0.063	0.034	0.007	0.016	0.180	0.054	0.180
kf	0.027	0.016	0.035	0.011	0.011	0.015	0.003	0.008	0.008	0.011	0.029	0.011
fk	0.003	0.004	0.004	0.063	0.063	0.180	0.013	0.014	0.007	0.063	0.079	0.063
fm	0.035	0.014	0.035	0.007	0.007	0.014	0.047	0.180	0.079	0.007	0.034	0.007
pm	0.035	0.014	0.092	0.004	0.004	0.008	0.026	0.100	0.044	0.004	0.019	0.004
tm	0.075	0.029	0.075	0.012	0.012	0.024	0.078	0.300	0.079	0.012	0.057	0.012
km	0.172	0.066	0.172	0.004	0.004	0.009	0.029	0.110	0.048	0.004	0.021	0.004
fn	0.032	0.042	0.010	0.034	0.034	0.013	0.180	0.047	0.072	0.034	0.011	0.034
pn	0.032	0.042	0.025	0.019	0.019	0.007	0.100	0.026	0.040	0.019	0.006	0.019
tn	0.068	0.090	0.020	0.057	0.057	0.021	0.300	0.078	0.072	0.057	0.018	0.057
kn	0.156	0.207	0.047	0.021	0.021	0.008	0.110	0.029	0.044	0.021	0.007	0.021

Each of the similarity values in Table 7 was multiplied by the frequency for the corresponding attested cluster, and these frequency-weighted similarity values was then summed over all attested clusters. The results of these calculations are the phonotactic probability scores for each target cluster, shown in Table 8. Phonotactic probability scores are multiplied by 1000 for ease of interpretation; original values

were very small. The target clusters and their probabilities are presented with SSP-conforming clusters on the left and SSP-violating clusters on the right.

Table 8: Phonotactic probability by unattested (target) cluster

Cluster	**Phon. probability**	**Cluster**	**Phon. probability**
/pf/	4.1	/fp/	5.5
/tf/	3.8	/ft/	6.9
/kf/	3.8	/fk/	4.5
/pm/	20.5	/fm/	10.5
/tm/	16.4		
/km/	18.8		
/pn/	22.9	/fn/	13.3
/tn/	16.8		
/kn/	20.1		