12

Part title

12

Speech Perception, Well-formedness and the Statistics of the Lexicon

---

JENNIFER HAY, JANET PIERREHUMBERT,
AND MARY BECKMAN

X.1 Introduction

The speech literature abounds in evidence that language-specific phonotactic patterns affect perception. Phonotactics affect placement of phoneme category boundaries (Massaro and Cohen 1983), segmentation of nonce forms (Suomi *et al.* 1997), and speed and accuracy of phoneme monitoring (Otake *et al.* 1996). Papers in previous volumes in this series (Pierrehumbert, 1993; Treiman, 1996) have provided evidence that perceived well-formedness of phoneme combinations is related to their frequency in the language. Coleman (1996) also found that speakers rated neologisms with attested clusters higher than those containing unattested clusters.

These results indicate that speakers generalise over the entries in their lexicons, and respond differently to patterns which are exemplified versus ones which are not. However, patterns may be exemplified to different degrees. This raises the question of whether knowledge of phonotactics is categorical, distinguishing only possible from impossible forms (as predicted by classical generative models), or whether it is gradient, tracking lexical statistics more

finely. Some evidence is available from studies which report different outcomes for high and low-frequency configurations.

Jusczyk *et al.* (1994) found that nine month old infants prefer frequent phonotactic patterns in their language to infrequent ones. Saffran *et al.* (1996a) showed that eight month old infants are sensitive to transitional probabilities in nonsense speech streams. Saffran *et al.* (1996b) show similar sensitivity in adults. Treiman *et al.* (forthcoming) found that high frequency rhymes were judged better, and were more likely to be preserved in blending tasks, than low frequency rhymes. Vitevitch et.al. (1997) demonstrate that subjects rate nonsense words with high-probability phonotactics more highly than nonsense words with low-probability phonotactics, and processing times are also faster for the high probability set. Pitt and McQueen (1998) explore a phoneme boundary effect reported in Elman and McClelland (1988). They show that the decision between /t/ and /k/ is biased by transitional frequencies, and argue for a model in which transitional frequencies are encoded in a pre-processor which parses the speech signal for access to the lexicon.

Coleman and Pierrehumbert (1997) found that rates of acceptance of neologisms as possible English words correlated with log likelihood, as determined by a probabilistic parse of the form. A low r^2 for this correlation indicates their parser did not adequately model all the important factors. However their major claim – that well-formedness of a neologism reflects its cumulative likelihood as a function of its subparts – has been subsequently validated by Frisch *et al.* (forthcoming).

These studies all make comparisons between attested patterns and either less attested or unattested patterns. Thus, they do not distinguish between two alternatives for the status of unattested patterns. One possibility is that they are a smooth extrapolation – the limiting case of less and less attested patterns, as expected if the phonological grammar is a simple projection of the lexical statistics. The other possibility is that unattested patterns are processed in a qualitatively different manner, supporting models in which lexical statistics contribute more indirectly.

This paper explores the perception and well-formedness of nonsense words containing nasal-obstruent (NO) clusters. Morpheme internally, these clusters are subject to a homorganicity constraint in English, which would be represented in a conventional phonology by a feature spreading rule. Yet such a rule does not do justice to the lexical statistics. The strength of the homorganicity requirement depends on the manner of the obstruent and the place of articulation of both the nasal and the obstruent. Some NO clusters are therefore extremely frequent (e.g. /nt/), others are unattested (/mθ/) , and yet others fall between these two extremes (/nf/). Because NO clusters are a phonetically coherent set, and sample the range of frequencies finely, they make

it possible to assess not only the existence, but also the mathematical character, of perception and well-formedness effects related to lexical statistics.

In all experiments reported here, subjects heard nonsense words containing NO clusters, and rated them as possible additions to the English vocabulary. In the first and third experiments, they also transcribed what they had heard in ordinary spelling. We created NO clusters by cross-splicing because the phonetic interpretation of naturally produced ill-formed clusters is problematic. They may prove disfluent because the speaker has little practice with them. Or they may display so much coarticulation that their identity is unclear.

Experiment 1 demonstrates that nasal homorganicity is psychologically real, gradient, and related to lexical frequency. Both the well-formedness judgements and the pattern of corrections in the transcriptions support this conclusion. The next two experiments follow up residual issues related to these results.

First, there was a remote possibility that the quality of cross-splicing was correlated with lexical frequency. The well-formedness ratings would then display the observed pattern even if subjects only attended to the speech quality and not to the phonotactics. Experiment 2 eliminated this possibility by inducing multiple parses for ambiguous compounds. A single stimulus (e.g. zanplirshdom) is rated better with a likely parse (zan-plirshdom) than with an unlikely parse (zanp-lirshdom).

Second, the two unattested clusters in Experiment 1 received anomalously high ratings. Experiment 3 explores the hypothesis that unattested clusters are vulnerable to both reanalysis of the place of the nasal, and to morphological decomposition. We allow for the possibility that the stimulus is interpreted as affixed (as in "camp#er") or as a compound ("sweet#pea"). The well-formedness ratings are found to be predicted by the log probability of the best morphological parse of the word transcriptions. In Section 3 we argue that these results support a model in which perception, production and well-formedness depend on the statistics of the lexicon.

X.2 Experiments

Experiment 1

Experiment 1 involved five series of nine trochaic nonsense words, with each of the nine words containing a different target nasal-obstruent cluster. None of the words begin with a real word substring; we also tried to avoid beginnings which were reminiscent of real words. All non-target onsets, rhymes, and phoneme-to-phoneme transitions are attested. It was not possible to design balanced

stimuli in which no word ended in a substring constituting a real word. However, the pitch accent was on the first syllable, and the second syllable was clearly subordinated.

Transcriptions of the stimuli are shown in Table X.1, with the log probabilities of the target clusters, given the set of bisyllabic monomorphemic trochees[1]. Two of the clusters have a probability of zero, and so the log cannot be calculated for these — their log probability is represented in the table as simply ln(0). Similarly, on the figures to follow, these clusters appear to the left of a disjunction on the x axis, indicating that no log probability value was calculated for these stimuli.

All calculations presented in this paper are based on type frequency in the CELEX lexical database. We believe that it is type frequency, rather than token frequency, which is most directly related to phonotactic well-formedness. The stimuli here were not constructed to directly test this hypothesis, and in fact, for the set of nasal-obstruent clusters discussed here, type and token frequency are highly correlated. However post-hoc analysis revealed that despite this fact, type frequency out-performs token frequency in predicting the distribution of well-formedness judgements in our data.

Table X.1

Set 1	**Set 2**	**Set 3**	**Set 4**	**Set 5**	**ln P**
zæntəˈ	stɹɪnti	krɛntɪk	gɹɑntəlt	slɛntu	-4.2
zæmpəˈ	stɹɪmpi	krɛmpɪk	gɹɑmpəlt	slɛmpu	-4.5
zænfəˈ	stɹɪnfi	krɛnfɪk	gɹɑnfəlt	slɛnfu	-6.87
zæmfəˈ	stɹɪmfi	krɛmfɪk	gɹɑmfəlt	slɛmfu	-7.16
zæmkəˈ	stɹɪmki	krɛmkɪk	gɹɑmkəlt	slɛmku	-7.57
zæmsəˈ	stɹɪmsi	krɛmsɪk	gɹɑmsəlt	slɛmsu	-8.26
zænkəˈ	stɹɪnki	krɛnkɪk	gɹɑnkəlt	slɛnku	-8.26
zænpəˈ	stɹɪnpi	krɛnpɪk	gɹɑnpəlt	slɛnpu	ln(0)
zæmθəˈ	stɹɪmθi	krɛmθɪk	gɹɑmθəlt	slɛmθu	ln(0)

All stimuli were created by cross-splicing. First, a naive male speaker of General American English produced nonsense words containing homorganic clusters. These nonsense words mixed beginnings from one stimulus set with endings from another set; for example, the words included *zænti* and *strɪmpəˈ* in order to support creation of *zænpəˈ*. Two stems for each set were excised: one in which the nasal originally appeared before a stop, and one in which it appeared before a fricative. Word endings were excised starting at the voiceless

obstruent. Stimuli were constructed by splicing the stem to the relevant ending. Three randomised blocks of all 45 stimuli were presented to 11 subjects, who judged the acceptability of each word on a scale from 1 to 10. A scale was used so that artificial categoriality would not be induced by the task. Subjects then wrote how they thought the word would be spelled in ordinary English spelling.

The results provide two types of evidence that the mental representation of phonotactics reflect lexical frequency. First, a *post hoc* tabulation of the transcribed data revealed that most reanalyses were from a less frequent cluster to a more frequent cluster (409 improving reanalyses vs. 147 worsening reanalyses).

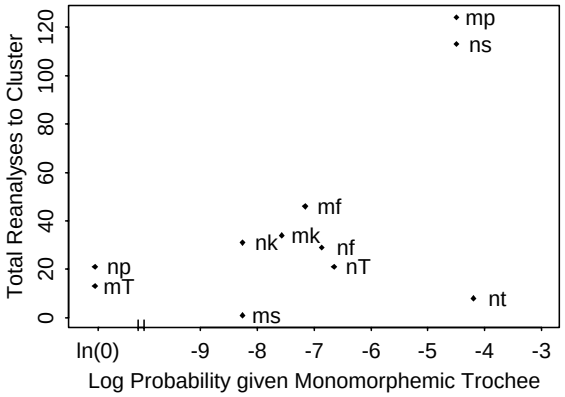

Figure X.1: Distribution of the outcomes of analysis

The rate of improving reanalyses was negatively correlated with the log probability of the actual input cluster (r^2 = .64, df = 7, p < 0.01); worse clusters were corrected more often[2]. The rate of worsening reanalyses was uncorrelated with log probability. This result indicates that lexical frequencies are active in the perception-production loop that resulted in the transcription. If reanalyses occurred entirely at random, the overall pattern of reanalyses would be towards the middle of the frequency range. Regression towards the mean would occur, because the lowest frequency clusters can only be reanalysed upwards, and the highest frequency clusters can only be reanalysed downwards. If reanalyses were based only on acoustic similarity, they would not correlate with cluster frequency.

Figure X.1 shows the distribution of outcomes of reanalysis. The log probability of each cluster in trochaic mono-morphemes is graphed against how often that cluster was the outcome

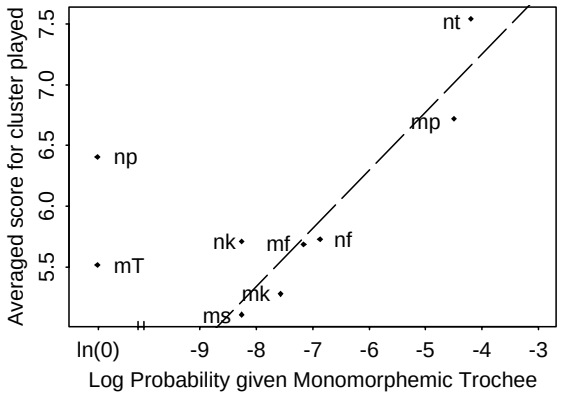

Figure X.2: Distribution of well-formedness judgements

of reanalysis. For example, the log probability of /mp/ is -4.5, and there were 124 cases in which a cluster other than /mp/ was reanalysed as /mp/. Note that 'T' is used to represent /θ/ in this graph. The graph contains some clusters which were not explicitly included in the experiment. For example the cluster /ns/ did not appear in the stimuli, but it appears on this graph because of the high frequency with which other clusters (such as /ms/) were reanalysed as /ns/. The upper left-hand quadrant of this graph is empty, as expected if reanalysis is a function of both acoustic similarity and lexical frequency. High frequency clusters attract responses, but only if they are acoustically similar to the speech signal. The cluster /nt/, for example, is high frequency, but was not similar to any cluster in the stimuli. Low frequency clusters are not outcomes of reanalysis, no matter how similar they are to the stimulus.

The well-formedness judgements also reflect the influence of lexical frequency. Excluding the unattested clusters, the mean acceptability of each cluster was highly correlated with its log frequency ($r^2 = .87$, $df = 5$, $p < .003$). A gradient dependence was also found within subjects, with 10 of the 11 subjects showing $r^2 > .64$ and $p < .03$.

The two unattested clusters (/ŋp/ and /mθ/) showed anomalous behaviour. The mean rating for /ŋp/ was 6.41, the mean rating for /mθ/ was 5.52, whereas the mean ratings for the lowest two attested clusters were 5.28 and 5.11. The overall picture of well-formedness ratings is shown in Figure X.2. To the right of the discontinuity in the x axis, we see the gradient relationship between log probability and well-formedness, as summarised by the regression line. The two unattested clusters are shown to the left of the discontinuity. They lie above the regression line. One hypothesis for explaining this boomerang shape might be that the unattested clusters were analysed on perception as higher frequency clusters. But this does not fully explain their behaviour. On an analysis of the ratings of the clusters actually transcribed (not shown), the /mθ/ and /ŋp/ clusters are still anomalous; in many instances, subjects heard and accurately transcribed the sequences, but still rated them highly.

To explain this result, we hypothesised that words transcribed with unattested clusters were interpreted as containing an internal boundary. In experiment 3 we explore this hypothesis, and find that it explains the response pattern very well.

Experiment 2

Experiment two was designed to eliminate any possibility that the phonetic quality of the cross-splices was responsible for the results. Nonsense compounds were created in which the affiliation of an obstruent is ambiguous,

either ending the first word or beginning the second. If the splicing is responsible for the effects observed in experiment 1, then we expect subjects to rate both parses of the same phonetic item identically. If, however, the effects are due to gradient phonotactic well-formedness, we expect different ratings for the two parses; and this difference in ratings should be a function of the difference in probability between the two parses.

Each NO cluster was used in two stimulus types. In one type, the obstruent could be interpreted as either ending the first syllable, or providing an onset for the following syllable (e.g. "zampirshdom" may be analysed as "zamp-irshdom" or as "zam-pirshdom"). In the second type, the obstruent either ends the first syllable, or is the first phoneme in a cluster onset for the second (e.g. "zamplirshdom" may be interpreted either as "zamp-lirshdom" or as "zam-plirshdom"). Including both simple and complex onsets in the stimulus set allowed us to maximise the difference in probability between competing parses, and to ensure that for some pairs the affiliation of the obstruent to the coda was the more probable parse, and for others the onset parse was more probable. Table X.2 shows the stimuli, with the log expected probability of the parses[3].

Table X.2

	compound	ln(prob)	compound	ln(prob)
Type 1:	zan-sirshdom	-8.79	zans-irshdom	-9.64
	zam-pirshdom	-8.97	zamp-irshdom	-8.43
	zan-tirshdom	-9.25	zant-irshdom	-8.96
	zam-firshdom	-9.50	zamf-irshdom	(-18.78)
	zan-pirshdom	-8.76	zanp-irshdom	(-18.78)
	zam-tirshdom	-9.46	zamt-irshdom	(-18.78)
Type 2:	zan-swirshdom	-11.18	zans-wirshdom	-11.51
	zam-plirshdom	-10.32	zamp-lirshdom	-8.61
	zan-twirshdom	-11.74	zant-wirshdom	-10.77
	zam-flirshdom	-10.45	zamf-lirshdom	(-18.78)
	zan-plirshdom	-10.10	zanp-lirshdom	(-18.78)
	zam-twirshdom	-12.02	zamt-wirshdom	(-18.78)

Following Coleman and Pierrehumbert (1997), the expected probability for the critical region of *zam-plirshdom* is calculated by taking the product of the probability of an /æm/ rhyme and the probability of a /pl/ onset, given our corpus of monomorphemes.

Each cross-spliced stimulus was then presented in two discourse contexts. The same token *zamplirshdom*, for example, was presented separately in each of the contexts shown in (2)a and b.

2. a. This is a zam, and this is a plirshdom. A plirshdom for a zam is a zam-plirshdom.
   b. This is a zamp, and this is a lirshdom. A lirshdom for a zamp is a zamp-lirshdom.

As in experiment 1, the NO clusters were created by cross-splicing from homorganic environments.. Each of the subparts in the contextual sentence was spliced out from the compound, and so represented identical phonetic material.

Subjects heard and read the stimuli in context. A written version was provided to reduce the complication of reanalysis of low frequency forms. The compound appeared in bold, and nine subjects were asked to rate it from one to seven, according to "how acceptable it would be as a word form for the English of the near future".

The different priming conditions induced different well-formedness judgements for the same phonetic stimuli. Furthermore, the difference in judgements between competing parses is predicted by the difference in log expected probability between those two parses (r^2=.73, df = 10, p< .001). That is, the well-formedness ratings were better for the more probable parse, and the larger the difference in probability between competing parses, the larger the difference in well-formedness judgements between them. This result provides clear validation of the cross-splicing technique. Subjects' well-formedness ratings are related to the probability of the nonsense form. Even when subjects are presented with identical stimuli, well-formedness ratings shift if the probability of the form is manipulated.

Experiment 3

As discussed, the attested clusters in experiment 1 showed remarkably uniform behaviour. However the judgements for /mθ/ and /np/ were anomalously high. We hypothesised that the high ratings were due to morphology. We originally sought to explore clusters in monomorphemes, but cannot be sure that the stimuli received a monomorphemic parse. The clusters may have been perceived as preceding an affix ("zamp#er", like "camp#er"), or as bridging a boundary between morphemes ("strin#pea", like "sweet#pea). Experiment 3 was designed to test this hypothesis.

We included all voiceless labial and coronal NO clusters[4] – a total of 14, including five unattested clusters. We chose the three sets from experiment 1 that received the highest ratings: zæNOɚ, stɪNOi and krɛNOɪk. All three have lax front vowels in the strong syllable, our probability calculations take account of this fact. This reflects an attempt to use the most narrow description

of our stimuli which does not make the render the universe of comparison so small as to make the estimates unstable, and thus to provide the most precise description of the data possible without sacrificing generalisability across the sets. Fresh cross-spliced stimuli were constructed. This gave 52 stimuli, which were presented in block randomised order three times each. 9 subjects rated the words from 1 to 7, and spelled them, using the same instructions used in experiment 1.

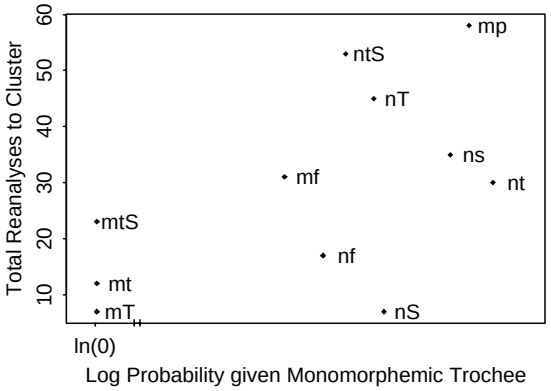

Figure X.3 : Distribution of outcomes of analysis

Figure X.3, corresponding to Figure X.1 for experiment 1, shows the outcomes of phonemic reanalysis. The overall counts of errors are not comparable across the two experiments because of differing numbers of stimuli and subjects. This Figure has the same empty upper-left quadrant, and supports the conclusion that reanalyses tend to be towards acoustically similar clusters which are more frequent in the lexicon. Note that in this, and subsequent graphs, "S" is used for /ʃ/, and "T" for /θ/.

Figure X.4, corresponding to Figure X.2 for experiment 1, shows the ratings of the new stimulus set. As in experiment 1, the unattested clusters differ in their perceived well-formedness. Also, the best unattested clusters are better than the worst attested clusters. However the careful orthogonal design of this stimulus set has not caused the responses to fall neatly on a line, but rather has produced a graph with scatter, reminiscent of the data in Coleman and Pierrehumbert (1997). The r^2 for the nonzero clusters is .49, with $p < .04$ ($df = 7$).

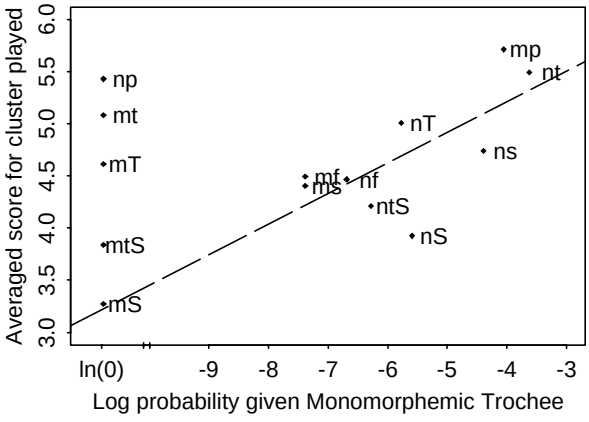

Figure X.4: Distribution of well-formedness judgements

Plots we do not show reveal that the

scatter is reduced a little by looking at the frequency of the cluster actually transcribed. Some portion of the time, low frequency clusters were reanalysed, and the rating (which occurred before the transcription) may be based on this reanalysis. This reduces some of the scatter, but certainly not all. And there is still considerable variation in the well-formedness of the zero frequency clusters. The data becomes much more orderly, only when we also allow for the possibility of competing morphological parses.

There are two possible parses in addition to a monomorphemic parse: a compound parse, in which the boundary splits the cluster, (krem#pick, like drum#pad), and a affixed parse in which the boundary follows the cluster (zamp#er, like camp#er.) We wish to explore the idea that parses such as these compete in terms of likelihood.

Table X.3 shows the raw counts of monomorphemic and affixed trochees, and of monosyllabic words, with lax front vowels. These counts are taken from CELEX. Monosyllables are included because each syllable of a bisyllabic compound is a monosyllabic word, and we calculate compound probabilities based on the assumption that any two monosyllabic words can combine to form a compound. That is, compound probabilities are assumed to be roughly approximated by probabilities across word boundaries. Assuming the three cases in table X.3 exhaust the possibilities, we can translate them into probabilities, as shown in the third column[5]. This gives a rough estimate of the overall probability of each of the parses, given the first syllable is strong, and contains a lax front vowel.

Note that these are straight counts of type frequency in the CELEX lexicon. There is no attempt to model the effects of morphological productivity, and so the counts may be slightly conservative. There may be some words ending in #er, for example, which some subjects have in their lexicon, yet are not represented in CELEX. However, we do not want to make the assumption that subjects generalise over *all possible* words with #er, whether or not they have ever encountered them. That is, we are trying to approximate the results of a generalisation over an existent lexicon, not over the potentially infinite set of words which could be licensed by morphological productivity.

Table X.3

	count	prob
monomorphemic:	1609	.457
affixed:	866	.246
monosyllabic:	1048	.297

We can now estimate the probability of each cluster given each parse, by simply calculating the proportion of types for any given parse, which contain the relevant cluster. For example, of the 866 affixed trochees with lax front vowels, nine contain an /mp/ cluster directly before the morpheme boundary. The probability of an /mp/ medial cluster given an affixed trochee with a lax front vowel is therefore 9/866. For each cluster, the overall probability of it occurring in a given parse can be estimated by the probability of the parse times the probability of the cluster occurring, given the parse. So the probability of encountering /mp/ in an affixed trochee with a lax front vowel is .246*9/866. These, and the analogous calculations for the other two analyses for /mp/ are summarised below. They show that, for /mp/, the best parse is as a monomorphemic trochee, with a probability of .00795276.

Monomorphemic analysis:
(za**mp**er **,** like pa**mp**er)

P(*monomorphemic trochee* | *lax front vowel*) x
P(*/mp/ medial cluster* | *monomorphemic trochee with lax front vowel*)
= .457 * (28/1609)
= 0.00795276

Analysis as a CC# suffix:
(za**mp**#er **,** like ca**mp**#er)

P(*bimorphemic trochee* | *lax front vowel*) x
P(*/mp/ cluster before #* | *bimorphemic trochee with lax front vowel*)
= .246 * (9/866)
= 0.00256

Analysis as a trochaic compound:
(kre**m#p**ick**,** like dru**m#p**ad)

P(*monosyllabic word* | *lax front vowel*) x
P(*/m/ coda* | *monosyllabic word with lax front vowel*) x
P(*second syll is a monosyllabic word given the first was*) x
P(*/p/ onset* | *monosyllabic word*)
= .297 * (44/1048) * 1 * (174/3880)
= 0.000559

We completed these three calculations for each of the 14 clusters. Figure X.5 shows how the probabilities of bimorphemic parses compare to those of monomorphemic parses.

On the x axis is the probability of the target cluster in a monomorphemic word. On the y axis is the probability in a bimorphemic word, with open squares representing the case of words with a boundary after the cluster (as in *camp#er*), and filled squares representing the case in which the boundary splits

the cluster *(*as in *drum#pad*). If the cluster probabilities were the same for bimorphemic as for monomorphemic parses, the points would fall on the line x=y. But they are not the same.

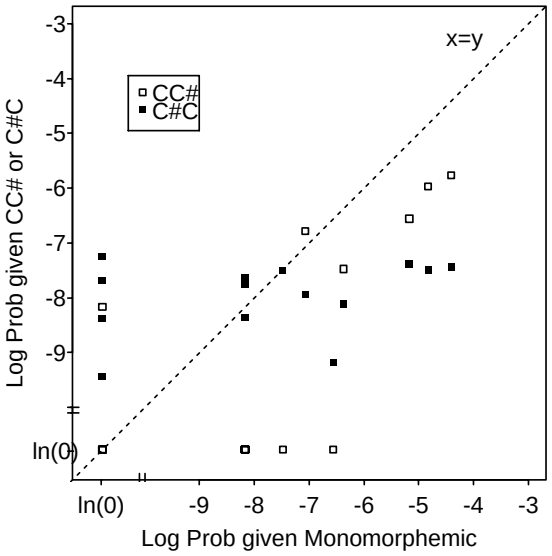

Figure X.5: log probability of clusters given monomorphemic vs log probability given affixed or compound parses

Across word boundaries consonants combine independently[6]. The probabilities for clusters in monomorphemes are therefore polarised compared to those in compounds (filled squares). Some clusters are less probable morpheme internally than the most improbable cluster across a word boundary. But the most probable morpheme internal clusters are more probable than any cluster crossing a word boundary. The parses involving an affix (open squares) fall between the monomorphemes and the compounds. The range in the y dimension is greater than for compounds (it extends to include zero probabilities), but less than for monomorphemes. The probabilities of clusters on either bimorphemic analysis are not correlated with the probability of the monomorphemic analysis. Both filled and open squares fall in a horizontal band.

The pattern in Figure X.5 is, we would argue, a fundamental characteristic of phonology. It would be logically possible to design a language in which phonological restrictions across word boundaries were strong and pervasive. But human language is not like this. Because words have arbitrary meanings, and coherence of meaning dominates how words are combined, phoneme transitions across word boundaries are closer to being statistically independent than transitions within morphemes.

Suppose that listeners choose the most probable analysis of what they hear. If a cluster is highly probable morpheme internally, then no bimorphemic analysis of that cluster can ever be more probable. However, if the cluster is

improbable morpheme internally, then a bimorphemic analysis might be more probable. Table X.4 shows the probability of the winning parse for each cluster.

Table X.4

Cluster	Prob. of Best Parse	Best Parse
nt	0.0122132	(monomorphemic)
mp	0.00795276	(monomorphemic)
ns	0.00568056	(monomorphemic)
nʃ	0.00170416	(monomorphemic)
nθ	0.00142014	(monomorphemic)
ntʃ	0.00113626	(CC#)
np	0.000711707	(C#C)
nf	0.000568056	(monomorphemic)
ms	0.00048528	(C#C)
mt	0.000459572	(C#C)
mf	0.000430647	(C#C)
mʃ	0.000234679	(C#C)
mtʃ	0.000231392	(C#C)
mθ	0.000080344	(C#C)

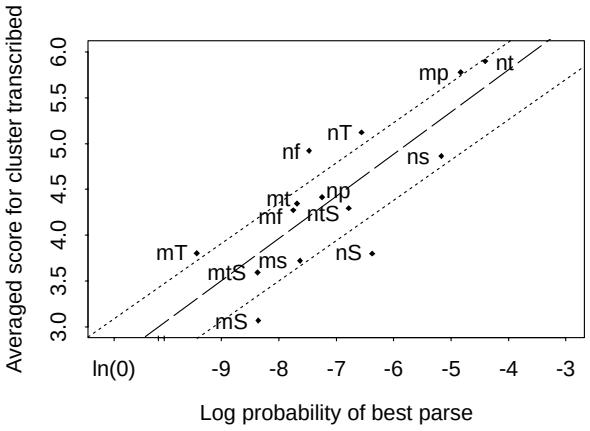

Figure X.6: log probability of the most likely parse for each cluster, vs well-formedness judgements for clusters perceived as that cluster

Figure X.6 shows that viewing our data in this light renders them very orderly. On the x axis is the log probability of the most likely parse for each cluster. On the y axis is the mean well-formedness judgement for stimuli (rightly or wrongly) transcribed with that cluster. Each time /np/ is transcribed as /mp/, for example, the related well-formedness score contributes to the

average well-formedness of /mp/ rather than /np/. The data are now more linear than in Figure X.4, with the regression line for the whole data set having an r^2 of .65, df = 12, $p < .0005$. This gradience is also present within individual subjects. Individual r^2s range from .28 to .71, and p values from .05 to .0002.

Examination of the residuals from this regression line shows an additional factor at work. Clusters containing strident obstruents (/s/, /ʃ/ or /tʃ/) fall below the regression line, whereas those containing other obstruents fall above it. This pattern reflects an effect of the Obligatory-Contour-Principle (OCP). OCP-Place refers to a tendency to avoid excessively similar consonants at the same place of articulation in close proximity (see McCarthy 1986). The "striN" and "zaN" stems begin with coronal stridents, and additional coronal stridents appear to be avoided even as far away as the beginning of the next syllable. Support for this interpretation can be found by comparing overall ratings for the three stems. The stem "kreN" does not contain a strident, so we predict that judgements for this set will not show a difference between stridents and other obstruents. Indeed, the difference in average well-formedness rating between stridents and non-stridents is only 0.07 for the "kreN" set, but 1.43 for the "zaN" set, and 1.39 for the "striN" set. This is consistent with findings by Pierrehumbert (1994) and Berkley (1994) indicating that OCP-Place operates across intervening phonemes in English.

When we fit separate lines through the strident and nonstrident points of Figure X.6, very high r^2 values are obtained. These lines are shown in dots; for stridents, $r^2=.8$, df=4, $p<.02$; for nonstridents, $r^2=.93$, df=6, $p<.0001$. The overall well-formedness reflects a cumulative effect of the local probability of the parse and the long-distance factor of the OCP.

The well-formedness ratings in experiment 3 reflect the most probable analysis of the stimulus. There is active competition amongst multiple analyses of the same stimulus, and the listener probabilistically imputes a phonemic and morphological analysis. As a result, judgements are well-behaved when plotted against the probability of the best analysis of the cluster transcribed, but poorly behaved when plotted against the morpheme-internal probability of the original cluster played.

X.3 Discussion and Conclusion

This study was undertaken to evaluate the relationship between well-formedness and frequency in the lexicon. The results support a model in which well-formedness is directly related to the perceived likelihood of the form. Furthermore, this relationship is gradient rather than categorical.

In the model we would propose, lexical probabilities figure twice. First, they influence perception and reproduction of the stimuli. Second, they

determine perceived well-formedness. The percept of a stimulus is a probabilistic function of its acoustic character and of the likelihood of its components. The best analysis of a stimulus may involve relabelling phonemes and/or imputing morpheme boundaries. That the same probabilities figure twice – both in perception and in judging the result of the perception – supports models in which perception, production, and well-formedness all depend on lexical frequency. That well-formedness is based on the optimal analysis supports models in which analyses of the signal compete, and in which recognising a signal as having a particular phonological form is equated with the triumph of that form in the competition with its alternatives. Models that have this property include both connectionist models (e.g. Rumelhart and McClelland, 1986) and Hidden Markov models (see Rabiner and Juang, 1986 for an overview).

The interaction of NO likelihood and the OCP broadly supports the claim of Coleman and Pierrehumbert (1997) that well-formedness reflects the cumulative effect of the likelihood of the subparts. However, the Coleman and Pierrehumbert model does not handle the effects we have found here. Their model provides for the probabilistic interaction of syllable onsets and rhymes as components in a metrical tree. However, the NO clusters investigated here crosscut the syllable structure, as they include the coda of one syllable and the onset of the next. That such strong results have been obtained for NO clusters indicates that not only syllabic components, but also junctures, are important cognitive elements. Contrary to Levelt (1989) and Schiller (1997), junctural configurations appear to be just as cognitively important as the onset and rhyme configurations that the junctures crosscut.

The Coleman and Pierrehumbert model also does not provide for the interaction between NO phonotactics and the OCP. Their model provides only for the interaction of independent components combined in sequence. In the present data, both the NO phonotactics and OCP-Place target the post-nasal obstruent. The failure to model junctural effects and overlaid generalisations is probably one reason for the high degree of scatter in Coleman and Pierrehumbert's results. Another is that they do not model the reanalysis of stimuli as phonologically better forms.

One of the main goals of our study was to investigate the status of unattested clusters. Are these clusters categorically different from attested clusters, or can zero frequency be viewed as the limiting case of low frequency? Under the model we propose, unattested clusters do differ from attested ones in that they are not exemplified in the lexicon. Since stimuli are analysed with reference to the lexicon, unattested clusters must be coerced onto a form with non-zero probability. This coercion may involve reanalysis of phonemes, or it may involve imputing an internal boundary. This coercion is not unique to unattested clusters however. Low frequency clusters are also probabilistically

reanalysed as more probable ones, and are likely to receive a bimorphemic parse. In these respects, unattested clusters do indeed show the limiting behaviour of less and less probable clusters.

A possible objection to this model is its apparent prediction that unattested combinations can never be recognised as single words. This prediction is obviously falsified in our own experiment, since all of our stimuli were nonsense words and therefore represented combinations of 5 to 8 phonemes which are not exemplified in the lexicon. However, the results lead us to believe that subjects judged these as monomorphemic, and common experience shows that such words can be internalised readily and added to the English vocabulary. This objection rests on the assumption that probabilities are computed over large, detailed, phonological fragments. If the relevant probability for each cluster were its probability in light of the entire phonological description up to that point in the word, then all of the clusters in our experiment would have probability zero. By the time /m/ of "strimpy" is reached, the cumulative probability of the analysis is zero. No words begin in "strim"[7]. The ability of the subjects to reliably assess differences in likelihood indicates that probabilities over large fragments are not relevant to this task. To explain knowledge of phonotactics, it is necessary to posit abstraction over the lexicon. Thus, our work broadly provides new evidence for Pitt and McQueens' claim (against Elman and Mclelland) that knowledge of phonological grammar abstracts over the lexicon. Phonotactic knowledge cannot simply consist of cumulative probability calculated from the hypothesised word-onset.

We ourselves made an abstraction over the lexicon when we computed the probabilities of various parses. Were the universes we selected for computing probabilities cognitively realistic? We suggest three ways in which they were. First, the phonological descriptions we evaluated were formally simple. Second, they represented descriptions which would plausibly be recovered bottom up from the speech signal[8]. Third, they provided sets of large enough size that probabilities could be reliably estimated. Each of these factors affects the degree to which a computation is robust. And we might expect listeners to be most finely attuned to probabilistic patterns which are robust: simple descriptions, which are easily recoverable and statistically reliable. Because these factors were in confluence, we are not in a position to offer conjectures about which is most important. If they are naturally in confluence – all appearing together when examples of psychologically real probabilities are found – then this would be an important fact about cognition.

References

Baayen, R.H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX Lexical Database (Release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

Berkley, D. M. 1994. Variability in Obligatory Contour Principle Effects. Papers from the 30th Regional Meeting of the Chicago Linguistic Society, Part 1

Coleman, J. S. (1996) The psychological reality of language-specific constraints. Paper presented at the Fourth Phonology Meeting, University of Manchester, May 1996

Coleman J. and J. Pierrehumbert (1997) Stochastic phonological grammars and acceptability. SIGPHON 1997.

Elman and J. McClelland (1988) "Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes," Journal of Memory and Language 27, 143-165.

Frisch S., N. Large, and D.B.Pisoni, D.B (forthcoming). Perception of wordlikeness: Effects of segment probability and length on subjective ratings and processing of non-words. Journal of Memory and Language.

Jusczyk, P. W. Luce, P. A. , and Charles Luce -J. (1994) "Infants Sensitivity to phonotactic patterns in the native language," Journal of Memory and Language 33. 630-645

Levelt, W.J.M. (1989) Speaking: From Intention to Articulation. MIT Press.

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. Journal of Experimental Psychology: Human Perception and Performance, 9, 753-771.

McCarthy, J. (1986) ``OCP Effects: Gemination and Antigemination," Linguistic Inquiry 17, 207-263.

Mermelstein. J (1975). Automatic Segmentation of Speech. Journal of the Acoustical Society of America, 58:880-883.

Otake, T., K. Yoneyama, A. Cutler and A. van der Lugt. (1996) The representation of Japanese moraic nasals. Journal of the Acoustical Society of America. 100 (6) 3831-3842

Pierrehumbert, J. (1994) Syllable structure and word structure: a study of triconsonantal clusters in English. In P. Keating (ed) Phonological Structure and Phonetic Form. Papers in Laboratory Phonology III. Cambridge University Press, Cambridge.

Pitt, M.A. & McQueen, J.M. (1998). Is compensation for coarticulation mediated by the lexicon? Journal of Memory and Language 39(3):347-370..

Rabiner, L & Juang, B.H. (1986) An Introduction to Hidden Markov Models. IEEE ASSP Magazine. (1) 4-16.

Rumelhart, D.E., J.L. McClelland and the PDP Research Group (1986), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations, Cambridge, MA: MIT Press

Saffran, J., R. Aslin, and E. Newport (1996a) "Statistical Learning by 8-Month Old Infants," Science 274, 1926-1928,

Saffran, J., E. Newport, and R. Aslin (1996b) "Word Segmentation: The Role of Distributional Cues," Journal of Memory and Language 35 606-621.

Schiller, N.O. (1997) The role of the syllable in speech production. Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. PhD Dissertation, Nijmegen University.

Shipman, D.W. and Zue, V.W (1982) "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems", Conference Record, IEEE International Conference on Speech Acoustics and Signal Processing, Paris, France, 546-549.

Suomi, K., J.M. McQueen and A. Cutler (1997), Vowel Harmony and Speech Segmentation in Finnish. Journal of Memory and Language, 36, 422-444.

Treiman, R. B. Kessler ,S. Knewasser, and R. Tincoff (forthcoming) Adults' sensitivity to phonotactic probabilities. To appear in Papers in Laboratory Phonology V.

Vitevitch, M.S., P.A. Luce, J. Charles-Luce and D. Kemmerer (1997), Phonotactics and Syllable Stress: Implications for the Processing of Spoken Nonsense Words. Language and Speech 40(1), 47-62.

Warner (1998) Dynamic cues in speech perception and spoken word recognition. PhD. Dissertation, University of California at Berkeley

---

[1] All probabilities reported in this paper were calculated using the CELEX Lexical Database (Baayen et al 1995). Our corpus of monomorphemes includes all words coded as monomorphemic in CELEX, as well as many words coded as having "obscure morphology" or as "possibly containing a root". Three linguists independently identified words in the latter two categories which they considered to be multimorphemic. Any word identified by any of the linguists as multimorphemic was omitted from the corpus. The researchers also rejected several words coded as monomorphemic, including reduplicative forms (e.g. *tomtom*), and adjectives relating to places or ethnic groups (eg *Mayan*).

[2] In this calculation, unattested clusters were treated as if they occurred just once in the lexicon, to avoid taking log of zero. Also note that if /nk/ was transcribed as "ɲk", this was not counted as a reanalysis, even though it may well have been heard as /ŋk/, but not recorded by the English spelling. This r^2 is therefore a conservative estimate.

[3] To avoid taking log of zero, the expected probabilities for zero frequency forms was set to ln(0.000000007)=-18.78. This is the probability a compound would receive if there was just one pair of words in the corpus which could combine to create a compound with the relevant characteristics.

[4] Velars were eliminated because English spelling does not record the place of articulation of a nasal before a velar.

[5] This is a slight over-simplification, as it omits marginal possibilities such as compounds where the cluster ends the first word (camp#out).

[6] While external sandhi provides an exception, there is no reason to believe it is relevant to this data set.

[7] Our reviewers point out that 'strim" and "strimmer" are, in fact, words in British English. It is reasonably safe for us to assume, however, that these were not items in our subjects' lexicons.

[8] See Shipman and Zue (1982) for evidence that major class features for consonants and major groupings of vowels can be recovered bottom up from the speech signal, and Mermelstein (1975) for evidence of syllable count recoverability.