Constituents can exhibit partial overlap: Experimental evidence for an exemplar approach to the mental lexicon[1]

Vsevolod M. Kapatsinski
University of New Mexico

1 Introduction

Exemplar models of long-term memory have proposed that complex units (constituents) emerge from co-occurrence of smaller units (Bybee and Scheibman 1999, Bybee 2002). Put in another way, if two units frequently occur together, they fuse into a larger unit. As long as the frequency of co-occurrence is high enough, the units will fuse, even if the process results in partially overlapping units (the existence of partially overlapping units has been posited by Skousen 1989, 2002a for phonological units and by Bod 1992, 1998 for syntactic ones).

Bodies and rimes of syllables (a.k.a. syllabic constituents) have been argued to emerge from patterns of segment co-occurrence. Thus, Yoon and Derwing (2001) have demonstrated that for monolingual English speakers the vowel in a monosyllabic word is more strongly connected to the coda than to the onset while the opposite is the case for monolingual Korean speakers. Bilingual English-Korean speakers exhibit the English pattern in English and the Korean one in Korean. Evidence comes from the pause insertion task where, when forced to pause within a syllable, speakers pause between the onset and the nucleus in English and between the nucleus and the coda in Korean. Furthermore, English speakers find learning a game where they have to convert a C_1VC_2 syllable produced by the experimenter into a C_1VC_1VC nonword harder than a game where the same C_1VC_2 stimulus has to be changed into a $C_1VC_2VC_2$ nonword while the opposite is true for Korean speakers. Finally, English speakers judge pairs of CVC stimuli to be more similar when they differ in the nucleus and the coda than when they differ in the nucleus and the onset. The opposite is the case for Korean speakers. Kessler and Treiman (1997) have hypothesized that the rime emerges as a unit in English because the coda is somewhat predictable given the nucleus while the nucleus is not predictable given the onset.

While Yoon and Derwing (2001) have argued that English syllables have an onset-rime structure and Korean syllables have a body-coda structure, the data do not rule out the possibility that both English and Korean have a body-rime syllable structure with bodies having higher resting activation levels than rimes in Korean and rimes having a higher resting activation level than bodies in English.

[1] My thanks go to Joan Bybee, Jill Morford, Caroline Smith and Rena Torres-Cacoullos for helpful comments on earlier versions of this paper, and to Lev Blumenfeld for an interesting discussion of the findings.

If this were the case, both English and Korean syllables would contain overlapping constituents.

To demonstrate the existence of an emergent complex unit, it is necessary to show that speakers of the language are sensitive to 1) the frequencies of co-occurrence and/or transitional probabilities between the smaller units from which the complex unit is purported to emerge; 2) the frequency of the complex unit itself; and 3) the frequencies of co-occurrence and/or transitional probabilities between the purported complex unit (not/in-addition-to its parts) and other units.

The first of these questions has been addressed by Coleman and Pierrehumbert (1997) and Frisch et al. (2000) who have shown that English speakers find words containing rare segments less acceptable than words containing frequent segments with overall acceptability of a word being predicted very well by the product of the frequencies of the segments comprising the word.

The second question has been considered by Cutler et al. (1987, experiment 3) who were presented with visual stimuli of the form CV- or CVC- and asked to produce one multisyllabic word per stimulus. They found that the subjects' completions were more likely to have the form CVCV- than CVCC-. They explained the data by pointing out that there are more multisyllabic words in English that start with CVCV than those that start with CVCC, suggesting that CVCV and CVCC are sublexical phonological units.

However, an alternative explanation is possible. The subjects may have no preference as to whether to produce CVCV or CVCC: as long as the number of CVCV-initial words that start with the stimulus is larger than the number of CVCC-initial words starting with the stimulus, they would still be more likely to come up with a CVCV-initial word than a CVCC-initial word when presented with the stimulus just by chance.

To see whether this is a viable explanation for Cutler et al.'s data, I have conducted a search for the stimuli used by Cutler et al. in the word-initial position in multisyllabic words in the Switchboard Corpus (Godfrey et al. 1992), a spoken corpus of American English and the same corpus used by Cutler et al. (1987) to ascertain that CVCV is more frequent than CVCC in English. For the CVC stimuli, 71% of multisyllabic words containing the stimuli started with CVCV even if compounds were counted and postvocalic preconsonantal /r/ was counted as a consonant rather than as part of the vowel, compared to 78% in experiment 3 of Cutler et al. (1987). For the CV stimuli, when compounds were excluded and /r/ was counted as part of the vowel, 74% of multisyllabic words containing the stimuli started with CVCV, compared to 80% in Cutler et al. (1987).[2] In neither

[2] When compounds were included and /r/ was counted as a codal consonant, 69% of the words beginning with the CV stimuli began with CVCV, a value that is significantly different from the experimental value of 80% at the .05 level. There is some evidence that /r/ should in fact be considered part of the nucleus: it patterns differently from other postvocalic consonants in speech errors (Stemberger 1983), Treiman (1984) found that, when asked to insert a slash into a written syllable, subjects are less likely to insert the slash after the vowel when the following consonant is

case was the difference between the experimental result and the percentage expected by chance statistically significant. Therefore, the results reported by Cutler et al. (1987) do not provide unambiguous evidence that English speakers are sensitive to the frequencies of sublexical phonological units above the segmental level.

Stronger evidence for sensitivity to syllabic constituent frequencies has been provided by Treiman et al. (2000). In a series of experiments, Treiman et al. have shown that native English speakers rate CVC nonwords as more word-like if they have frequent rimes than if they have rare rimes. In addition, when forced to pause inside the stimulus, the speakers are less likely to pause after the vowel if the rime is high-frequency than if it is low-frequency. The frequency of nuclei and codas was controlled in the experiments. However, rime frequency has a correlation of 1 with between-segment transitional probability in the experiment. Therefore, the results could be due to transitional probabilities between segments and not to rime frequency.

In fact, it appears very difficult to manipulate complex unit frequency independently of the frequency of the frequency of its parts and transitional probabilities between them: the frequency of a string is driven up by both the frequencies of its parts and the transitional probabilities between them.[3] Therefore, sensitivity to the frequency of the complex unit alone does not provides unambiguous evidence for the unit being a unit: the potential influence of both part frequency and transitional probability are hard to rule out.

One might object that transitional probability is a poor predictor of chunking and thus should not be considered a confound. One case where transitional probability makes wrong predictions is apparent in the data presented in Bybee (2002: 125) where the fusion of the auxiliary *will* with the preceding (pro)noun, rather than the following verb is shown to be correlated with the fact that the most frequent preceding units occur with *will* more frequently than the most common following units do. Differences in the frequency of co-occurrence in the study are a direct consequence of differences in overall frequencies of the preceding units and the following ones.

---

a liquid than when it is a nasal or an obstruent, Dow (1987) found that mismatches in the nucleus, including mismatches in postvocalic preconsonantal liquids, were more salient than mismatches in the coda.

[3] The only possibility for distinguishing string frequency of a two-part whole from the product of string frequencies of the parts and the transitional probability between them lies in the role of the frequency of the second part, which is multiplied by the frequency of the whole to yield the product of the frequencies of the parts and the transitional probability between them. By coordinating reducing whole frequency with increasing second-part frequency (and vice versa) one could vary whole frequency without varying the product of part frequencies and transitional probability, e.g. the string frequencies of 'I mean' and 'on the' are similar: 20343 and 20387 respectively in the spoken component of the British National Corpus (Leech 1992, frequencies calculated via VIEW, Davies 2005), while the frequency of 'mean' is much smaller than that of 'the' (23468 and 409966 respectively), hence the product of part frequencies and transitional probability is much larger for 'on the' than for 'I mean'.

Because the most frequent units preceding *will*, e.g. *I*, *they*, are more frequent than *will* is, transitional probability is higher after *will* than before *will*, wrongly predicting that *will* should fuse with the following, rather than the preceding unit. A quick search of Switchboard (Godfrey et al. 1992), carried out by the present author, revealed that the transitional probability before *will* is .015 while transitional probability after *will* is .116. Furthermore, transitional probability is a poor predictor of constituent boundaries in preposing languages, like English (Kapatsinski, forthcoming).

However, the problems with transitional probability can be eliminated if 1) transitional probability is counted away from the unit under consideration for the purposes of determining which adjacent unit it will be chunked with, i.e. assuming that the word under consideration will chunk with whichever adjacent word co-occurs more reliably with the word under consideration,[4] and 2) it is counted away from content words for the purposes of constituent structure determination, a proposal supported by evidence from eye tracking, which shows that readers focus on content words and perceive adjacent function words with peripheral vision (cf. Hoffman 1998, Rayner 1998 for reviews). With these modifications, transitional probability is highly correlated with string frequency. Hence, sensitivity to the frequency of a string does not necessarily mean that the string is stored as a unit but may rather mean that the component units are strongly interconnected.

Thus, it is still an open question whether syllables contain 1) syllabic constituents (e.g. Kessler and Treiman 1997, Yoon and Derwing 2001) or 2) simply segments and transitions with variable probability between them (Adams 1981, Seidenberg 1987). Furthermore, if syllables do contain syllabic constituents, it is not clear whether the constituents can exhibit partial overlap, as predicted by exemplar models (Bod 1992, 1998, Skousen 1989, 2002a).

It is only by looking at whether speaker/hearers are sensitive to the co-occurrence relations between the proposed complex unit and other units that we can unambiguously determine whether the unit exists or not because co-occurrence relations between a complex unit and other units can be dissociated from the frequency of the unit and from the co-occurrence relations of the unit's parts. This question has not been addressed in the literature at all, and it is this question we will consider in this study.

In particular, we examine the question of whether a nonce Russian verb bearing the stem-extending suffix *-i-* or *-a-* is more acceptable 1) if its rime co-occurs with the stem extension it is bearing than if its rime co-occurs with the competing stem extension and 2) if its body co-occurs with the stem extension it is bearing than if its body co-occurs with the competing stem extension. If both rimes and bodies are found to have an effect within a single set of stimuli within a

---

[4] In this case, transitional probability would have a perfect correlation with two-word string frequency.

single language, this would provide support for the existence of overlapping units and thus for an exemplar model of memory.

Stem extensions in Russian derive a verbal stem from a nominal root. For instance, *mot* 'a spendthrift' → *motat*^{j} 'to waste', *svet* 'light' → *svetit*^{j} 'to shine', *pot* 'sweat' → *pot*^{j}*et*^{j} 'to sweat', *ʃtraf* 'a fine' → *ʃtrafovat*^{j} 'to fine' where -*t*^{j} is an infinitival marker.

The reasons Russian verbal stem extensions -*i*- and -*a*- provide a good domain to test our hypothesis are that 1) there are two competing stem extensions that are approximately equal in both type and token frequency: Kapatsinski (2005) has shown that each stem extension accounts for about a third of the types in the reverse dictionary of Russian (Zaliznjak 1977) and a third of the tokens in the modern, 7,000,000-word Ogonek Corpus (SFB-441 2000); 2) the stem extensions are approximately equal in productivity and are much more productive than the other stem extensions (-*ova*-, -*eva*-, -*nu*-, -*e*-, among others), as shown by both elicited production and acceptability ratings (Kapatsinski 2005, 2005b, forthcoming b).

2 Methods

For every possible CV body and every possible VC rime in Russian, all verbs with monosyllabic roots containing the rime or body within the root and listed in the 125,000-word reverse dictionary of Russian (Zaliznjak 1977) were noted. For both rimes and bodies, three sets were created: constituents that overwhelmingly occurred with -*i*-, those that overwhelmingly occurred with -*a*-, and those that did not occur in monosyllabic verbs. The absolute frequency of occurrence in monosyllabic verbal roots was controlled for bodies when testing the effect of body associations and for the rimes when testing for the effect of rime associations.

Nine stimulus categories were created: 1) -*a*-associated body, -*a*-associated rime, 2) -*a*-associated body, -*i*-associated rime, 3) -*a*-associated body, non-associated rime, 4) -*i*-associated body, -*a*-associated rime, 5) -*i*-associated body, -*i*-associated rime, 6) -*i*-associated body, non-associated rime, 7) non-associated body, -*a*-associated rime, 8) non-associated body, -*i*-associated rime, and 9) non-associated body, non-associated rime. For instance, /la/ occurs in 6 bisyllabic verbs bearing -*a*- (*labat*^{j}, *lakat*^{j}, *lajat*^{j}, *latat*^{j}, *laskat*^{j}, *lapat*^{j}) and only 2 bisyllabic verbs bearing -*i*- (*ladit*^{j}, *lazit*^{j}) and is therefore labeled as co-occurring with -*a*-. On the other hand, /al/ occurs only in verbs bearing -*i*- (*valit*^{j}, *ʒalit*^{j}, *xvalit*^{j}, *palit*^{j}, *ʃalit*^{j}) and is therefore labeled as co-occurring with -*i*-. Therefore, in the stimulus root *lal*, the body co-occurs with -*a*- while the rime co-occurs with -*i*-. Frequency estimates and co-occurrence statistics were type-frequency-based and were derived from Zaliznjak (1977) and the Ogonek Corpus (SFB 441, 2000).

Co-occurrence statistics were also collected for the phonemes comprising the syllables from the same sources. The statistics were based on the form of the

roots inside verbs rather than when they occurred independently as nouns. The reason this was done is that there is much support for the idea that speakers use product-oriented and not source-oriented generalizations (see Wang and Derwing 1994, Bybee 2001). Furthermore, doing this allows for more robust generalizations because there are more verbs than there are nouns which have served as bases for verbs.[5] In addition, measuring type frequency in a source-oriented manner makes the type frequency of *-a-* relatively low, failing to account for its high productivity (Kapatsinski 2005b: chapter 8).

Thirteen adult native Russian speakers participated in the experiment. Nonsense nouns followed by a verb formed from it via either *-i-*, *-a-*, *-ova-*, or *-eva-* were presented to them. They were asked to rate the pairs on a 10 point scale ranging from 1=”nobody would form this verb from this noun” to 10=”everyone would form this exact verb from this noun”. Subjects were asked to imagine that the nouns are recent borrowings and they need a verb to describe an event involving the noun.

The stimuli were presented in a pseudorandomized order with no stimuli that had a body or rime favoring the same stem extension occurring in a row and no stimuli with the same rime or body occurring within five stimuli of each other. Half of the non-associated stimuli were stressed on the first syllable and half on the second. In unambiguously associated stimuli, stress was assigned to match the majority of the associates. With ambiguously associated stimuli, i.e. stimuli in which the body is associated with a different extension than the rime, two differently stressed variants were used if the *-a*-bearing associates and the *-i*-bearing associates differed in typical stress location. For analyzing the impact of the body, stimuli stressed where words sharing the body with it were stressed were used. For analyzing the impact of the rime, stimuli stressed where words sharing the rime with them were stressed were used. We eliminated stimuli where the body-based gangs and rime-based gangs required different stress, e.g. *tsodatʲ*: *ˈtsokatʲ* vs. *xo ˈdiтʲ*, *bro ˈdiтʲ*, *ro ˈdiтʲ*, *vo ˈdiтʲ*, *plo ˈdiтʲ*, *go ˈdiтʲ*. We also eliminated potential stimuli that contained phonotactically illegal sequences, e.g. /*gɨ/. The stimuli are presented in table 1.

[5] Note that as long as generalizations are allowed to compete and are rewarded for having a high type frequency, as in the Rule-based Learner (Albright and Hayes 2003), product-oriented generalizations will outcompete source-oriented ones.

Table 1. The full set of CVC nonce roots whose bodies and rimes exhibit reliable associations with -i- or -a- or do not occur in monosyllabic verbal roots.

Body co-occurs with	Rime co-occurs with							
	-a-		*-i-*			Does not occur	
-a-	map matʃ tim pim tit	tiʃʲ tsoʃʲ rɨt sɨt	mal mam lan	ʒub ʒuz ʒud	sɨb lar lal	tsof ʒuf rɨʃʲ	tʲuf rɨn rɨf
-i-	ʃʲap	ʃʲatʃ xoʃʲ	ʃʲal ʃʲam ʃʲar xoʃ xoz	tud tul dʲuz dʲur xotʃ	dʲud dʲul dʲuʃ xob	xof	dʲuf tuf
Does not occur	tʃap tʃatʃ sʲap sʲatʃ dʲap dʲatʃ	ʃʲev rʲoʃʲ bʲap bʲatʃ xʲev	tʃam sʲal sʲam sʲan sʲar dʲan	bʲam bʲan bʲar gʲes xʲes	rʲud sʲub dʲal dʲar dʲam	bʲuf	rʲuf sʲuf

3 Results and Discussion

Table 2 shows that when the frequency of the body is controlled verbs bearing *-i-* receive higher ratings if their root's body co-occurs with *-i-* than when it co-occurs with *-a-*, indicating that the body is a unit in Russian. Similarly, a verb bearing *-a-* is judged to be more natural if the body of its root co-occurs with *-a-* than if it co-occurs with *-i-*. The influence of the body is statistically significant when tested in an ANOVA that also includes respondent identity and rime's preference (p=0.022). Importantly, the difference cannot be due to differences in the frequencies of bodies associated with *-i-* and those associated with *-a-* because they were exactly the same and the bodies that favor *-i-* disfavor *-a-* and vice versa.

Table 2. Body influence on naturalness judgments

Body co-occurs with	Score for -i-bearing verbs	Score for -a-bearing verbs
-i-	5.08	5.01
-a-	4.85	5.27
Significance	p=0.022	

The results in Table 2 might be taken to imply that Russian is a body language, like Korean. However, Table 3 shows that this interpretation is incorrect because both the co-occurrence relations between root body and the suffix and those

between root rime and the suffix play a role in naturalness judgments. Verbs bearing *-i-* were judged to be more natural if the rime of their root co-occurred with *-i-* than if it co-occurred with *-a-*. Similarly, a verb bearing *-a-* is judged to sound more natural if its root's rime co-occurs with *-a-* than if it co-occurs with *-i-* ($p<0.0005$ in an ANOVA that also included body preference and respondent identity). Rime frequency was controlled in this analysis.

Table 3. Rime influence on naturalness judgments

Rime associated with	Score for -i-bearing verbs	Score for -a-bearing verbs
-i-	5.31	4.81
-a-	4.76	5.16
Significance	$p<0.0005$	

It is also possible a priori that the associations we investigate are really on the segmental level. However, neither the initial consonant, nor the vowel are reliable predictors of the differences between stimuli whose bodies favor *-a-* and those whose bodies favor *-i-*. The final consonant is a reliable predictor (in C_1VC_2, $p=.711$ for C1, $p=.234$ for V, $p=.039$ for C2) but its predictions are in the wrong direction (the mean score is 4.97 for the stimuli it disfavors and 4.7 for those it favors). This indicates that the factor is simply significant by virtue of a negative correlation with a real determinant. This conclusion is supported by the fact that including C2 in the same model with body and rime reduces C2 to insignificance while body and rime stay significant ($p=0.014$ for body, $p=0.002$ for rime, $p=0.171$ for C2). There is a significant interaction between C2, the body, and the rime ($p<0.0005$), which explains why C2 has a significant but negative correlation. The differences predicted by rime's associations (Table 3) are not predicted by any of the segmental variables ($p=0.309$ for C1, $p=0.164$ for V, $p=0.141$ for C2).

Finally, the possibility that the effects are due to the transitional probabilities between the final consonant of the root and the stem extension must be ruled out. Table 4 shows that transitional probabilities cannot predict the results because *-i-* is in fact more probable, given the root-final consonant, when the rime favors *-a-* than when the rime favors *-i-* while the difference between transitional probabilities in body-based conditions is not reliable. The calculations are in terms of token frequency due to the lack of machine-readable dictionaries of Russian and are based on the part of the Uppsala Corpus containing literary (as opposed to press) texts (Loenngren, 1993).

Table 4. Transitional probability of -a- divided by the transitional probability of -i- given the root-final consonant by experimental condition

The ratio of transitional probability of /a/ divided by that of /i/ given the preceding consonant when the constituent listed below is associated with the stem extension listed to the right	*-a-*	*-i-*
Rime	1.27	2.60
Body	1.11	1.14

Thus, overlapping units are necessary to describe the data. The existence of overlapping units is predicted only by exemplar models of lexical representation (Bod 1992, 1998, Skousen 1989, 2002a) and thus lends them strong support.[6]

4 Implications for studies of morphological productivity

The finding that bodies can form associations with suffixes contradicts a central assumption of the Rule-Based Learner (Albright and Hayes 2002, 2003). The model relies on a particular similarity metric to determine which words in the lexicon are similar to the nonce stimulus. The model needs to make this determination to decide which suffix to add to the stimulus and to predict naturalness ratings for stimuli bearing the suffix. The metric first compares the final phonemes of the root of the stimulus and the roots of all the words in the lexicon and then proceeds backwards through the stimulus root, comparing phonemes one-by-one. If at any point there is a mismatch, the comparison process stops and does not compare any phonemes further away from the root-suffix boundary than the nearest mismatched phoneme. An illustration is provided in Table 5.

[6] As suggested to me by Lev Blumenfeld (p.c.), it is possible that the 'rimes' are really crosssyllabic units. That is, the rimes are rimes if the syllable boundary in the verb falls at the root boundary but sequences of nucleus of the first syllable and the onset of the following syllable if onset maximization applies regardless of the root boundary. Derwing (1992) found that, when forced to insert a pause in a C(C)(C)VCV word, English speakers produce C(C)(C)V.CV much more often than C(C)(C)VC.V but only if there is no morpheme boundary after the last consonant. If there is a morpheme boundary there, C(C)(C)VC.V is much more frequent than C(C)(C)V.CV, suggesting that syllables end at morpheme boundaries and hence that our rimes really are rimes. More work is needed on this issue but the existence of overlapping constituents would be supported either way.

*Table 5. Relative similarity of English verbs to the nonce stimulus /*twɪnk/ *determined by the Rule-Based Learner (Albright and Hayes 2002, 2003)*[7]*. Words in bold share the beginning but not the end with the stimulus.*

4 phonemes shared	3 phonemes shared	2 phonemes shared	1 phoneme shared	Nothing shared
wink	sink	bank	bake	match
	drink	thank	take	agglutinate
	blink	spank	trick	veer
	link	rank	lick	**twist**
	think	crank	**tweak**	**twich**
	stink	chunk	stick	**win**
	shrink	debunk	poke	**twinkle**
/(Y)Xwink/	/(Y)Xɪnk/	/(Y)Xnk/	/(Y)Xk/	/(Y)X/

Some evidence contradicting this hypothesis is provided by Skousen (2002b:31-34) who found that the past tense of the rare Finnish verb *sorta* 'to oppress' is *sorti*, despite the fact that verbs ending in /rta/ overwhelmingly favor changing the /t/ to /s/, thus voting for *sorsi*. Skousen suggests that the reason *sorti* is preferred by the speakers is that there is a very large number of verbs that have the vowel /o/, none of which display the /t/-/s/ alternation. Skousen was able to show that the outcome is predicted by the Analogical Modeling algorithm. However, given that *sorta* is an existing verb, it is also possible that the subjects just remembered the past tense form associated with it. As Skousen (2002b:37-39) points out, the ability of the model to predict the forms of existing verbs is simply measuring the degree of regularity in that area of the lexicon.

As shown in table 4, in the model proposed by Albright and Hayes (2002, 2003), sharing the body does not increase the similarity between the stimulus and a word in the lexicon for the purposes of suffix assignment, as long as the rime is not shared. Therefore, how frequently the stimulus's body co-occurs with a particular suffix should make no difference to how acceptable the stimulus sounds when bearing the suffix, a prediction that is disconfirmed by our results.

5 Conclusion

This paper has introduced a new way to test the psychological reality of complex units. We have argued that complex unit frequency is inherently confounded with frequencies of the component units and the transitional probabilities between them, especially if transitional probability is measured in a way consistent with

[7] Here I am omitting featural similarity between the first mismatched phonemes for the sake of simplicity. X is any phoneme other than the one present in that position in the nonce stimulus, and Y is any phoneme, including phonemes that are in the nonce stimulus.

the evidence from eye tracking and corpus data on syntactic constituent formation. Thus, sensitivity to the frequency of a complex unit is always open to interpretation in terms of sensitivity to lower-level units. However, there is an unambiguous indicator of unithood for complex units. This indicator is subjects' sensitivity to the co-occurrence relations (whether defined in terms of string frequency or transitional probability) between the complex unit and other units.

Using this new method of testing unithood, we have examined the question of whether bodies and rimes of Russian verbal roots are associated with stem extensions with which they frequently co-occur. We found reliable correlations between acceptability of a nonce verb and whether or not the body and the rime of its root usually co-occur with its stem extension. Thus, support for both bodies and rimes as units was found. This finding is consistent with proposals made by Vennemann (1988) and Kessler and Treiman (1997) that nuclei are connected to both onsets and codas with connections of varying strength. It is also consistent with data presented by Treiman et al. (2000), who found that English speakers are sensitive to frequencies of both rimes and bodies. However, unlike the model proposed by Vennemann (1988), and in accordance with the proposal by Skousen (1989, 2002a), the model proposed here posits higher-level body and rime units in addition to the trainable connections between onsets, nuclei and codas.[8] The existence of these higher-level units allows trainable connections to exist between those units and other nodes in the network.

The results are consistent with findings reported by Bod (1998:51-68) that including overlapping syntactic constituents in a model of the speaker/hearer's syntacticon, i.e. his/her memory of the syntactic structures s/he has encountered, leads to improved parsing of not previously encountered sentences, compared to a model that does not store overlapping constituents. The model that stored overlapping constituents achieved 85% parsing accuracy compared to 69% accuracy achieved by the model that did not store overlapping constituents, where correct parses were defined as being parses equivalent to those produced by native speakers who were not aware of the models' predictions.

The finding that bodies can form associations with suffixes is inconsistent with the similarity metric proposed by Albright and Hayes (2002, 2003). However, the findings are fully consistent with exemplar models, showing both the existence of connections based on co-occurrence (e.g. Bybee and Scheibman 1999, Bybee 2001, 2002, Kapatsinski forthcoming c) and the emergence of partially overlapping constituents (Bod 1992, 1998, Skousen 1989, 2002a). A prediction that has been made by both Bod and Skousen and one that we have not yet tested is whether constituents could also be non-continuous with, for instance, the onset and the coda fusing into a unit. This issue remains for future research.

---

[8] Skousen's model is not formalized in terms of a network. It is clear that units (what Skousen calls contexts and supracontexts) can form associations with other units, e.g. a supracontext may be associated with a particular morpheme. It is not clear, however, whether (supra)contexts are automatically associated with any (supra)contexts they co-occur with.

References

Adams, M. 1981. What good is orthographic redundancy? In *Perception of Print*, edited by H. Singer and O. J. L. Tseng, 197-221. Hillsdale, NJ: Erlbaum.

Albright, A., and B. Hayes. 2002. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the 6th Meeting of the ACL Special Interest Group in Computational Phonology*, edited by M. Maxwell. Philadelphia: ACL.

Albright, A., and B. Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. Cognition 90: 119-61.

Bod, R. 1992. A computational model of language performance: Data oriented parsing. *Proceedings of COLING '92*.

Bod, R. 1998. *Beyond Grammar: An Experience-Based Theory of Language*. Stanford, CA: CSLI.

Bybee, J. L. 2001. *Phonology and Language Use*. Cambridge: Cambridge University Press.

Bybee, J. L. 2002. Sequentiality as the basis of constituent structure. In *The Evolution of Language out of Pre-language*, edited by T. Givon and B. F. Malle. Amsterdam, Philadelphia: John Benjamins.

Bybee, J. L., and J. Scheibman. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in English. Linguistics 37: 575-96.

Coleman, J., and J. B. Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. Computational Phonology 3: 49-56.

Cutler, A., D. Norris, and J. Williams. 1987. A note on the role of phonological expectations in speech segmentation. Journal of Memory and Language 26: 480-7.

Davies, M. 2005. VIEW: Variation In English Words and phrases. http://view.byu.edu

Derwing, B. L. 1992. A "pause-break" task for eliciting syllable boundary judgments from literate and illiterate speakers: Preliminary results from five diverse languages. Language and Speech 35: 219-35.

Dow, M. L. 1987. *On the Psychological Reality of Sub-syllabic Units.* Ph.D. Thesis: U of Alberta.

Frisch, S. A., N. R. Large, and D. B. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. Journal of Memory and Language 42: 481-96.

Godfrey, J. J., E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. IEEE ICASSP: I517-20.

Hoffman, J. E. 1998. Visual attention and eye movements. In *Attention*, edited by H. Pashler, 119-53. San Diego: Psychology Press.

Kapatsinski, V. M. 2005. Characteristics of a rule-based default are dissociable: Evidence against the Dual Mechanism Model. In *Formal approaches to Slavic linguistics 13: The South Carolina Meeting*, edited by S. Franks, F. Y. Gladney, and M. Tasseva-Kurktchieva, 136-46. Ann Arbor, MI: Michigan Slavic Publications.

Kapatsinski, V. M. 2005b. *Productivity of Russian Stem Extensions: Evidence for and a Formalization of Network Theory*. M.A. Thesis: University of New Mexico.

Kapatsinski, V. M. Forthcoming. Measuring the relationship of structure to use: The determinants of recycle in repetition repair. BLS 30.

Kapatsinski, V. M. Forthcoming, b. To scheme or to rule: Contra the Dual Mechanism Model. BLS 31.

Kapatsinski, V. M. Forthcoming, c. Frequency, Neighborhood Density, Age-of-Acquisition, and Lexicon Size Effects in Priming, Recognition and Associative Learning: Towards a Single-Mechanism Account. HDLS VI.

Kessler, B., and R. Treiman. 1997. Syllable structure and the distribution of phonemes in English syllables. Journal of Memory and Language 37: 295-311.

Leech, G. 1992. 100 million words of English: the British National Corpus. Language Research 28: 1-13.

Loenngren, L. 1993. *A Frequency Dictionary of Modern Russian. With a Summary in English.* Uppsala Universitet: Uppsala, Sweden. Corpus available at http://heckel.sfb.uni-tuebingen.de/cgi-bin/cqp.pl.

Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin 124: 372-422.

Seidenberg, M. S. 1987. Sublexical structures in visual word recognition: access units or orthographic redundancy? In *Attention and Performance XII: Reading*, edited by M. Coltheart. London: Erlbaum.

SFB 441, Project B1. 2000. *Ogonek 1996-2000*. http://heckel.sfb.uni-tuebingen.de/cgi-bin/cqp.pl.

Skousen, R. 1989. *Analogical Modeling of Language*. Dordrecht: Kluwer.

Skousen, R. 2002a. An overview of Analogical Modeling. In *Analogical Modeling: An Exemplar-based Approach to Language*, edited by R. Skousen, D. Lonsdale, and D. B. Parkinson, 11-26. Amsterdam, Philadelphia: John Benjamins.

Skousen, R. 2002b. Issues in Analogical Modeling. In *Analogical Modeling: An Exemplar-based Approach to Language*, edited by R. Skousen, D. Lonsdale, and D. B. Parkinson, 27-50. Amsterdam, Philadelphia: John Benjamins.

Stemberger, J. P. 1983. *Speech Errors and Theoretical Phonology*. Bloomington, IN: IU Linguistics Club.

Treiman, R. 1984. On the status of final consonant clusters in English syllables. Journal of Verbal Learning and Verbal Behavior 23: 343-56.

Treiman, R., B. Kessler, S. Knewasser, R. Tincoff, and M. Bowman. 2000. English speakers' sensitivity to phonotactic patterns. In *Laboratory Phonology V: Acquisition and the Lexicon*, edited by M. B. Broe and J. B. Pierrehumbert, 269-82. London: Cambridge University Press.

Vennemann, T. 1988. The rule dependence of syllable structure. In *On language: Rhetorica, Phonologica, Syntactica: A Festschrift for Robert P. Stockwell from his Friends and Colleagues*, edited by C. Duncan-Rose and T. Vennemann, 257-83. London: Routledge.

Wang, H. S., and B. L. Derwing. 1994. Some vowel schemas in three English morphological classes: Experimental evidence. In *In Honor of Professor William S.-Y. Wang: Interdisciplinary Studies on Language and Language Change*, edited by M. Y. Chen and O. C. L. Tseng, 561-75. Taipei: Pyramid Press.

Yoon, Y. B., and B. L. Derwing. 2001. A language without a rhyme: Syllable structure experiments in Korean. Canadian Journal of Linguistics 46: 187-237.

Zaliznjak, A. A. 1977. *Grammaticheskij Sovar' Russkogo Jazyka: Slovoizmenenie*. Moscow: Russkij Jazyk.