Acta Linguistica Hungarica, Vol. 58 (4), pp. 353–391 (2011)
DOI: 10.1556/ALing.58.2011.4.1

GRADIENT PHONOTACTIC ACCEPTABILITY
A CASE STUDY FROM SLOVAK

ZSUZSANNA BÁRKÁNYI

Research Institute for Linguistics
Hungarian Academy of Sciences
Benczúr u. 33.
H–1068 Budapest
Hungary
bzsu@nytud.hu

Abstract: Phonotactic well-formedness judgments are usually gradient, the theoretical interpretation of which is controversial in the phonological literature. In this study we present experimental evidence from Slovak that speakers do have intuitions about unattested grammatical forms as well as attested marginal ones and these intuitions can be modeled fairly closely by gradient phonotactic learners like, for instance, the Hayes – Wilson Phonotactic Learner. Our results suggest that in gradient phonotactic judgments the knowledge of the relative probability of various combinations of natural classes plays a decisive role. We pay special attention to sonority reversal clusters in Slovak and claim that these sequences, although attested in the language, are on the verge of grammaticality and thus prone to change.

Keywords: phonotactics, Slovak, gradience, sonority reversal clusters, analogy

1. Introduction

Western Slavic languages, among them Slovak, are well known for allowing “exotic” consonant clusters word-initially that are absent in other Indo-European languages, and are rare cross-linguistically, like[1] S *tkať*, Cz *tkáti*, P *tkać* ‘to weave’ (OCS tъkǫ); S *ktorý*, Cz *který*, P *który* ‘which’ (OCS kъt-); S *pstruh*, Cz *pstruh*, P *pstrąg* ‘trout’ (OCS pьstrъ ‘checkered’); S *lkať*, Cz *lkát*, P *łkać* ‘cry, mourn’ (OCS lъk-); S *ľnúť*, Cz *lnouti*,

[1] “S” stands for Slovak, “Cz” for Czech, “P” for Polish and “OCS” for Old Church Slavonic.

1216–8076/$ 20.00 © 2011 Akadémiai Kiadó, Budapest

P *lgnąć* ‘stick to’ (OCS lei-/lь-); S *mnohý*, Cz *mnohý*, P *mnogi* ‘many’ (OCS mъnogъ); S *mladý*, Cz *mladý*, P *młody* ‘young’ (OCS moldъ), etc. By word-initial onset clusters in this study we mean consonants adjacent on the surface without making reference to any formal syllable theory—see also section **2.3**.

Among (Western) Slavic languages the study of Polish word-initial clusters has been in the focus of phonological interest for over 50 years (see Kuryłowicz 1952; Rubach–Booij 1990; Cyran–Gussmann 1999; Rowicka 1999; Scheer 2004, etc.). Phonologists generally view the patterning of word-initial consonant clusters in a language as the result of a categorical grammar, and aim to account for existing and non-existing consonant sequences in terms of natural classes and rules or constraints that regulate them. However, any analysis of Slavic—especially Polish or Czech—word-initial consonant clusters to date has failed to fully identify existing and non-existing clusters as natural classes. As Scheer (2006, 2) puts it: “Distribution is truly anarchic, i.e. lexical accident.” In Scheer (2007) he concludes that there are and can be only two types of word-initial onset grammars: “TR-only” languages (where “T” stands for any obstruent and “R” for any sonorant) like English, where sonority increases in all word-initial clusters, and “anything-goes” languages where all possible combinations of T’s and R’s can occur, and the absence of any is a lexical accident—Slovak according to this analysis belongs to the second group.

We agree that word-initial consonant clusters in Slovak cannot be accounted for in a categorical model where a cluster either should exist and is perfectly well-formed, or should not, and is totally ill-formed. It is also true that the list of word-initial clusters whose type frequency is really low is surprisingly long. Does this mean that a linguist can do nothing more than list the existing clusters? We hope not. Does the existence of words like *lkať* ‘cry’, *ľpieť*[2] ‘stick to’ in Slovak mean that native speakers would accept any word-initial sonority reversal sequence? Do people think

[2] Examples are given in Slovak orthography unless we find it crucial to provide IPA symbols. The letters *ť, ď, ľ* represent palatalized/(pre)palatal [tj]/[c], [dj]/[ɟ] and [lj]/[ʎ], respectively; the wedge (or hachek) also signals palatal quality, thus: *š* = [ʃ], *ž* = [ʒ], *č* = [ʧ], *dž* = [ʤ], *ň* = [ɲ]. *c* stands for [ʦ], while *dz* represents [dz]. *ch* signals the voiceless velar fricative [x], whereas *h* is realized as the **voiced** laryngeal fricative [ɦ]. *y* is used for [i] (to indicate, after alveolars, that the preceding sound is not palatalized). *ä* is pronounced as [æ] or [ɛ], whereas *ô* stands for the diphthong [u̯o]. In the orthography of this language an acute accent over vowels (and syllabic consonants, such as *ĺ*) signals length.

that *#lk-* is just as good as *#pr-*? Are native speakers' judgments about phonotactic well-formedness really categorical or do they impose some structure on the data to distinguish between common vs. rare vs. absent sequences? We cannot answer these questions without an experiment. The issue interesting to phonologists is whether unattested sequences, that is to say "accidental gaps", are really well-formed in the language, or even more interesting, whether it is possible that some attested sequences are felt ill-formed or on the verge of grammaticality.

The aims of the present study are to show that although Slovak has more liberal word-initial onset phonotactics than most Indo-European languages, it is not an "anything goes" language, and can better be accounted for by gradient grammatical models. These claims are based on empirical data gathered from 38 native speakers of Slovak.

The article is organized as follows: in section **2** we give a brief introduction to earlier analyses of Slovak onset phonotactics and a description of the database used for modeling speakers' judgments; in section **3** the experiment is presented, in section **4** our results are discussed and in section **5** we compare the experimental results with two computational models.

2. Word-initial consonant clusters in Slovak

2.1. Earlier descriptions of word-initial clusters

Slovak descriptive grammars (e.g., Pauliny 1979) generally mention well-formed syllables and word-initial consonant clusters only in connection with syllabification. That is to say, they focus on the syllabic division of word-internal consonant clusters. One of the main discrepancies between authors is the extent to which morphology should be taken into account in the syllabification of intervocalic consonant clusters.

Pauliny defines two basic syllable types: "free syllables" and "bound syllables". Free syllables for him are those where there is no morphological boundary between the segments, except for those where a vowel-initial derivational suffix is attached to the root—these cases are also considered free. Bound combinations are those which contain a morphological boundary, or which are "rare and easy to list". So Pauliny in his classification makes use of morphology and type frequency but not token frequency. Regarding the C_1C_2V type he says that there are nine logically possible C_1C_2V combinations,[3] but he considers only four of them to be free: (i) fricative + stop; (ii) fricative + sonorant; (iii) stop + sonorant; (iv) /v/, /m/ + sonorant. We can say that Pauliny's 'free' onsets are unmarked in the sense that they are cross-linguistically frequent. Group (i) basically corresponds to word-initial sibilant + stop clusters which is the most common violation to sonority sequencing (see the next section). Groups (ii) and (iii) are typologically the most canonical branching onsets. *v*-initial clusters are discussed in connection with the role of morphology in section **2.3** and *m*-initial clusters are dealt with in more detail in **4.2.3**. So we can say that although in a very implicit way, Pauliny gives a grammatical account of word-initial consonant clusters in Slovak, as he groups clusters into phonologically defined subtypes and their frequency in the lexicon is in close connection with the subtype they belong to.

The only work focusing on the syllable structure of Slovak in a generative framework is Rubach (1993). Syllabification in Slovak according to him obeys universal principles like the Sonority Sequencing Generalization and language-specific principles like the Obstruent Sequencing Principle, which says that "with obstruents there is no requirement for sonority distance" (*op.cit.*, 213). Syllabification for Rubach is cyclic and prefix boundaries block the application of the algorithm. Word-final consonants can be extrametrical, word-initial ones are attached to the syllable node by the Initial Adjunction rule (*ibid.*, 239). This rule is ordered before the Liquid Syllabification rule; this is how in words like *ľpieť* 'stick to' the word-initial liquid escapes becoming syllabic. Another restriction he mentions is "not only identical but also near-identical consonants are not permitted to form onsets and codas. Thus no word may begin with **sš*, **zž*, **vf*, or **gk*" (*ibid.*, 214). Rubach's analysis is not as permissive as Scheer's, but it largely overgenerates in the sense that it does not constrain the existence of many more different obstruent clusters. Rubach, obviously did not mean to explain gradual well-formedness judgments of Slovak onsets, and he did not want to capture the type frequency of existing onsets. We may assume, though, that the more complex machinery is needed for the generation of certain syllables the more marked they are. By applying the Initial Adjunction rule all liquid + obstruent onsets can be generated and the vast majority of such clusters should be considered accidental gaps in Rubach's analysis. In this study we do not aim to give

[3] Stop + stop, stop + fricative, stop + sonorant, fricative + fricative, fricative + stop, fricative + sonorant, sonorant + sonorant, sonorant + stop, sonorant + fricative.

a formal account of word-initial onset clusters in Slovak, rather we want to test native speakers' intuitions about typologically marked and unmarked clusters both of which are attested in Slovak and we claim that these intuitions can be fairly closely approximated by non-categorical grammatical models which apply features and natural classes.

2.2. Sonority reversal clusters

One of the most general cross-linguistic patterns of syllable phonotactics is the generalization that the segment highest on the sonority scale constitutes the syllable peak and all the other segments are organized around the nucleus in such a way that sonority increases towards the peak and decreases as we move away from it. This generalization is known in the literature as the Sonority Sequencing Principle (SSP), which was noticed as early as Sievers (1881). More recently, many attempts have been made to formalize it—e.g., Hooper (1976); Kiparsky (1979); Steriade (1982); Selkirk (1982); Clements (1990). Although the exact nature of sonority is still a controversial issue (see, for instance, Ohala 1990) and whether there is a single universal sonority scale or there are language-specific sonority hierarchies is also debated, it is unquestionable that the most common cross-linguistic generalizations of syllable phonotactics can be captured in terms of sonority.

The most common exceptions to sonority sequencing are word-initial $s+\mathrm{C}$ clusters ('s' stands for a sibilant fricative; this is Pauliny's group (i)). These clusters are generally regarded as unmarked in the phonological literature. Most analyses stipulate special syllabification rules or representations to allow for such clusters. Wright (2004, 35) reformulates the SSP as "a perceptually motivated and scalar constraint in which an optimal ordering of segments is one that maximises robustness of encoding of perceptual cues to the segmental makeup of the utterance". In this model $s+\mathrm{C}$ clusters are not exceptional any more since the internal cues in the frication noise of the sibilant are salient enough to identify it, and it is perceptually distant from the following C as long as it is not another fricative.

The cross-linguistic markedness of other sonority reversal clusters, and the fact that they are felt to be ill-formed,[4] sometimes even by speak-

[4] Berent et al. (2007) in a series of experiments with English and Russian native speakers conclude that English speakers' knowledge about the markedness of on-

ers of languages which marginally contain such clusters (as will be shown further on), remains unexplained by Wright (2004) as well. In this paper by 'sonority reversal clusters' we understand a word-initial sonorant followed by an obstruent, as well as sonority plateau clusters, sequences consisting of two sonorants. Scheer (2007) gives a detailed account of these conspicuous word-initial consonant clusters in Slavic where sonority decreases towards the nucleus. The situation within Slavic languages in this respect is scalar (*op.cit.*, 5): Bulgarian, Macedonian, Slovenian and Bielorussian do not have any sonority reversal clusters; others have "almost none" (Upper Sorbian 4, Lower Sorbian 1, Kashubian 4); Slovak represents a midpoint with 'some'; 'quite some' is the label for Ukranian (12) and Russian (16) and 'a whole lot' for Polish (20) and Czech (28).[5] Scheer lists 8 such clusters for Slovak, the database we used, the *Short Dictionary of the Slovak language* (Kačala–Pisárčiková 1987) (ShDSL) contains only 6 (listed further below), the words *msta* 'vengeance' which is normally used as *pomsta* and *mša* 'mass' nowadays used as *omša* are not included in our corpus. This discrepancy probably does not have any bearing on the experiment we conducted and the conclusions we draw.

Scheer dispenses with the idea that the patterning of #RT clusters 'should be explained at all: the clusters and gaps are not enforced by grammar; rather, they are the result of lexical accident' (Scheer 2007, 8), since all Slavic #RTs are rooted in the Common Slavic #R-yer-T sequences and are created by the loss of the yer. So in those Slavic languages which did not react against the new clusters arising in this way, we do not expect any co-occurrence restrictions between the two consonants. Scheer also says that on phonological grounds the frequent occurrence of #*m*T is much less expected than #*n*T which does not occur in any of these languages. He adds that on diachronic grounds this is also explained: in Common Slavic there were many #*m*-yer-T sequences, while there happened to be no #*n*-yer-T sequences, with a yer in weak position. Indeed it seems to be a synchronically unnatural phonotactic constraint since nasal homorganicity is cross-linguistically favoured over non-homorganic nasal + C clusters (this is phonetically motivated, see for instance, Blevins 2004 and the references therein). However, it has also been observed that languages disfavor homorganic consonants flanking very short vowels,

set clusters might reflect universal markedness constraints with sonority reversal clusters being the most marked.

[5] In Bosno-Serbo-Croatian all #RT-clusters are #*r*T-clusters, where the *r* is syllabic.

which Slavic yers in weak position surely were. Remember that Slovak generally disfavors near-identical consonants word-initially (section **2.1**). A counterexample are coronal clusters: *tl*- and *dl*-initial words are fairly well-attested in Slovak (there is one word beginning with *tn*- and three words start with the sequence *dn*-). In their sonority profile these clusters are the most unmarked branching onsets. Non-coronal clusters are limited by OCP-Place in Slovak, too. Furthermore, *m* is unquestionably more obstruent-like than *n* in Slovak (and in other languages as well)—we tackle this issue in section **4.2.3**.

As it has been mentioned in section **1**, Scheer concludes—and gives a CVCV phonology account of the conclusion—that there are and can be only two types of word-initial onset grammars: “TR-only” languages like English, where sonority increases in all word-initial clusters, and ‘anything-goes’ languages where all possible T and R combinations can occur, and the absence of any is a lexical accident.[6] He brings a strong argument from Russian. Russian freely borrowed form Caucasian, mostly Georgian, words with initial #RT clusters, which were absent in Russian. Although there are not many such borrowings and they are all proper names, they are not altered and receive regular inflection. It is true that English is very unlikely to borrow such words and leave them unmodified. Does this necessarily mean that Russian could borrow any cluster? Do speakers think that these words are just as good as any other Russian words, or is it precisely their phonotactics that defines that they belong to a different group in the lexicon, that of ‘proper names’ or ‘foreign names’? How exactly are they pronounced? Although we will not answer these questions regarding Russian, we will look at what Slovak grammarians think of these clusters and will also discuss what the results of our experiment suggest regarding the status of sonority reversal clusters in Slovak.

Pauliny claims that in two-member consonant sequences the first consonant cannot be *n*, [ɲ], *l*, [ɦ], *r, j*, [c] and [ɟ], but he does not discuss words like *ľstivý* ‘false’, *ľpieť* ‘stick to’, etc. in any detail, only says that they are ‘not of native origin’ (1979, 195). He touches upon one sonority reversal example when discussing three-member clusters: *mdloba* ‘loss of consciousness’ which he considers a loan from Czech. This observation

[6] A counterexample that immediately comes to mind is Ancient Greek, which has a small number of #TT clusters and does not allow sonority reversal clusters. Scheer says that there is a historical explanation for this. Unfortunately, we cannot test whether Ancient Greek speakers could freely borrow #*r*T clusters.

is somewhat surprising since the word is documented on the territory of Slovakia throughout the centuries (in the 17th, 18th and 19th centuries as well) according to the *Historical Dictionary of the Slovak Language* (Blanár 2005) (HDSL). (We return to sonority reversal clusters in section **4.2.4**.) This can be viewed as a claim referring to the marginality of these sequences. We can see that these highly marked sonority reversal clusters are felt to be marginal by grammarians, or else they need sophisticated extra machinery in a generative analysis to be accounted for. The opposing view is that since there is no reasonably economical way to characterize the set of attested clusters in Slovak (and Slavic in general), one has to allow everything. After giving a description of the corpus we used in the simulations presented in section **5** and as a point of reference, let us look at what native speakers' intuitions are about common, rare and absent word-initial consonant clusters in Slovak.

2.3. The corpus

The corpus we used as a point of comparison and as training data in the simulations presented in section **5** is based on ShDSL, which contains over 50,000 entries out of which we could use 29,907 onsets. There is ample phonological evidence (stress system, vowel lengthening, metrics) that a liquid trapped between two consonants in Slovak is syllabic, in traditional terms it means that it is the head of the syllable in question. In this study we decided to discard those consonant sequences which contained a syllabic liquid, so that the clusters under scrutiny are true onsets. This means that all the clusters included are followed by a vowel. The palatal and non-palatal lateral liquids are merged, and all appear as ‘*l*’.[7] Our onset corpus contains both morphologically simple (*vdova* ‘widow’) and morphologically complex forms (*v+ behnúť* ‘to run in’, in + run) as long as they are listed as lexical items in ShDSL. A shortcoming of the corpus is that it does not contain sequences which are created by a single-consonant preposition like *k* ‘to’, *s* ‘with’, *z* ‘from’, *v* ‘in’ followed by practically any other consonant or consonant cluster (except for an identical one, in those cases a vowel-final allomorph appears, e.g., *ku Katke* ‘to Cathy’). So the onset sequence [kx], for instance, does not appear in the training data in spite of the fact that this sequence is not rare in Slovak phonological words (*k* [x]*ate* ‘to the hut’, *k* [x]*lapovi* ‘to the man’, etc.).

[7] Western Slovak dialects do not distinguish the two.

3. Experiment

The goal of the present experiment was to find out whether certain existing and non-existing word-initial consonant clusters are well-formed in Slovak, or more precisely, how well-formed they are; in other words, to test the types of information that speakers use to generalize. The following 52 consonant clusters were tested:

(1) Onsets tested in the experiment
bd, bd[ʤ], bzd[r], [x]k[r], [ʧ]k, [ts]z, [ɟ]g, dl, [dz], f[ʃ], gd[r], [ɦ]d, k[x]pl, kf, kt, ktk, ktv, lk, lm, lp[r], m, md, ml, mn, ms, mst, mst[r], mz, mzd, [ɲ], nd, nm, n[r], ns, nz, pl, p[r], ps, p[ʃ]t[r], pt, pv, [r], [r]n, [r]p, s[x], [ʃ]p, st[r], t[x], tk, tkl, zbl, z[ɦ]

We included cross-linguistically unmarked onsets which are well-attested in Slovak like *m,* [r]*, p*[r]*, pl, st*[r]; typologically marked but well-attested sequences like *dl, f*[ʃ], *ml, z*[ɦ]; typologically marked onsets which are rare in Slovak, too, e.g., [dz], *bd, kt, tk*; we could not find typologically unmarked clusters that are rare in Slovak since Slovak has a fairly permissive consonant phonotactics. We also included unattested sequences some of which can be considered accidental gaps like *pt* and *kf*, others can systematically be ruled out like homorganic non-coronal clusters and near identical sequences (*pv* and [ts]*z*) or three-stop/(affricate) clusters like *ktk* and *bd*[ʤ]. A fair number of sonority reversal or sonority plateau clusters also appear in the experiment some of which are attested in Slovak (*lk, mzd*), although in a very limited number, others are missing (*lm, mst, n*[r], etc.) We were especially curious to see what native speakers' intuitions are regarding these clusters and whether there are well-formedness constraints applying to them which can be stated in terms of the segmental make up of these clusters. The clusters respect the voicing requirement on obstruents, so no voiced-voiceless or voiceless-voiced obstruent sequences are included.

We regard the inclusion of endpoint clusters, i.e., unmarked well-attested (e.g., *m, p*[r]*, st*[r]) and ill-formed unattested onsets (e.g., [ts]*z, ktk*, *bd*[ʤ]) into this kind of experiments essential since this helps speakers to anchor their ratings to numerical points on the scale provided. See Albright (2009) on the issue, too.

3.1. Method

Data collection occurred at two locations: 19 subjects took part in the experiment at the Komenský University in Bratislava (Western Slovakia) and 20 subjects participated from the University of Prešov (Eastern Slovakia). All of them were university students aged 18–23, they were all native speakers of Slovak coming from the area of data collection, i.e., Western and Eastern Slovakia, respectively. They were ignorant as to the purpose of the experiment, especially as the person collecting the data was a fluent but non-native speaker of Slovak. All the testing procedure occurred in Slovak. None of the participants reported any hearing or speaking impairment. Subjects were paid a modest sum for their participation.

3.1.1. Test words

Participants, obviously, cannot be asked to judge the well-formedness of isolated consonant clusters, they have to be presented with words. In this experiment the 52 tested onsets appeared with the following four endings: *eva*, *i*[ʃ]*a*, *obo* and *ek*[i]. These are typical sounding, well-formed but not too frequent endings, all of them occur mostly in nouns (-V*bo* in adverbs too) and none of them is a derivational suffix (they were chosen on the basis of the *Inverse Dictionary of Slovak* by Mistrík 1976). All tested onsets appeared with all four endings, which means that subjects were presented 208 test words (e.g., *meva, miša, mobo, meky, pteva, ptyša, ptobo, pteky*,[8] etc.)

3.1.2. Mode of presentation

Stimuli were presented both visually and acoustically at the same time twice in the carrier sentence:

(2) meva To je meva.
‘meva This is meva.’

Test sentences were read aloud by a trained male native speaker in his early forties, only clearly and correctly pronounced stimuli were used in the test. This means that none of the clusters were ‘repaired’ in the acoustic material by vowel insertion or cluster simplification or in any other way. The reason why stimuli were presented both visually and acoustically is as follows. If stimuli are presented only acoustically, informants might mishear the test word and rate a different cluster. This problem can be avoided if informants are asked to repeat test words, and only

[8] Slovak orthography was respected during the presentation of stimuli.

correctly repeated words are taken into account—see Albright–Hayes (2003), for instance. If test words are presented only in writing, they might be misread, perceived with a more 'native-like' consonant cluster, or vowel insertion—so again speakers might rate a different cluster. Presenting stimuli both visually and acoustically at the same time makes it very unlikely that low-probability clusters are repaired perceptually, i.e., that participants misinterpret test words and make false judgments. Test words were presented in randomized lists. After hearing a test word twice, participants had to rate it on a scale of 1 to 7 with 1 being totally ill-formed and 7 being perfectly well-formed in Slovak. The following instructions were given (in Slovak):

(3) Instructions

The task awaiting you is the following: You will hear 208 invented words, each word twice, then you will have some time to rate the word from 1 to 7, whether these words sound good in Slovak. If you think that the word sounds perfectly fine, it could be a Slovak word, then mark 7. If you think that the word sounds horrible, it's absurd, it could never be a Slovak word, then mark 1. If you think that the word sounds a bit strange, but you can still imagine it as a Slovak word, then mark, for instance, 4.

1	2	3	4	5	6	7
very bad absurd			strange but imaginable			very good sounds like Slovak

After the instructions three examples followed: *roha* (marked 7), *lrgona* (marked 1) and *sčus* (marked 5), and then came the test itself.

A reviewer raises the question whether test words could be interpreted as personal names since names frequently show other phonotactic restrictions than native words. We cannot exclude this possibility, and the possibility that test words were interpreted as brand names, for instance. Slovak names, however, usually contain unmarked, frequent onsets and speakers of Slovak are not used to encountering names with marked initial consonant clusters as most European languages have more restrictive word-initial onset phonotactics.

The fact that all test words had a perfectly well-formed Slovak-sounding "ending" might bias our results towards the top of the scale. Coleman and Pierrehumbert (1997) found that the probability of the worst part of a word is not the best score of acceptability, that is to say, the frequency of well-formed parts may ameliorate the unacceptability of the worst part. On the other hand, most subjects cannot disregard the fact that these are invented, non-existent words in Slovak, and judge even the most typical sounding test word somewhat ill-formed. Keeping this in mind, we can claim that the hierarchy, the differences we obtain between the test words must be due to their onset clusters mainly, since the rest of the word is identical in all the cases and is repeated throughout the test which probably distracts speakers' attention from the ending.

4. Results and discussion

The very first question that arises is whether results obtained in Bratislava[9] and in Prešov can be treated together or we must count with a significant dialectal difference that does not allow pooling data collected at these two locations. We assumed that differences between Eastern and Western dialects do not influence phonotactic well-formedness judgments significantly. In order to check this, we performed an h^2 test[10] for each test word and each cluster, as well as a chi-square test of homogeneity[11] for each cluster. Results confirm our predictions: there are virtually no differences between the results obtained in Bratislava and those in Prešov, so from now on *results* always refers to experimental data obtained from 38 subjects disregarding the place of data collection.

In experiments targeting phonotactic well-formedness judgments the results obtained from subjects are generally pooled and simply the averages are taken (Albright–Hayes 2003 referred to in Albright 2009; Coleman–Pierrehumbert 1997; Frisch et al. 2000), or raw ratings are scaled to the interval 0 to 1 (as in Bailey–Hahn 2001; Scholes 1966). Standard deviation in these cases is concealed, so the average for a cluster which is rated solely at 3 or 4 by speakers might be 3.5, while we can get the same result for a cluster which is rated from 1 to 7. Another problem we see is that with only 20–30 speakers one extreme rating might significantly bias the average. In order to avoid the latter problem we could consider the median instead of the mean, but as the median is always an integer this would make it more difficult to set up a hierarchy between the clusters under scrutiny as we would get many clusters with the same

[9] We had to discard one test due to some technical error, so we were left with 18 subjects from Bratislava.

[10] The score for h^2 was 0.10–0.16 in 9 out of 208 test words, in the rest we got an even lower value, which means that location and rating are completely independent of each other.

[11] In none of the 52 clusters did we get a significant difference between the results from the two locations.

median, while the mean lets us posit a finer hierarchy. Although keeping in mind the doubts mentioned, we will also use the mean in our computations, as customary, and return to the question of dispersion in the discussion part.

We also checked whether endings influenced speakers' judgments significantly. Test words ending in *-eva* were ranked 5.2% above average, those ending in *-ek*[i] 4.5% above average, words ending in *-i*[ʃ]*a* were ranked 1.8% below average and those ending in *-obo* were ranked 8% below average. An ANOVA on mean ratings with onset and ending as factors shows a significant interaction ($p < .001$). This is not surprising, though, since even if one particular item has a deviant rating it might yield a significant interaction. (Remember that we had 208 test words, 7904 answers). Therefore, we checked effect size as well, and as we expected, the effect of onset is huge ($F(51, 7684) = 156.96, \eta^2 = 0.51$, which is a large effect) relative to that of ending ($F(3, 7684) = 40, \eta^2 = 0.015$) and the interaction of onset and ending ($F(153, 7684) = 3, 11, \eta^2 = 0.058$ which does not even reach medium size effect).[12] A Kruskal–Wallis post hoc test revealed that the ratings of *-eva* and *-ek*[i] words did not differ significantly from each other, while the other two endings differed from each other as well as from *-eva* and *-ek*[i]. Results obtained from *-eva* and *-ek*[i] words only, however, do not differ in any meaningful way from those coming from all the test words. Therefore, *results* will refer to all the test words from now on.

Below we present the ratings in decreasing order. The first column contains the average ratings from 38 speakers of the 52 onsets tested on a 1–7 scale. The second column shows the averages of every single test word, the third column contains the type frequency of the onset cluster in ShDSL. These are the counts in the corpus of training data that was used for the simulations presented in section **5**. The fourth column shows the scores assigned by the Hayes–Wilson Phonotactic Learner to the tested onsets and the fifth column contains the scores from the Generalized Neighborhood Model simulation. Note that the learners do not work on a 1 to 7 scale. The score assigned by HWPL range between 0–1 and those by GNM between 0.011–0.239 in this particular case.

[12] With scores statistically normalized in order to minimize the differences between the four endings, we get very similar results: the effect of onset is $F(51, 7904) = 149, \eta^2 = 0.503$ and the effect of the ineraction of onset and ending is $F(3, 7904) = 0.066, \eta^2 = 0.056$.

(4) Ratings of the top 25% of the tested onset clusters

Overall average		Word averages	Corpus	HWPL	GNM
[r]	**6.480**	reky 6.21, reva 6.342, riša 6.526, robo 6.842	1698	1	0.239
m	**6.139**	meky 6.5, meva 5.5 miša 6.947, mobo 5.594	1185	1	0.218
pl	**5.900**	pleky 5.736, pleva 6.605, pliša 5.54, plobo 5.702	269	1	0.163
st[r]	**5.816**	streky 6.552, streva 5.737, striša 5.579, strobo 5.394	179	1	0.044
[ɲ]	**5.691**	neky 5.526, neva 6.684, niša 5.895, ňobo 4.658	740	1	0.203
p[r]	**5.421**	preky 5.842, preva 5.734, priša 5, probo 5.105	2069	1	0.18
[ʃ]p	**5.263**	špeky 6.684, špeva 5.394, špiša 4.921, špobo 4.052	100	1	0.091
[dz]	**5.252**	dzeky 5.263, dzeva 5.921, dziša 4.757, dzobo 5.052	1	0.487	0.182
dl	**5.166**	dleky 5.026, dleva 5.378, dliša 4.842, dlobo 5.421	16	0.662	0.101
ml	**5.132**	mleky 5.71, mleva 5.526, mliša 4.5, mlobo 4.789	60	0.788	0.101
zbl	**4.750**	zbleky 5.237, zbleva 4.868, zbliša 5.158, zblobo 3.737	7	0.711	0.039
ps	**4.355**	pseky 4.421, pseva 4.658, psiša 4.053, psobo 4.289	29	0.505	0.098
s[x]	**4.132**	scheky 3.421, scheva 4.184, schyša 4.71, schobo 4.21	44	1	0.074

(5) Ratings of the second 25% of the tested onset clusters

Overall average		Word averages	Corpus	HWPL	GNM
z[ɦ]	**4.072**	zheky 4.132, zheva 3.71, zhyša 4.421, zhobo 4.026	52	0.787	0.075
mn	**4.013**	mneky 4.184, mneva 4.052, mnyša 4.052, mnobo 3.763	27	0.812	0.083
kt	**3.842**	kteky 3.579, kteva 3.71, ktyša 3.789, ktobo 4.289	12	0.642	0.077
f[ʃ]	**3.711**	fšeky 4.368, fševa 4.237, fšiša 3.184, fšobo 3.052	43	0.741	0.068
pt	**3.711**	pteky 4.321, pteva 3.71, ptyša 4, ptobo 2.789	0	0.31	0.097
bd	**3.645**	bdeky 4, bdeva 3.473 , bdyša 3.921, bdobo 3.184	1	0.492	0.076
p[ʃ]tr	**3.526**	pštreky 4.5, pštreva 3.21, pštriša 2.921, pštrobo 3.473	1	0.165	0.014
mzd	**3.395**	mzdeky 3.447, mzdeva 3.395, mzdyša 3.316, mzdobo 3.421	1	0.312	0.024
mz	**3.349**	mzeky 3.816, mzeva 3.158, mziša 3.71, mzobo 2.71	0	0.439	0.079
bzd[r]	**3.257**	bzdreky 3.605, bzdreva 3.421, bzdriša 3.5, bzdrobo 2.5	0	0.202	0.011
[tʃ]k	**3.166**	čkeky 3.184, čkeva 3.316, čkyša 3.131, čkobo 3.027	1	0.5	0.069
t[x]	**3.158**	tcheky 2.842, tcheva 3.105, tchiša 3.368, tchobo 3.316	1	0.489	0.065
[ɦ]d	**2.875**	hdeky 2.895, hdeva 3.237, hdyša 2.71, hdobo 2.658	0	0.337	0.067

(6) Ratings of the third 25% of the tested onset clusters

Overall average		Word averages	Corpus	HWPL	GNM
md	**2.862**	mdeky 3.052, mdeva 3.289, mdyša 2.5, mdobo 2.605	0	0.458	0.075
gd[r]	**2.849**	gdreky 2.658, gdreva 3.658, gdriša 2.658, gdrobo 2.421	0	0.227	0.032
nr	**2.803**	nreky 2.815, nreva 3.158, nriša 2.342, nrobo 2.894	0	0.309	0.098
tk	**2.757**	tkeky 2.5, tkeva 3.184, tkyša 2.842, tkobo 2.5	8	0.595	0.075
rn	**2.684**	rneky 2.921, rneva 3.105, rnyša 2.631, rnobo 2.079	0	0.381	0.078
mst	**2.678**	msteky 3.421, msteva 2.79, mstyša 2.421, mstobo 2.079	0	0.235	0.025
tkl	**2.664**	tkleky 2.473, tkleva 3.526, tkliša 2.421, tklobo 2.237	1	0.429	0.036
pv	**2.618**	pveky 2.921, pveva 2.579, pviša 2.816, pvobo 2.158	0	0.331	0.107
kf	**2.539**	kfeky 2.368, kfeva 2.579, kfiša 2.737, kfobo 2.473	0	0.284	0.069
lm	**2.539**	lmeky 2.421, lmeva 2.552, lmiša 3, lmobo 2.184	0	0.349	0.078
mst[r]	**2.477**	mstreky 2.783, mstreva 2.552, mstriša 2.368, mstrobo 2.21	0	0.235	0.012
ms	**2.454**	mseky 3.105, mseva 2.421, msiša 2.184, msobo 2.105	0	0.258	0.075
ktv	**2.205**	ktveky 2.21, ktveva 2.405, ktviša 2.473, ktvobo 1.737	0	0.431	0.027

(7) Ratings of the bottom 25% of the tested onset clusters

Overall average		Word averages	Corpus	HWPL	GNM
nz	**2.204**	nzeky 2, nzeva 1.894, nziša 2.579, nzobo 2.342	0	0.135	0.076
nd	**2.158**	ndeky 2.263, ndeva 3.263, ndyša 2.026, ndobo 2.079	0	0.174	0.073
lk	**1.98**	lkeky 2.105, lkeva 1.947, lkyša 2.026, lkobo 1.842	1	0.188	0.078
nm	**1.961**	nmeky 2.079, nmeva 2.053, nmša 2.053, nmobo 1.658	0	0.225	0.078
ns	**1.961**	nseky 2.21, nseva 2.026, nsiša 1.632, nsobo 1.973	0	0.069	0.073
[r]p	**1.921**	rpeky 2.19, rpeva 1.842, rpiša 1.71, rpobo 1.947	0	0.225	0.094
[x]k[r]	**1.854**	chkreky 2.27, chkreva 1.71, chkriša 1.789, chkrobo 1.658	0	0.286	0.036
[ɟ]g	**1.671**	ďgeky 1.263, ďgeva 1.816, ďgyša 1.579, ďgobo 2.026	0	0.191	0.058
lpr	**1.645**	lpreky 1.868, lpreva 1.868, lpriša 1.526, lprobo 1.316	0	0.225	0.047
[ts]z	**1.625**	czeky 1.579, czeva 1.868, cziša 1.421, czobo 1.632	0	0.069	0.068
k[x]pl	**1.454**	kchpleky 1.395, kchpleva 1.868, kchpliša 1.342, kchplobo 1.21	0	0.072	0.012
ktk	**1.344**	ktkeky 1.316, ktkeva 1.447, ktkyša 1.289, ktkobo 1.324	0	0.209	0.02
bd[ʤ]	**1.219**	bdžeky 1.184, bdževa 1.324, bdžiša 1.289, bdžobo 1.079	0	0.032	0.012

In the remaining part of this section we will discuss tested consonant sequences by type, paying special attention to sonority reversal sequences and will use the arbitrary labels 'well accepted' for the top 25%, 'accepted' for the second quarter, 'on the verge of acceptability' for the third 25% and 'rejected' for the bottom quarter.

4.1. Singletons

The following singletons appeared in the experiment: [r], *m*, [ɲ] and [dz]. The first two are typical end-point clusters whose type frequency is high (within the first ten according to ShDSL),[13] and as expected, speakers rank them high; these are the only two onsets that have a score over 6. [ɲ] figured in the experiment because palatals are cross-linguistically less frequent but in Slovak [ɲ] is a frequent word-initial onset as the negation of verbs occurs by adding the prefix *ne-* to the verb in question (*dať* 'give' *nedať* 'not to give'), and is ranked high in our experiment, too. [dz] on the other hand, occurs once in ShDSL word-initially and it is in the word *zeta*, the name of the Greek letter, that is to say, although [dz] is an existing Slovak consonant it does not occur word-initially in the native vocabulary.[14] As we can see in (4), speakers feel that the lack of word-initial [dz] in Slovak is a lexical accident, it is well accepted.[15] The following questions arise, though: why do people rank [r]-initial words higher than *m*-initial (or *p*[r]-initial words) and why do these words not receive a top score? It seems that people are reluctant to give non-words a top score even if they are perfectly well-formed with a frequent, canonical pattern.[16] It could be illuminating to make speakers rank existing words according to how "typical sounding" they are. We might receive the same hierarchy as with non-words but with higher actual scores. A possible explanation is that [r] is a more frequent word-initial onset in Slovak than *m*, which would be a clearly lexical effect and tells us nothing about the onset grammar of Slovak.

4.2. Two (and three)-consonant clusters

4.2.1. Rising sonority sequences

Stop + (non-homorganic) liquid sequences are generally considered the most typical, least marked two-consonant onset clusters. This is borne

[13] The type frequency of initial onset clusters in ShDSL, which was used as training data in the modeling experiment as well, can be found in the Appendix.

[14] There are dialects where a *d* before *e* and *i* palatalizes to [dz] rather than [ɟ].

[15] Probably because common last names start with *dz* (e.g., *Dzurinda*, former prime minister). In Eastern Slovak it replaces [ɟ].

[16] Two test words *robo* and *miša* can actually be nicknames in Slovak; they were among the highest ranked words together with *pleva, meky, špeky, riša* and *reva.*

out in Slovak as well. Both *p*[r] and *pl* have high type frequency and are ranked high by speakers. We do not have an answer to why *p*[r] is ranked lower by our speakers than *pl* (since *p*[r] is a more frequent initial onset), it might just be a random effect, a task related phenomenon inevitable in non-word testing. The difference between the two clusters is statistically not significant (we performed independent t-tests).

Contrary to most Indo-European languages, in Slovak no true anti-homorganicity requirement is observable on word-initial coronal clusters, *tl* and *dl* are attested—with a type frequency of 26 and 16, respectively, in ShDSL—and well accepted onsets, which is mirrored by our speakers' judgments: *dl* scored 5.166. We do not find labial onset sequences word-initially in Slovak, though. Pauliny simply lists a handful of stop + sonorant clusters, with labial clusters among them, which do not occur in his corpus, but does not give a principled explanation for their absence. We think that a sequence of homorganic non-coronal consonants in the onset are disfavoured in Slovak. According to our informants the unattested *pv* is on the verge of acceptability (2.618). It violates an OCP-place constraint, at the same time it might get some support from other attested stop + sonorant sequences in the lexicon like *pl, pn, tv, kv*, etc., which are well-formed word-initial clusters in Slovak, on the one hand, as well as from homorganic *tl* and *dl*, on the other.

4.2.2. Obstruent + obstruent sequences

The most common O + O onset clusters in Slovak, in accordance with universal patterns, are sibilant + stop sequences. Note that *s/z* is a preposition and a verbal prefix in this language and as such it can form a cluster with almost any consonant (except *s, z* and [ʃ], [ʒ]). Here we refer to sibilant + stop + [r]/*l* sequences, too, as these are cross-linguistically the most frequent three-consonant word-initial clusters and they are decomposible to the unmarked sibilant + stop and stop + liquid clusters. These sequences are ranked high by speakers, with the voiced sequence *zbl* receiving a somewhat lower score than voiceless *s* + stop clusters (*st*[r] and [ʃ]*p*).

Sibilant + fricative clusters are typologically more marked but they are well-attested in Slovak and the tested onsets (*s*[x] and *z*[ɦ]) fall into the 'well accepted' and 'accepted' region in accordance with the universal implicational hierarchy saying that the existence of fricative + fricative sequences in a language implies the existence of fricative + stop clusters (see Morelli 1999 and the references therein). The voiced sequence receives a somewhat lower score despite the fact that it has more occurrences in the corpus. The relative markedness of voiced fricatives as opposed to voiceless ones can be viewed as a grammatical effect and is aerodynamically motivated because the production of friction noise and voicing involves conflicting aerodynamic requirements (see Shadle 1985 and Stevens et al. 1992 for more details). The longer the fricative (cluster) the more difficult its realization is compared to the voiceless fricative (cluster).

If the first fricative is not a sibilant, the judgment of the cluster deteriorates in accordance with cross-linguistic tendencies. (Similarly to *s/z, f/v* is also a preposition and a verbal prefix in Slovak, so the type frequency of *f/v* + C clusters is high in the language.) The attested fricative + fricative (*f*[ʃ]), the attested affricate + stop ([tʃ]*k*) and the unattested fricative + stop ([ɦ]*d*) clusters were accepted by speakers with the voiced cluster receiving somewhat lower score. In this case we do not know to what extent this result is due to the voicing or to the unattestedness of the latter cluster. As for three-member fricative-initial clusters we tested [x]*k*[r] and it was rejected (1.854). This cluster is decomposable into the unattested (and untested) but acceptable [x]*k* cluster and the attested and canonical *k*[r]. Why is it still rejected? As it has been mentioned in **4.2.1**, in Slovak the canonical sibilant + stop + sonorant sequences are well-attesed word-initially. In case the first fricative is a non-coronal sibilant it must be voiceless (e.g., [ʃ]*k*[r]-) in all other cases, i.e., the sequence starts with a different voiced fricative (e.g., *vd*[r]-) or the second member is not a stop (e.g., *fsp-*) the form is morphologically complex and contains a word-initial prefix *s/z* or *f/v.* So there are no three-consonant clusters in Slovak like *[ʒ]*d*[r]-. Speakers when assessing the well-formedness of certain forms rely on their knowledge of morphology as well (a grammatical effect again) which in this case tells them that the word-initial sequence [x]*k*[r] should be ill-formed in Slovak.

In this test we included the following attested stop + stop clusters:—the number in brackets shows the type frequency of the cluster in the training corpus—*kt* (12), *tk* (8) and *bd* (1); and we also included the unattested onset *pt.* Interestingly, speakers did not rate these clusters in accordance with their type-frequency or whether they are attested or unattested, but according to their "well-formedness" on perceptual grounds. Although this point is in need of further research, it seems that a non-coronal stop followed by a coronal is a "better" cluster than a coronal stop followed by a non-coronal. There is evidence that labial closure may precede coronal closure, hiding the coronal gesture and resulting in a labial only percept (Browman–Goldstein 1990). We assume that this finding can be extended to velars as well. Byrd (1996) shows that gestural overlap in coronal-labial stop clusters is greater than in labial-coronal clusters, and Byrd (1992) used articulatory synthesis to show that a completely articulated alveolar stop is not perceived by listeners if it is substantially overlapped by a velar stop. In our study the sequences *kt, pt* and *bd* have the average ranking of 3.842–3.645 and fall into the 'accepted' range, while *tk* gets only 2.757 points being on the verge of acceptability.

It is generally assumed (see, e.g., Morelli 1999) that the existence of stop + stop clusters in a language implies the existence of stop + fricative clusters. This is also borne out in Slovak since the latter are slightly more frequent in the language than stop + stop clusters, with *ps* (29) being the "best" and *t*[x] (1) the "worst". Note that except for *t*[x] C_2 in all such attested clusters is a sibilant, so we might assume a constraint in Slovak which demands the second member of such clusters to be a sibilant. The experiment contained the attested sequences *ps* and *t*[x] as well as the unattested sequence *kf*. Speakers consider *ps* (4.355) well-acceptable, *t*[x] is accepted, while *kf* (2.539) is on the verge. In light of these results we should say that the fewer the tokens on the basis of which speakers could generalize the more the importance of the actual occurrences in the lexicon grows.

In this section we must mention another absolutely ill-formed cluster in the experiment: [ɟ]*g* which was rejected by speakers (1.671). Slovak does not allow palatal + C sequences monomorphemically. There are some morphologically complex forms like *Pe*[c]*o* 'Pete' → *Pe*[c]*k*o 'Pete diminutive', but word-initial clusters with a C_1 palatal are excluded. This has a historical explanation and seems to be an active synchronic constraint in the language. Coronal palatals in Slovak are the result of palatalization by a palatal vowel (*e, i* or [j] in diphthongs). At the same time when yer-dropping could have created such clusters a depalatalization occurred in these clusters (there are a few exceptions with *l*, for example *ľnúť* all mentioned in the Introduction). Remember that a lone [ɲ] was well accepted by our speakers.

4.2.3. *m* + C clusters

We discuss *m* initial onsets separately, not within the group of sonority reversal clusters because they seem to pattern with obstruent-initial clusters rather than sonorant-initial clusters. This asymmetrical behavior of *m* is not unique to Slovak or Western Slavic languages—see, for instance, Rowicka (1999) on Polish. Van der Torre (2003) gives evidence that Dutch *m* patterns with obstruents. He gives a Government Phonological representation to explain why *m* is more consonantal, i.e., is a stronger licenser than *n*. It is also observed—by Botma and Smith (2007) on the basis of Maddieson (1984)—that voiceless nasals are more likely to be bilabial, or in implicational terms, if a language has a voiceless coronal or velar nasal, it will also have a voiceless labial nasal. Maddieson (1984) reports a similar pattern for breathy voiced nasals. Zuraw (2007, 281) anchors these observations to phonetic facts. She says that "[i]t can be argued that [n] is more vowel-like than [m] because nasal antiformants that might interfere with vowel-like formant structure are higher for [n] than for [m]". Zuraw talks about nasals in pre-vocalic postion which is not the case here, so the point is in need of further acoustic analysis.

Returning to *m*-initial clusters in the present experiment, if for the moment we assume that *m* is an obstruent, the results reflect universal tendencies: the *m* + liquid cluster (*ml*) is ranked highest (5.132), the *m* + nasal cluster (*mn*) has less counts in the corpus and gets a lower score; *m* + voiced obstruent clusters follow: *mzd, mz* and *md* with the last one on the verge of acceptability. Note that only *mzd* is actually attested in Slovak with a type frequency of 1. Those clusters in which the obstruent does not agree with *m* in voicing (*ms*, *mst* and *mst*[r]) are ranked even lower (2.678–2.477) still being on the verge of acceptability. None of these initial clusters is attested in our corpus, although, some corpora may contain the word *msta* 'vengeance' that is very obsolete and dialectal, substituted in Modern Slovak by the form *pomsta.* The four-member cluster *mst*[r] is discussed in section **4.3**. These results suggest that speakers make their decisions on grammatical grounds.

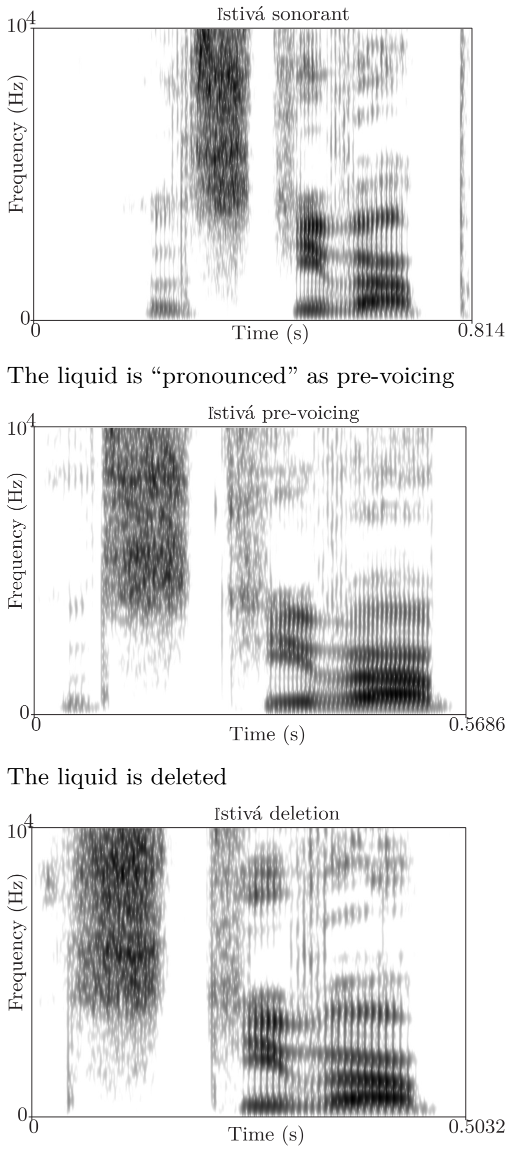
4.2.4. Sonority reversal clusters

In this section we will discuss sonority reversal clusters and sonority plateau clusters consisting of two sonorants, except for *m* + C clusters, which are dealt with in section **4.2.3**. Sonority plateau clusters starting with a coronal sonorant (*nr, rn*) are ranked quite low (2.803–2.684), they are on the verge of acceptability. The only sonority reversal cluster in this group is *lm* as long as we classify *m* as an obstruent. None of these clusters exists in the language. True sonority reversal clusters (*nz, nd, lk, nm, ns,* [r]*p* and *lp*[r]) are rejected by speakers (2.204–1.645). Recall that the only clusters of this type attested in Slovak and included in the test material are *lk* and *lp*. These results suggest that not only forms that are perfectly well-formed can exist in a language. Forms that are more offending, forms that native speakers reject as possible words in their language can persist (for instance, due to 'historical accidents' like the creation of clusters through the loss of yers). These combinations, however, are more unstable, more likely to change: *ľnúť* 'stick to' and *ľpieť* 'stick to' are in modern colloquial Slovak substituted by *lipnúť* 'stick to'; *mša* 'mass' is used as *omša, msta* 'vengeance' became *pomsta* (the prefix was borrowed from the perfective verbal form *pomstiť* 'to avenge'), *rmútiť* 'to grieve' disappeared and *smútiť* 'to grieve' took its place. Further examples from the history of Slovak are *rvať* 'to fight' which became *ruvať*; *mgla* 'fog' through the form *mlha* (which could have survived with a syllabic *l*) evolved with a metathesis into *hmla*; although the form *mdloba* 'loss of consciousness' persists in the language, the verbal pair *mdlieť* ∼ *omdlievať* 'to faint perf. ∼ imperf.' became *omdlieť* ∼ *omdlievať*. The word *ľstivý* 'false' is documented from the 17th century on and is an attested word to date. The cluster #*l*[ʒ] appears three times in our corpus in the words *lživý* 'mendacious', *lžidemokracia* 'false democracy' and *lžimorálka* 'false moral'. According to HDSL *lživý* is well documented from the 15th century on. This seems to be the 'best' sonority reversal cluster in Slovak. ShDSL considers the form *lkať* 'cry' poetic, according to HDSL it is a Czech word documented only in the 18th century. What is the actual realization of *l* in sonority reversal clusters? We could not find any reliable phonetic accounts of the realization of *l* in this position. We suspect that in spontaneous speech on those rare occasions when these words occur in an utterance-initial position, it varies to a great extent. The spectrograms in (8), overleaf, show the pronunciation of the word *ľstivá* 'false.fem.' by the same speaker in utterance-initial position. In (8a) the liquid is pronounced as a true sonorant, in (8b) a short prevoicing is observable before the sibilant, the traces of an *l* gesture which is practically unperceivable. In (8c) we can see a realization where the violating liquid is deleted. The last cluster of this type in our corpus is *rmut* 'fruit pomace', a technical term according to ShDSL, which does not appear in HDSL.

4.3. Three and four-consonant onsets

The only three-member clusters that have not been discussed in the previous sections are *ktv*, *tkl*, *gd*[r], *ktk* and *bd*[ʤ]. The latter two are ruled out

(8) The pronunciation of *ľstivá* ‘false.fem.’ in utterance-initial position



(a) The liquid is pronounced as a sonorant

(b) The liquid is “pronounced” as pre-voicing

(c) The liquid is deleted

by the onset grammar of Slovak which does not allow three-obstruent clusters and as expected are unanimously rejected by speakers. The standard deviation in the case of *ktk* is 0.703, for *bd*[ʤ] 0.576, while for two-member obstruent clusters it is much higher (*kt* 1.583). This is because their judgment is almost categorical, they get the score of 1 or 2. The unattested cluster *ktv* is decomposable to *kt* (12) and *tv* (24) and also reinforced by the attested clusters *tkl* (1) and *tkv* (1). Speakers consider both unattested *ktv* and attested *tkl* being on the verge of acceptability, with *tkl* ranked somewhat higher. Note, however, that unattested cluster *gd*[r] (2.849) is rated higher than attested *tkl* which is probably due to *d*[r] being an absolutely canonical word-initial sequence (*gd* is not attested but acceptable in this position).

All four-consonant onsets in Slovak fit into the labial + sibilant + stop + sonorant template. The vast majority of these words contain the prefixes *v*/[f] + *s*, *z* or the complex prefix *vz*/[fs], which is followed by a 'regular' branching onset. Two exceptions are *pstruh* 'trout' and *p*[ʃ]*tros* 'ostrich', which do not start with a prefix but also conform to the above mentioned template. Speakers indeed rated non-existent *bzd*[r] which fits into the four-member cluster template considerably higher (3.257) than the unattested and non-fitting sequence *k*[x]*pl* (1.454), *mst*[r] (2.477) lies on the verge of acceptability as it conforms to the four-consonant template but contains a quasi voicing violation—see **4.2.3**.

4.4. Summary

In this section we presented experimental data from 38 native speakers of Slovak on the acceptability of word-initial onset clusters. Those cases are especially interesting which are cross-linguistically marked but attested in Slovak, or else which are absent even in Slovak. We saw that in some cases speakers' judgments even of unattested forms can be captured by grammatical constraints (which are based on phonetic facts), like in the case of stop + stop sequences or fricative-initial clusters. In other cases morphology clearly plays a role (three-member fricative-initial clusters). Yet in other cases it seems that speakers' judgments are based on what is attested in the lexicon (stop + fricative clusters). These two "forces" (grammar and lexicon) are, naturally, closely related since phonotactically unmarked forms generally outnumber marked ones. It is surprising that speakers rank [dz] so high. Our findings on sonority reversal clusters are especially interesting because they suggest that Slovak "happens to have" these highly offending clusters, but is slowly "repairing" them, that is to say, they exist in the language, but speakers reject them. Let us now proceed to the modeling of these results.

5. Modeling experimental data

Much research focuses on the issue of gradient intuitions in generative linguistics (see, for instance, Hayes–Wilson 2008 and the references therein). The question is not whether we find gradience because we do. Even where only binary responses are allowed, when averaged across speakers, gradient well-formedness judgments are achieved. The question is how this relates to the phonotactic grammar.

We can still maintain the idea that the grammar is categorical: gradience comes from performance, and it is a task-related phenomenon. Grammar describes what is possible and what is not, whereas in a blick test—when speakers are tested on the well-formedness of nonce words—speakers rather judge what is probable, or more precisely how probable a sequence is. (Schütze 1996 discusses the issue in detail.) At the other extreme of this range we find the approach according to which there is no grammar. The acceptability of novel words relies solely on the lexicon, that is, on support from existing words (e.g., Greenberg–Jenkins 1964; Ohala–Ohala 1986). Speakers try to recognize novel words as real words, and their judgments depend on how much support these words get from the lexicon, so their intuitions are extracted from the data only. It is very difficult (if not impossible) to experimentally differentiate between these two views, since decision-making in both cases relies exclusively on the lexicon.

It is also plausible that grammar itself is gradient (e.g., Coleman–Pierrehumbert 1997; Frisch et al. 2000; Albright–Hayes 2003, etc.). So gradient grammaticality judgments reflect gradient grammaticality intuitions, that is to say, a gradient grammar. There is extensive literature on both the "lexicon-only" (second approach) and "phonotactic knowledge" (third approach) models, but few works contrast the two. Bailey and Hahn (2001) explicitly contrast the two approaches and conclude that lexical influences play a bigger role in sequence typicality than phonotactic probabilities. Albright (2006), on the other hand, concludes after comparing two lexical, three sequential and a "hybrid" model on two sets of phonotactic well-formedness judgment data, those of Bailey–Hahn (2001) and Albright–Hayes (2003), that "gradient acceptability reflects knowledge about the relative probability of different combinations of natural classes, not knowledge of words directly" (Albright 2006, 12).

The models that are compared to experimental data in the present paper are the Hayes–Wilson Phonotactic Learner (Hayes–Wilson 2008), which is a "phonotactic model", and the Generalized Neighborhood Model[17] by Bailey and Hahn, which is a "lexical model". We chose these two models to contrast our experimental data with because they are both fairly recent, well-elaborated models, which were available to us, they are both designed to model gradient phonotactic intuitions, but represent opposing approaches to gradient phonological judgments: the Hayes–Wilson learner is a phonotactic model, while GNM is a lexicon-only model. Both models use features and natural classes defined by those features. We fed both models with the same segment inventory and feature matrix.[18] The feature matrix we used is based on Pauliny (1979) and Rubach (1993) with smaller changes. The most important change, and the least standard feature matrix, is provided for [v]. It is adopted from Padgett (2002) who introduces the feature [wide]—relying on phonetic facts—to explain the special behavior of *v* in Russian. This segment shows a similar double-faced phonological behavior in Slovak as well (for details refer to Bárkányi–Kiss 2010) so we assume that it is best analyzed as [+son, −wide] in this language, too. The total number of natural classes is controlled for by using both contrastive and privative underspecification as in Hayes–Wilson (2008).

We, obviously, do not want to claim that these are definitely the best or only available algorithms suitable for modeling phonotactic grammars. We leave it open for further research to try other models on our data.

5.1. The Hayes–Wilson Phonotactic Learner

We will not go into (mathematical or computational) details about the models; we just give a brief summary here. The interested reader is referred to the original papers. HWPL comes up with grammars that are composed of numerically weighted constraints, and the well-formedness of an output is formalized as a probability determined by the weighted sum of its constraint violations. The constraints are selected according to an inductive constraint-finding algorithm. Constraints are free to refer to all the featural, structural and other distinctions made by the representations, and thus permit multiple overlapping characterizations of

[17] An implemented version of the latter was kindly provided to us by Adam Albright.

[18] The training data and the feature matrix used in the simulations can be found in the appendix, together with the 25 most important constraints that HWPL comes up with (note that this is not the order of discovery).

phonological forms, so it is the natural classes determined by the features, rather than the features themselves that determine the content of a constraint. The learner only learns Markedness constraints; inputs do not play a role. The system selects constraints in order of generality. Shorter constraints are treated as more general. There is a trade-off between the size of the constraint (which is usually set to refer to 2 or 3 natural classes) and specificity. In this study the learner was set to learn a grammar with 200 constraints and refer to 3 natural classes. The learner makes use of contrastive underspecification and is allowed to refer to the complementation operator as well. Contrary to Hayes and Wilson (2008), though, we found that it did a better job on the Slovak data without referring to complement classes. The procedure of constraint weighting is an iterated hill-climbing search, designed to maximize the probability of the learning data. It is computed according to the principle of maximum entropy. The search is determined at each stage by calculating a local gradient based on the difference between the observed value and the expected value for each constraint. The observed value is determined by inspection of the learning data, while the expected value is calculated based on the grammar already learnt. The complete process of learning alternates between constraint selection and constraint weighting. When a new constraint is selected, all constraints are reweighted. The learning algorithm ends when the search fails to return a new constraint at the least stringent accuracy level, or when the grammar reaches the stipulated grammar size (as in our case). Since constraint selection is stochastic in this model, the learner learns a slightly different grammar at each run. We ran the learner five times with these settings (and present here one as an example), but the predictions it made when tested with the onsets from the experiment was basically identical in all the cases.

5.2. The Generalized Neighborhood Model

The standard measure of sequence typicality in lexical models is based on the notion of lexical neighborhoods, which relies on the single phoneme edit distance (Luce 1986). By this metric, a neighbor is any word that can be derived by substituting, deleting, or inserting a single phoneme, and the number of such neighbors is the neighborhood density of an item. Although this is a crude approximation, it has been successful in the study of lexical processing. This procedure, however, fails to take similarity between phonemes into account. Furthermore, it has a sharp cutoff, simply ignoring all words outside the single phoneme edit distance.

In order to overcome these problems, Bailey and Hahn adapt Nosofsky's (1986) Generalized Context Model (GCM) in which neighbors vary on a continuous scale of similarity. Instead of imposing a sharp distinction between neighbors and non-neighbors, all relevant items are neighbors to some degree, and the model categorizes novel exemplars based on their aggregate similarity to sets of existing exemplars. In GNM the degree of lexical support for a novel word is proportional to the weighted sum of the perceptual similarities of the novel word to each existing word. The distance between novel and existing words is calculated by finding their minimum string edit distance, which involves finding the optimal alignment between the segments of the novel word and the existing word, combining the best pairing of phonetically similar segments and the fewest possible insertions and deletions. Phonetic similarity is assessed with a metric based on natural classes (Frisch et al. 2004); the model counts the number of shared and unshared natural classes of two phonemes. The value it returns ranges from 0 for identical phonemes to 1 for phonemes that have no features in common, so are totally dissimilar. This variant of GNM gives no credit for shared phonemes in different word positions (so *stick* and *ticks* are no more similar than *trick* and *ticks*).

The relative cost of insertions and deletions compared to substitutions is determined empirically by post hoc fitting. The authors claim that several costs were computed and 0.7 was chosen because it gave the best fit to the oral word-likeness data they gathered. This parameter was left untouched in our simulations, although the task of matching entire words is different from this more restricted task of matching only consonant clusters. If this value is too low, even quite dissimilar consonants may get support from each other. This could be the reason why [dz] is ranked so high by GNM—see section **5.3**. On the other hand, if the insertion/deletion cost is too high, it is difficult for long clusters to get support from shorter ones (e.g., *str* from *st, tr*). This might be a reason of why three- and four-consonant clusters were underestimated in the simulation.

As token frequency has been found important in a wide range of lexical tasks, a frequency-weighing term is incorporated into the similarity equation as a quadratic function of the log token frequency of a given word, in this way GNM is able to capture both monotonic and non-monotonic frequency effects. The authors claim that the influence of a lexical neighbor is an inverted-U-shaped function of its token frequency. This means that most frequent and very infrequent forms contribute less, while middle frequency forms contribute most to determining similarity. The exact shape of the frequency effect must be determined by post hoc fitting. This, however, is not used in the present study for two reasons. Firstly, we could not find a good frequency dictionary or database of Slovak; secondly, Hayes and Wilson (2008), as well as Albright (2009), achieved better results in modeling phonotactic judgments if token frequency was not taken into account.

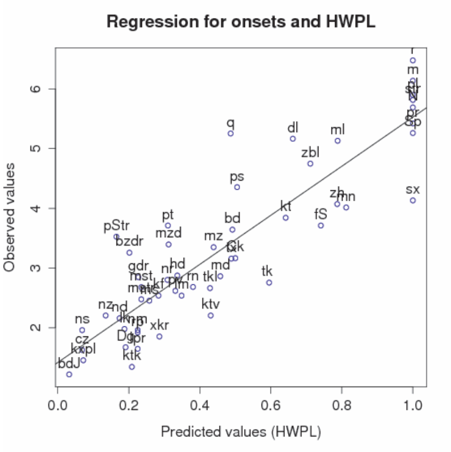
GNM has several free parameters, but the authors claim that "the extra complexity of the GNM is justified to the extent that it provides a better explanation for empirical data" (Bailey–Hahn 2001, 573). This makes the model somewhat difficult to test on empirical data, and with the use of "default" settings the model does not necessarily return the best-case performance, which might be a worry in this case, too. The coefficients A, B and C that are used in the quadratic equation for taking token frequency into account are turned off in the version we used (A and B are set to 0 and C is 1). The coefficient D is basically responsible for the "sensitivity" of the model, meaning how quickly the influence of less similar items drops off. In several runs, we found that the best-fit value of this parameter for our data was 2.

5.3. Discussion of modeling results

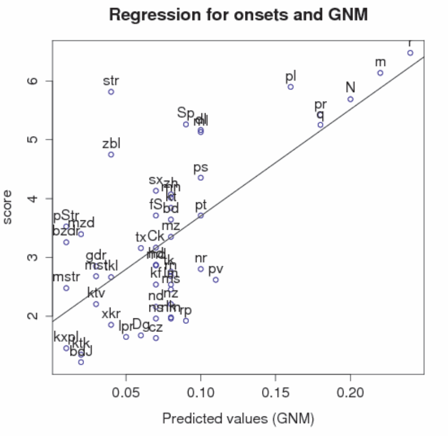
In (9) we show the performance of HWPL when contrasted with the results from our experiment on Slovak onsets: $r = .886$, $r^2 = 785$. In (10) we can see the performance of GNM: $r = .682$, $r^2 = 466$.

It is noteworthy that the strongest constraint referring to initial consonant sequences in Slovak that the Hayes–Wilson Phonotactic Learner comes up with is the prohibition on coronal sonorants as first members in a consonant cluster despite the 8 observed forms in the training data, which is in accordance with speakers' intuition about sonority reversal clusters. HWPL, however, assigns 1 to all canonical clusters and does not capture the differences between unmarked clusters we observed in people's judgments. This weakness of the model is due to the fact that it is biased towards generality, and it is compensated with its ability to perform fairly well on unattested and low type-frequency clusters. This

(9) Performance of the Hayes–Wilson Phonotactic Learner in predicting experimental data on Slovak onsets[19]



(10) Performance of GNM in predicting experimental data on Slovak onsets



learner seems to overestimate fricative-initial clusters. We found this out by fitting the model's predictions to subjects' ratings and calculated the residuals. We have no explanation for this. It does not discover the four-consonant "template" either, this is not so surprising, though, since the model was set to discover constraints of three natural classes. The fact

[19] Special characters used in the regression charts: "S" stands for [ʃ], "C" for [ʧ], "c" for [ʦ], "J" for [ʤ], "D" for [ɟ], "N" for [ɲ], "x" for [x], and "h" for [ɦ]. Note that "q", completely counterintuitively, signals [dz].

that [dz]—“q” in (9) and (10)—is underestimated is understandable; it is more surprising that speakers rank it so high despite the fact that it occurs only in a single foreign word. The model’s fairly good performance on attested and unattested marked clusters suggests that a learner which employs features and natural classes gives a reasonable approximation of how subjects generalize to novel items (involving both attested and unattested sequences).

Looking at the scatter plot in (10) we can see that GNM captures the hierarchy between well-formed clusters, a point where HWPL fails, and ranks even [dz] really high as native speakers did. Probably it gets support from [ts] whose type frequency in our corpus is 293 and *z* with a type frequency of 1777. GNM, however, fails to differentiate between marked clusters as can be seen on the bottom left of the figure in (10), it considerably overestimates sonority reversal clusters (as these get some support from the lexicon) and [ts]*z* and [ɟ]*g* which are rejected by speakers and ranked low by HWPL. All the clusters this model overestimates contain a phonotactic violation which GNM cannot encode. We can also state that GNM does not “discover” the four-consonant template, and in general, it cannot deal with three-consonant clusters either, which might partly be a problem of parameter setting as mentioned in **5.2** and again the model’s inability of encoding the phonotactic restrictions referring to three-consonant sequences. However, even if we disregard long onsets, it is obvious that the model cannot differentiate between marked and illegal forms, especially in those cases when the cluster is unattested because it cannot encode phonotactic violations, a finding very similar to that in Albright (2009). It seems that the two models represent different ways of judging the well-formedness of onset sequences, both of which are reflected in speakers’ ratings.

A multiple regression model reveals that if both HWPL and GNM are combined we get numerically an even better fit to the Slovak data ($r = .903$, $r^2 = .816$) Our findings are in accordance with Shademan (2007) who claims that acceptability ratings are the result of grammatical and analogical mechanisms acting together, with grammar being stronger. She also suggests that real word fillers bias speakers to a processing mode where the lexicon plays a more important role. Albright (2009) could not replicate this effect. In our experiment no real words appeared so speakers were not biased to the analogical or lexical mode of processing even if such bias exists.

6. Conclusions

We are aware that the theoretical interpretation of gradient well-formedness judgments remains controversial, and do not want to take strong positions on whether such judgments are based solely on the knowledge of the lexicon, or solely on phonotactic grammar. It seems that both contribute probably with grammar being more important, so the knowledge of the relative probability of various combinations of natural classes have more influence on phonotactic judgments than simply the knowledge of words in the lexicon. We saw that with a gradient grammatical model based on phonological features and natural classes we can account for much of the data.

We have demonstrated that speakers do have intuitions about unattested grammatical forms as well as attested but marginal ones. We have also seen that forms on the verge of grammaticality that are rated low by native speakers and are rated low by a phonotactic model—like sonority reversal clusters in Slovak—can exist in a language but are prone to change.

References

Albright, Adam 2006. Gradient phonotactic effects: Lexical? Grammatical? Both? Neither? LSA talk handout, January 7, Albuquerque.

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

Albright, Adam – Bruce Hayes 2003. Rules vs. analogy in English past tenses: A computational/experimental study. In: Cognition 90 : 119–61.

Bailey, Todd – Ulrike Hahn 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods. In: Journal of Memory and Language 44 : 568–91.

Bárkányi, Zsuzsanna – Zoltán Kiss 2010. A phonetic approach to the phonology of *v*: A case study from Hungarian and Slovak. In: Susanne Fuchs – Martine Toda – Marzena Żygis (eds): Turbulent sounds. An interdisciplinary guide, 103–42. De Gruyter Mouton, Berlin & New York.

Berent, Iris – Donca Steriade – Tracy Lennertz – Vered Vaknin 2007. What we know about what we have never heard: Evidence from perceptual illusions. In: Cognition 104 : 591–630.

Blanár, Vincent (ed.) 2005. Historický slovník slovenského jazyka [Historical dictionary of the Slovak language]. Vydavateľstvo Veda, Bratislava.

Blevins, Juliette 2004. Evolutionary phonology: The emergence of sound patterns. Cambridge University Press, Cambridge.

Botma, Bert – Norval Smith 2007. A dependency-based typlogy of nasalisation and voicing phenomena. In: Bettelou Los – Marjo van Koppen (eds): Linguistics in the Netherlands, 36–48. John Benjamins, Amsterdam & Philadelphia.

Browman, Cathrine – Louis Goldstein 1990. Gestural specification using dynamically defined articulatory structures. In: Journal of Phonetics 18 : 299–320.

Byrd, Dani 1992. Perception of assimilation in consonant clusters: A gestural model. In: Phonetica 49 : 1–24.

Clements, George N. 1990. The role of the sonority cycle in core syllabification. In: John Kingston – Mary E. Beckman (eds): Papers in laboratory phonology I: Between the grammar and the physics of speech, 283–333. Cambridge University Press, Cambridge.

Coleman, John S. – Janet B. Pierrehumbert 1997. Stochastic phonological grammars and acceptability. In: Computational phonology. Third Meeting of the ACL Special Interest Group in Computational Phonology, 49–56. Association for Computational Linguistics, Somerset NJ.

Cyran, Eugeniusz – Edmund Gussmann 1999. Consonant clusters and governing relations: Polish initial consonant sequences. In: Harry van der Hulst – Nancy Ritter (eds): The syllable: Views and facts, 219–48. Mouton de Gruyter, Berlin & New York.

Frisch, Stefan A. – Nathan R. Large – David B. Pisoni 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. In: Journal of Memory and Language 42 : 481–96.

Frisch, Stefan A. – Janet B. Pierrehumbert – Michael B. Broe 2004. Similarity avoidance and the OCP. In: Natural Language and Linguistic Theory 22 : 179–228.

Greenberg, Joseph H. – James J. Jenkins 1964. Studies in the psychological correlates of the sound system of American English. In: Word 20 : 157–77.

Hayes, Bruce – Colin Wilson 2008. A maximum entropy model of phonotactics and phonotactic learning. In: Linguistic Inquiry 39 : 379–440.

Hooper, Joan B. 1976. An introduction to natural generative phonology. Academic Press, New York.

Kačala, Ján – Mária Pisárčiková (eds) 1987. Krátky slovník slovenského jazyka [Short dictionary of the Slovak language]. Vydavateľstvo Veda, Bratislava.

Kiparsky, Paul 1979. Metrical structure assignment is cyclic. In: Linguistic Inquiry 10 : 421–41.

Kuryłowicz, Jerzy 1952. Uwagi o polskich grupach spółgłoskowych [Remarks on Polish consonantal groups]. In: Biuletyn Polskiego Towarzystwa Językoznawczego 11 : 54–69.

Luce, Paul A. 1986. Neighborhoods of words in the mental lexicon. Technical report. Speech research Laboratory, Department of Psychology, Indiana University.

Maddieson, Ian 1984. Patterns of sounds. Cambridge University Press, Cambridge.

Mistrík, Jozef (ed.) 1976. Retrográdny slovník slovenského jazyka [Inverse dictionary of the Slovak language]. Univerzita Komenského, Bratislava.

Morelli, Frida 1999. The phonotactics and phonology of obstruent clusters in Optimality Theory. Doctoral dissertation, University of Maryland.

Nosofsky, Robert M. 1986. Attention, similarity, and the identification-categorization relationship. In: Journal of Experimental Psychology 15 : 39–57.

Ohala, John J. 1990. There is no interface between phonology and phonetics: A personal view. In: Journal of Phonetics 18 : 153–71.

Ohala, John J. – Manjari Ohala 1986. Testing hypotheses regarding the psychological reality of morpheme structure constraints. In: John J. Ohala – Jeri J. Jaeger (eds): Experimental phonology, 239–52. Academic Press, San Diego.

Padgett, Jay 2002. Russian voicing assimilation, final devoicing, and the problem of [v]. Ms. University of California, Santa Cruz.
(http://people.uscs.edu/ padgett/locker/newvoice.pdf)

Pauliny, Eugén 1979. Fonológia slovenského jazyka [The phonology of the Slovak language]. Slovenské pedagogické nakladateľstvo, Bratislava.

Rowicka, Grażyna 1999. On ghost vowels: A Strict CV approach. Holland Academic Graphics, The Hague.

Rubach, Jerzy 1993. The lexical phonology of Slovak. Clarendon Press, Oxford.

Rubach, Jerzy – Geert E. Booij 1990. Edge of constituent effects in Polish. In: Natural Language and Linguistic Theory 8 : 427–63.

Scheer, Tobias 2004. A lateral theory of phonology. Vol 1: What is CVCV, and why should it be? Mouton de Gruyter, Berlin & New York.

Scheer, Tobias 2006. Distributional gaps in Slavic initial clusters are accidental. Paper presented at Formal Approaches to Slavic Languages 15, Toronto 12–14, May.

Scheer, Tobias 2007. On the status of word-initial clusters in Slavic (and elsewhere). In: Richard Compton – Magdalena Goledzinowska – Ulyana Savchenko (eds): Annual Workshop on Formal Approaches to Slavic Linguistics. The Toronto Meeting 2006, 346–64. Michigan Slavic Publications, Ann Arbor.

Scholes, Robert J. 1966. Phonotactic grammaticality. Mouton, The Hague.

Schütze, Carson 1996. The empirical base of linguistics: Grammaticality judgments and linguistic methodology. The University of Chicago Press, Chicago.

Selkirk, Elisabeth 1982. Syllables. In: Harry van der Hulst – Norval Smith (eds): The structure of phonological representations, part II, 337–83. Foris, Dordrecht, Cinnaminson.

Shademan, Shabnam 2007. Grammar and analogy in phonotactic well-formedness judgments. Doctoral dissertation, UCLA.

Shadle, Christine H. 1985. The acoustics of frictaive consonants (RLE Technical Report 506). MIT Press, Cambridge MA.

Sievers, Edouard 1881. Grundzüge der Phonetik. Breitkopf und Hartel, Leipzig.

Steriade, Donca 1982. Greek prosodies and the nature of syllabification. Doctoral dissertation, MIT.

Stevens, Kenneth N. – Sheila Blumstein – Laura Glicksman – Marta Burton – Katheleen Kurowski 1992. Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. In: Journal of the Acoustical Society of America 91 : 2979–3000.

Torre, Erik Jan van der 2003. Dutch sonorants: The role of place of articulation in phonotactics (LOT Dissertation Series 81). Landelijke Onderzoekschool Taalwetenschap, Utrecht.

Wright, Richard 2004. A review of perceptual cues and cue robustness. In: Bruce Hayes – Robert Kirchner – Donca Steriade (eds): Phonetically based phonology, 34–57. Cambridge University Press, Cambridge.

Zuraw, Kie 2007. The role of phonetic knowledge in phonological patterning: Corpus and survey evidence from Tagalog. In: Language 83 : 277–316.

Appendix

Training corpus of Slovak word-initial onsets (based on ShDSL)

p 3378, p[r] 2069, v 2026, z 1777, [r] 1698, n 1372, k 1336, m 1185, d 1108, l 882, s 882, b 819, [ɲ] 740, [ɦ] 647, t 571, f 393, j 368, [ʃ] 365, [ʧ] 332, [ʦ] 293, st 287, t[r] 275, pl 269, [ʒ] 250, k[r] 238, [ɟ] 210, g 207, [c] 201, sp 182, st[r] 179, [x] 176, zv 174, sk 167, d[r] 166, sl 159, kl 152, dv 149, [ɦr] 138, sv 137, m[r] 132, zl 131, spl 128, z[r] 122, [ɦ]l 116, zm 108, b[r] 103, [ʃ]p 100, bl 98, [ʃ]t 96, sk[r] 88, sp[r] 81, kv 77, [ʃc] 75, vl 73, [xr] 72, s[c] 70, v[r] 70, [ʃ]k 68, [ʃ]tv 68, g[r] 67, zb 65, ml 60, [ʃ]k[r] 57, f[r] 54, zd 53, z[ɦ] 52, sm 52, [ʒ]52, [x]l 49, z[ɲ] 48, zn 47, s[x] 44, f[ʃ] 43, [ʃ]l 43, fl 42, zj 35, [ʃ]v 32, [ʦ]v 31, [ɦɲ] 30, [ʃtr] 30, ps 29, [ʧ]l 28, [ʃ]m 28, mn 27, s[ɲ] 26, tl 26, zd[r] 26, gl 24, [x]v 24, tm 24, tv 24, v 24, vn 24, [ɦ]n 23, k[ɲ] 20, sf 20, s[ʦ] 20, [ʒr] 20, vz 19, [ʤ] 18, vz[ɟ] 18, dl 16, f[c] 16, f[ʧ] 15, [ɦ]m 15, sn 15, zv[r] 15, [ʃr] 14, z[ɦr] 14, zvl 14, [ɦ]v 13, km 13, g[ɟ] 12, kt 12, s[ʧ] 12, [ʃ]n 12, zb[r] 12, [ʒ]l 12, [ʧr] 11, fp 11, fst 11, [x]m 11, fs 10, [ʒ]m 10, ft 9, v[ɟ] 3, vd 6, [ʧ]m 8, ks 8, stv 8, tk 8, bz 7, d[ɲ] 7, fsp 7, skv 7, [ʦc] 7, v[ɲ] 7, vzl 7, vz[r] 7, zbl 7, zml 7, fkl 6, sm[r] 6, [ʃ]kv 6, [ʃpr] 6, v[ɦ] 6, fp[r] 6, vz[ɲ] 6, z[ɟ] 6, [ʧ]v 5, fk 5, fsp[r] 5, s[x]v 5, s[r] 5, stl 5, stm 5, [ʦ]l 5, vm 5, zg 5, zm[r] 5, z[ʒ] 5, fpl 4, pn 4, vzb 4, z[ɦ]l 4, [ʧ]n 3, dn 3, [ʤ]b 3, fk[r] 3, fs[x] 3, fst[r] 3, kn 3, l[ʒ] 3, m[ɲ] 3, [ʃ]kl 3, [ʦ]m 3, vb 3, vz[ɦ]l 3, z[ɦ] [ɲ] 3, [ʧɲ] 2, [ʧ]p 2, [ʤ]g 2, f[x] 2, fsk 2, fsk[r] 2, fs[c] 2, f[ʦ] 2, gn 2, [xɲ] 2, [xc] 2, [xʦ] 2, k[ʃ] 2 mdl 2, p[ʃ] 2, vd[r] 2, vzm 2, v[ʒ] 2, v[ʒ]d 2, zg[r] 2, zhm, 2, zhn 2, zmn 2, bd 1, b[ɟ] 1, bzd 1, [ʧ]k 1, db 1, dm 1, dz 1, f[j] 1, fkv 1, f[ɲ] 1, fskl 1, fspl 1, fstl 1, f[ʃc] 1, ftl 1, ft[r] 1, g[ɲ] 1, gv 1, lk 1, ln 1, lp 1, ls[c] 1, mzd 1, p[x] 1, p[ɲ] 1, pst[r] 1, p[ʃ]t[r] 1, [r]m 1, sv[r] 1, [ʃʧ] 1, t[x] 1, tkl 1, tkv 1, [ʦ]n 1, [ʦɲ] 1, vj 1, vzn 1, v[ʒr] 1, zdn 1, zgl 1, zg[ɲ] 1, zvn 1, [ʒɲ] 1, [ʒ]v 1

Feature matrix used in the computer simulations

	cons	son	cont	wide	nas	voice	lab	cor	ant	strid	lat	dors
p	+	−	−	−		−	+					
t	+	−	−	−		−		+	+	−		
[ʦ]	+	−	−	−		−		+	+	+		
[ʧ]	+	−	−	−		−		+	−	+		
[c]	+	−	−	−		−		+	−	−		
k	+	−	−	−		−						+
b	+	−	−	−		+	+					
d	+	−	−	−		+		+	+	−		
dz	+	−	−	−		+		+	+	+		
[ʤ]	+	−	−	−		+		+	−	+		
[ɟ]	+	−	−	−		+		+	−	−		
g	+	−	−	−		+						+
f	+	−	+	−		−	+					
v	+	−	+	−		+	+					
s	+	−	+	−		−		+	+	+		
z	+	−	+	−		+		+	+	+		
[ʃ]	+	−	+	−		−		+	−	+		
[ʒ]	+	−	+	−		+		+	−	+		
[x]	+	−	+	−		−						+
[ɦ]	+	−	+	−		+						+
m	+	+		−	+		+					
n	+	+		−	+			+	+			
[ɲ]	+	+		−	+			+	−			
l	+	+		+	−			+	+		+	
r	+	+		+	−			+	+		−	
[j]	−	+		−	−			+	−			

The first 25 most important constraints HWPL learned

Constraint	Weight	Observed	Disc. order	Comment
1 *[+son +cor][ ]	6.274	8	11	prohibition on coronal sonorants as C_1
2 *[+voice][−voice]	6.158	0	3	voicing agreement in obstruent clusters
3 *[−ant −strid][ ]	4.786	0	36	no palatal stops as C_1
4 *[−voice][+voice]	4.697	0	2	voicing agreement in obstruent clusters
5 *[−ant][−son.+cont]	4.453	0	7	no palatals or alveo-palatals can precede a fricative
6 *[+dors] j	4.449	0	53	dorsals cannot be followed by [j]
7 *[+cont +dors][−son]	4.408	4	30	[ɦ] and [x] cannot be followed by any sonorant
8 *[−cont][−son +lab]	4.097	6	19	stops cannot be followed by a labial obstruent
9 *[−cont −ant][+cont]	3.733	5	74	(alveo)-palatal stops cannot be followed by a fricative
10 *[−cont][+nas]	3.715	104	73	stops cannot be followed by nasals
11 *[−cont. −voice][+dors]	3.663	13	72	voiceless stops cannot be followed by dorsals
12 *[+son][−ant]	3.475	22	65	sonorantscannot be followed by (alveo)palatals
13 *[+cont −voice +cor] r	3.473	19	92	no *sr* or [ʃ]*r*
14 *[−cont] j	3.329	0	37	stops cannot be followed by [j]
15 *[+lab][−cont]	3.246	98	68	labials cannot be followed by fricatives
16 *[−cont][−cont +ant]	3.245	13	88	stops cannot be followed by t/[ʦ]/d/[dz]
17 *[+voice −ant][+lab]	3.239	14	82	[ʤ]/[ɟ]/[ʒ] cannot be followed by a labial
18 *[+voice +dors][+cont]	3.228	14	87	g/[ɦ] cannot be followed by a fricative+*v*
19 *[−cont +strid][−nas]	3.192	44	76	[ʦ]/[ʧ]/[dz]/[ʤ] cannot be followed by *l*, *r*, [j]
20 *[+voice −ant] l	3.174	12	111	[ʤ]/[ɟ]/[ʒ] cannot be followed by *l*
21 *[+cont −voice] [+son −wide +cor]	3.134	57	75	voiceless fricatives cannot be followed by *n* and [ɲ]
22 *[−strid] l	3.056	51	90	t/[c]/d/[ɟ] cannot be followed by *l*
23 *[+voice +dors] m	2.943	17	117	no *gm* or [ɦ]*m*
24 *[−voice][+cont][+son]	2.911	6	54	a voiceless obstruent cannot form a cluster with a following fricative or *v* which is followed by a sonorant
25 *[−cont +cor][−wide +cor]	2.904	25	64	alveolar/alveo-palatal/ palatal obstruents cannot be followed by coronals except *l* and *r*