Varying Aspiration Levels of the /ř/ Phoneme: An Analogical Modeling Project

Jon Weatherford Stansell

Varying Aspiration Levels of the / ř / Phoneme: An Analogical Modeling Project

Jon Weatherford Stansell

Analogical modeling is the method by which a computer program takes an item (in this case, a word) for whom a certain characteristic is unknown and compares it against a data set of items with verifiable characteristics. This generates a percentage of similarity to those items that are phonetically alike, and thus generates the test item's probability of sharing their characteristics. This process replicates a neural process whereby, when we come to a place where known, immediately recallable language ends, we make analogies with known items to generate a like utterance to fill that obligatory context (such as generalizing -ed as a past tense marker for an unknown verb). This generates standard, rule-like behavior without actually referring to a "rule." My study is focused on exploring the phonetic environments that determine aspiration of one of the most unusual phonemes that I am aware of: the Czech language / ř / or r-hachek. This alveolar flap is found in no other language in the world and is little understood and even less well pronounced by non-natives, including phoneticians. This analogical modeling project supports the claim that there are three different / ř / -phoneme aspiration levels (heavy-, medium-, and light-), which were thought initially to be a function of the pre- / ř / phonemic environment. The analogical modeling program generated other phonemic concerns which also seemed to have an effect on predicting aspiration. One example of this is a /k/ directly before an / ř /, which showed a different aspiration than other consonant-initial / ř / clusters. The model shows many other phonemic environments to cause varying effects on the level of / ř / aspiration. This project shows that assigning "rules" for unknown phonemes in their phonetic environments may be more difficult than initially expected.

The / ř / Phoneme

Though normal pronunciation of / ř / is a simple alveolar flap, the main premise of this study is that clusters of phonemes will effect this aspiration and that like clusters will produce like aspiration levels. If the / ř / is clustered with consonants proceeding, the pronunciation is very forceful (/ vskřišeny /) in order to "push through" the whole cluster. Sometimes this phenomenon creates a semi-syllable where no vowel is present in the lexicography, as in / hřmělo /, which is written with two-syllables, yet is pronounced with the length of a three-syllable word. This heavy aspiration is quite evident in the normally quiet speech of the Czech people. A medium aspiration occurs when the / ř / finds itself in word-initial position, or when it is preceded by a long vowel: / á /, / é /, / í /, / ou /, / u' /, or / ý /. /Ř/ can also be pronounced so softly that it almost seems to not be there. This occurs in words where there is an / ř / preceded by a

short vowel: / a /, / e /, / I /, / o /, / u /, or /y/. These three different forms of aspiration -- heavy, medium, and light -- have a decided effect on the pronunciation of words. Though Czech speakers and scholars identify them all as the same phoneme, the differences in aspiration are clear. One goal of this study is to question whether these conditions listed above are the only factors in the phonemic environment affecting aspiration of / ř /, and whether these perceived "rules" can hold up under the scrutiny of analogical modeling.

Method

With the help of a native Czech, Zlátuše Durdová, I compiled a database of Czech words from two different sources, Čertův Kámen, a collection of Czech folk tales, and Země Lídi, a popular novel in translation. As she read two different folk tales and the first chapter of the novel, I used a Sony portable CD player with an Andrea systems microphone headset to make a tape recording, which lasted approximately one and one-half hours. I then transcribed her pronunciation of 290 Czech words containing the / ř / phoneme. These were then classified as containing a heavy-, medium-, or light-aspirated / ř /. Throughout the database, 16% of / ř / had light aspiration, 30% had medium, and 54% had heavy aspiration. This group of 290 words will now be referred to as "the data set," and will be the standard against which the analogical modeling program will compare the experimental words, or the "test set."

This test set consisted of 40 words from folk tales and a religious text. This may seem like only a few examples, but many of the words in these texts were strange or uncommon. After the first set of 20 had only three words that were not being predicted over 95% accuracy, I began purposely looking for these words, which would give unusual or interesting results. /Ř/ is often found in words with the common prefix /při/, so I stopped recording these because the analogical modeling program continued to report a 100% prediction every time. So many of these syllables occur in the data set (44 examples) that they formed a gang effect, which means that when a test word with a reoccurring phoneme or cluster of phonemes was run through the program, this data was more heavily weighted than other, less represented factors. The prefix /před/ is also commonly found in the data, so I stopped testing words that began with this conglomeration of phonemes because they all produced very high predictions (99.5%). The next twenty examples contained some very unusual words, showing gang effects and word doubling effects.

Results

The first ten words were from the folk tale book. Most of them behaved as expected, with proper aspiration being ascribed to the afore-mentioned phonetic environments, though two of them did not. / Zříditel / would usually be called a heavy-aspiration word because of the consonant cluster. However, the word / řídi / appears several times in the text and has medium aspiration, causing / zříditel / to have only 89% accuracy. Another word that has some ambiguity in its data is / křápla /. It would seem to be heavy-aspirated. But a few elements in the data set had medium aspiration when / ř / was combined with a t in front of it. This is where the extra

medium-aspirated prediction (14%) came from. When I listened to these words and transcribed them, they seemed a little lighter than the heavy-aspirated ones.

The second set of ten test items I examined came from a church publication and had several unique words that would be seldom found elsewhere. These were words for baptism (křest), resurrection (vzkříšeny), immersion (ponořením), and others. Strangely enough, the program had no problem predicting any of these, including the heavily clustered "vzkříšeny," but had a harder time predicting "věřit," or "to believe." This only had 80% light-aspiration predicted, with 2% medium- and 18% heavy-aspiration. This word was more ambiguous because the / ř / was followed by an í. All of the words that affected this prediction had this environment, but also had clustering or long vowels that gave them a heavy- or medium- aspiration. This shows that the program is able to look to phonemes following the / ř / to find analogical matches even though I gave more weight to the phoneme immediately preceding it. The word / řádu / is another example of this phenomenon, for it gathered some unusual attention from two words. Most of the matches were with words starting / ří /, and so they predicted a proper medium aspiration. Two of the matches, however, were / vpředu / and /předu /, which have heavy aspiration. The syllable / du /. in proper place made them analogical pairs with the focus word, which dropped the accuracy of this prediction from 100% to 88%.

Gang effects especially affected two small words of only four phonemes. When I ran the word / míře /, the program predicted an incorrect 66% light aspiration, while the other 33% was proper medium aspiration. Many of the gang words all had / m / in / ř / pre-initial position. For this reason, that variable became heavily weighted, even more than the long vowels in the other examples. Three of the main players in this gang effect also had an / e / following the /ř/ in common with the target word. Because of this gang, all of the more appropriate words were "bullied away" from the analogical set and the prediction was an incorrect light-aspiration. The word / řeki / was properly predicted as medium-aspiration despite some gang-effect problems it encountered in the data set. For light aspiration, it got 18%, and for heavy, it got 22%. The four main problems were the words / přelila /, /odměřeni/, / kořeny /, and / pobřěži /. These may not seem too similar to / řeki /, but following the / ř / are an e, then a consonant, than an I sound. This gang took almost nine percent of the prediction apiece. Little words like / řeki / and /míře/ could not fight the gang-effect.

Three words that were very similar began having a large affect on the prediction of several words. For instance, / zemřes / got 64% heavy aspiration, which is correct, but also 36% medium aspiration. The reason for this spillage were the three words /otevřela/, / nezávřeli /, and / závře /. These words have many phonemes similar to the target word, and since there are three of them, this gang effect makes them more powerful. The major difference is that, while they have a /v/, which is a fricative consonant, the target word has an / m /. This makes all the difference for pronunciation, but the model did not discern this because it was not coded in the data set.

These three words returned again and again to make large contributions towards the prediction of aspiration. / Vřelé / is another word effected by this gang. It took

lots of heavy aspiration (31%) because of its consonant before the / ř /. But / závře /, /otévřela/, and / nezavřeli / made this prediction into a correct medium-aspiration. Another word affected by this gang was /zemřeli/. It had a correct prediction, but the data would not have given it normally. The words / nezávřeli /, / otevřela /, and / závře / took over 77% of the prediction all by themselves. This was a heavy gang effect that worked in favor of the correct prediction. Sometimes, though, a negative effect was observed. The word / odepřel / went from perfect prediction to only 80% heavy aspiration. The causes of this were /otevřila/, which took 16% of the prediction, and / nezavřeli /, which took 8%. These two words which seemed similar to the analogical modeling program were only somewhat so.

/ Hořkos / was one of the words from the data set whose aspiration was difficult for me to define. The program predicted it quite well, with 72% medium aspiration and 28% heavy-. No item appeared in the analogical set for the light-aspiration, to which I thought this word belonged. The word requires a heavy or medium aspiration because of the 'k' immediately following the / ř /. The pronunciation was difficult to ascertain in this circumstance because the length of the vowel immediately preceding the / ř / had little or no influence on it. After the program defined it, I could hear this was correct.

Word repetition in the data set played a vital role in prediction. Though the word / řekl / is common, it did not appear in the data set, and so it was able to be affected by unfamiliar words. Three examples of /přišli/ formed a huge gang effect and swayed the prediction to an incorrect heavy-aspiration. They took almost 50% of the prediction to themselves, then a few other heavy-predicting words made up the rest of the 60% incorrect prediction. Medium aspiration was predicted 22% for the word /hřích/. This was a function of / dřiv / and /třiskne/. The word / drív / is located in the data set twice, with two different aspirations. Despite this contradictory data, the model still worked.

Conclusion

Based on these findings, I believe that this model works extremely well. In words that are common, this program predicts with amazing accuracy. In less-common words, it is still very close. With words that have unusual similarities to other words in the data set, the prediction can break down. Perhaps in another trial of this data, the fricative consonants should be coded, since that had a slight affect on the prediction of aspiration. Also, the aspiration of / ř / immediately following a k or a t is similar to the beginning of a new word, not the heavy aspiration one would expect. This model has shown that the letters immediately following the / ř / have as much to do with its pronunciation as the letters immediately preceding it do. Although the model tried to make matches with other elements in the word that did not relate to its aspiration, the program ran quite well. Perhaps using fewer letters in the data set would eliminate the propensity of the program to seek out unrelated phoneme clusters when looking for analogical matches. The model could, for example, be more robust if it only considered the two letters immediately before and after the / ř / to be important.

In future studies of this phenomenon, a digitized language-aspiration device could be used to determine the data set instead of

the human ear, which could be swayed by what the researcher wants to hear in the data. This would give the model a closer connection to the data, but it could also require some arbitrary divisions in the levels of aspiration. On the other hand, if the division of / ř / aspiration into three levels is a correct heuristic supported by actual divisions in the digitized language data, then the aims of this project would be met more productively and the predictions of aspiration for new words could be more thoroughly ascertained. To make a project of this scale would take more time and effort, but the benefits could be a narrowly defined aspiration level of / ř / for each of its many phonetic environments, which would expand the linguistic knowledge of Czech, a most unique Slavic language.

Questions

Dr. Don Chapman: So, when you assigned the aspiration levels, how did you determine them?

Jon Stansell: The different levels correspond to my perception of the relative force of aspiration. Through the course of the experiment, I came to some conclusions about the different phonological environments that seemed to produce a heavy, medium, or light aspiration. In the presentation, I said these were criteria for a word to be placed in its proper levels. This is incorrect. Upon reflection, I have realized that I did actually classify each word according to how I perceived it, which had a lot to do with the phonetic environment. However, in the case of many words, it went against my perceived 'rules' for classification. This is shown by some of the words with fricative-/ř/ clusters and with some of the /ř/-consonant cluster words. The analogical model shows these patterns in the data set, indicating that there were other factors with a major part in determining aspiration that I hadn't considered before. If I had tried a totally rule-based approach, I would have overlooked these unexpected determining factors. The analogical model showed that my initial conclusions were only partially correct.

Dr. John Robertson: Did you have to assign a Heavy, Medium, or Light level to each word in the data set and what is the relevance of that?

Jon Stansell: Yes. The analogical data set requires an assigning of aspiration. As Dr. Skousen has said, there are two parts to the data set: namely, the raw language data which is the input, and the output, which is the aspiration level. If I had not assigned a level to each word, there would have been no output variable for the data set. The model would not have been able to predict an output for the test data without this output variable. The model makes comparisons between data set input and a test word, assigning an outcome probability value for that word based on the relative occurrence of outputs from only those data set items that are analogically significant. The second part of your question refers to a seeming flaw in my research design, which is that I designed the levels of aspiration, I coded the data, and found that there were predictable patterns when the program was run. This would seem to suggest that these aspiration levels and their occurrence in the data set could have been a generation of my own mind. I do not claim to have a completely discerning ear; in fact, some of my classifications have been

swayed by knowledge of their phonemic environments. There are, however, enough examples of items which went against my initial reaction to show that I did not skew the data. In any case, the computer program did the work, so, supposing the data set was assigned a somewhat correct aspiration, the outcomes will be reliable. The patterns of correlation would show up in the outcome set and their percentages, despite what the final outcome would show. In fact, these patterns of correlation are an excellent way to see how other factors may influence aspiration levels and to check if the assigning of aspiration to the data set was correct. In future studies, I could use native speakers who were not 'tainted' by living in the U.S., as well as using a language digitizing device to precisely determine the aspiration. It may be that this device would reveal an aspiration continuum without any real breaks, or it could reveal my perception of the sound as being in three categories to be correct. I am not totally sure of this based on only one experiment, but I think the results show that there are some natural divisions, outside of my own theoretical biases.

Dr. Robertson: Well, in a perfect world, we could fly native speakers out here, control for every variable, and have equipment that would accurately measure everything we want. But this isn't a perfect world, and I think you've done well with the limitations of the project, considering that it was for a semester class.

h	uprostred	508	0.84
h	preskvouci	708	1.16
h	prezvacny	708	1.16
h	pred	1356	2.23
h	prehrozny	708	1.16
h	prostrednika	924	1.52
h	prida	724	1.19
h	pridal	724	1.19
h	pred	1356	2.23
h	prijde	724	1.19
h	prede	1356	2.23
h	pred	1356	2.23
h	predsevzeti	2400	3.95
h	predu	1356	2.23
h	vpredu	1356	2.23
h	prisli	484	0.80
h	prisla	484	0.80
h	prekrasne	708	1.16
h	pred	1356	2.23
h	predstavuje	2400	3.95
h	pred	1356	2.23
h	pristani		484
	0.80		
h	predpisy	2400	3.95
h	nejpriznejsimi	676	1.11
h	pred	1356	2.23
h	predstavovali	2400	3.95
h	prostrednictvim	924	1.52
h	prostrednictvim	924	1.52
h	prisli	484	0.80
h	pripraven	484	0.80
h	pred	1356	2.23
h	pred	1356	2.23
h	prede	1356	2.23
h	pred	1356	2.23
h	pred	1356	2.23
h	prispecha	484	0.80
h	pred	1356	2.23
h	prestani	708	1.16
h	pripravenim	484	0.80
h	presne	1044	1.72
h	presne	10441.72	
h	prestaveni	708	1.16
h	pred	1356	2.23
h	prede	1356	2.23
h	pribral	484	0.80
h	pred	1356	2.23
h	pripravou	484	0.80

Statistical Summary

h	heavy-asp	60784	100.00

Given Context: = = C m L # R S e = = = =

l	morska	32	3.54
l	mori	56	6.19
l	more	96	10.62
l	more	96	10.62
l	morem	96	10.62
m	vymenkare	48	5.31
l	morsti	32	3.54
m	kancellare	48	5.31
m	kancellare	48	5.31
m	sire	64	7.08
m	boure	48	5.31
l	morem	96	10.62
m	boure	48	5.31
l	samorejme	96	10.62

Statistical Summary

l	light-asp	600	66.37
m	medium-asp	04	33.63

Given Context: = = C h S o R C k S o C s

m	cukrarke	96	11.27
h	horcici	136	15.96
m	s bourkami	112	13.15
h	prerostle	52	6.10
m	sirku	96	11.27
m	vymenkarkou	120	14.08
h	predsevzeti	16	1.88
m	narku	96	11.27
h	predstavuje	16	1.88
h	predstavovali	16	1.88
m	bourki	96	11.27

Statistical Summary

m	medium-asp	616	72.30
h	heavy-asp	23	27.70

Given Context: C z S e C m R S e C X = =

h	mrizi	8	0.73
m	problemy resit	36	3.28
h	strese	88	8.03
h	krese	88	8.03
h	prisli	40	3.65
m	zavre	128	11.68
m	otevrela	32	2.92
h	prisli	72	6.57
h	prisla	72	6.57
h	prisel	72	6.57
m	nezavreli	192	17.52
h	prisel	72	6.57
h	prisel	72	6.57
h	prisli	72	6.57
h	poprisene	52	4.74

Statistical Summary

m	medium-asp	3	88	35.40
h	heavy-asp		708	64.60

Given Context: C z S e C m R S e C l L #

h	mrizi	8	0.70
h	krizi	24	2.09
h	veprova	112	9.76
h	krizi	24	2.09
h	priletne	16	1.39
m	zavre	128	11.15
h	strevic	48	4.18
m	otevrela	308	26.83
m	nezavreli	468	40.77
m	vytvarily	12	1.05

Statistical Summary

m	medium-asp	916	79.79
h	heavy-asp	232	20.21

Given Context: = = = = = = R S e C k C l

m	traslavy	40	16.67
h	strikla	8	3.33
m	traslice	40	16.67
h	prisli	40	16.67
h	prisla	40	16.67
h	prekrasne	16	6.67
h	prisli	40	16.67
m	rekni	16	6.67

Statistical Summary

m	medium-asp	96	40.00
h	heavy-asp	144	60.00

Given Context: = = = = C h R L # C & = =

h	mrizi	100	3.88
h	krizi	100	3.88
m	driv	100	3.88
h	krizi	100	3.88
h	strikla	100	3.88
h	prisli	100	3.88
h	strika	100	3.88
m	nejdriv	100	3.88
h	skripa	100	3.88
h	krizem	100	3.88
h	krizaly	100	3.88
m	triskne	100	3.88
h	driv	100	3.88
h	prihrbena	100	3.88
h	spatrite	100	3.88
h	kriz	100	3.88
h	pristoupil	100	3.88
h	pritelem	100	3.88
h	tricet	100	3.88
h	prilezitost	100	3.88
h	triciti	100	3.88
h	tricet	100	3.88
h	trφcet	100	3.88
h	nepritelem	100	3.88
h	poprisene	100	3.88

Statistical Summary

m	medium-asp	300	11.63
h	heavy-asp	2280	88.37