Grouping morphologically complex words in the mental lexicon: Evidence from Russian verbs and nouns

Natalia Slioussar
HSE Moscow & St. Petersburg State University
slioussar@gmail.com

Anastasia Chuprina
St. Petersburg State University
a.o.chuprina@gmail.com

1 Introduction

Frequency is known to play a crucial role in lexical access. The notions primarily discussed in the literature are form frequency, (whole) word frequency and morpheme frequency, e.g. root frequency. In numerous studies (Alegre & Gordon, 1999; Baayen & al. 2007, a.m.o.), these characteristics were manipulated to find out whether various word forms are decomposed during lexical access or are stored and can be accessed as a whole. Similar issues arise when we turn from inflection to derivation, at least with semantically transparent derivates (Niswander-Klement & Pollatsek, 2006; Taft 2004, a.m.o.).

2 Our study

Some morphologically complex words were shown to be accessed as a whole (then their own frequency played a crucial role), the others were demonstrated to be decomposed (then root frequency and the frequency of the word they are derived from was important). Both options are available in some models: the one that is more efficient in a particular case wins. However, the picture may be more complex in morphologically rich languages. If a word has many inflectional forms or derivates that are stored as a whole, they probably form groups, and lexical access to this word may depend on the properties of such groups. Our hypothesis is that if a word has a large group of morphologically complex derivates which are relatively semantically transparent, access to and storage of this word would depend on the properties of this group even though the derivates do not necessarily undergo the process of decomposition. We explored this question in our study on Russian.

2.1 Experiment 1

Method. We conducted a lexical decision experiment using *E-Prime* software. Participants were 27 speakers of Russian (age: 19-52 years, 20 female). Materials were 18 triplets of unprefixed imperfective verbs and 12 pairs of unprefixed deverbal nouns. Word frequency, length and CV structure were matched inside triplets and pairs, while the summed frequency of the corresponding prefixed verbs and nouns was different for every verb and noun inside a triplet/pair (as shown in Table 1). Word frequency information was taken from the *The Frequency Dictionary of the Modern Russian Language* (Lyashevskaya & Sharoff, 2009).

word	*letters (in Cyrillic)*	*word F (ipm)*	*summed F of prefixed words*	*group*
torčat' to stick out	7	86,3	2,0	1
dyšat' to breath	6	90,8	29,4	2
platit' to pay	7	89,0	86,3	3
roždenie birth	8	98,5	35,8	1
javlenie apparition	7	94,3	297,5	2

Table 1. An example of stimuli for Exp.1.

It is important to note that prefixed verbs are derived from unprefixed ones, while prefixed deverbal nouns are not (they are derived from prefixed verbs). For verbs, we also counted derivates with the reflexive postfix *-sja*. We made a simplification not taking suffixes into account because, firstly, suffixes change the inflectional class the word belongs to and often cause stress shifts and various alternations, and, secondly, most unprefixed verbs have dramatically more derivates created by prefixation than by suffixation.

In total, every participant saw 54 verbs in infinitive and 24 nouns in nominative singular form, and 78 nonce stimuli. They were shown on

In Vito Pirrelli, Claudia Marzi, Marcello Ferro (eds.): *Word Structure and Word Usage*. Proceedings of the NetWordS Final Conference, Pisa, March 30-April 1, 2015, published at http://ceur-ws.org

the computer screen for 500 ms or until a response button was pressed. If no button was pressed, participants saw a blank screen for up to 2 s. After a response was given or after these 2,5 s were over, an interstimulus interval was initiated and then the next trial began.

Results and discussion. We analyzed participants' question-answering accuracy and reaction times. All participants gave at least 85% of correct answers (92,4% on average); trials with incorrect answers were excluded from further analysis. We also discarded all RTs that exceeded 1,5 s, as is customary in many such studies (e.g. Alegre & Gordon, 1999). In total, 0,3% of reactions to real stimuli were discarded.

We demonstrated that RTs for verbs differ significantly depending on the summed frequency of corresponding prefixed (and postfixed) verbs (repeated measures ANOVA, $F1(2,52) = 8,66$, $p = 0,001$, $F2(2,34) = 4,99$, $p = 0,013$), but RTs for nouns do not. Average RTs for different groups of verbs and nouns are given in Tables 2a and 2b.

group	*av. F (ipm)*	*av. summed F of prefixed words*	*av. RT (ms)*
1	40,1	11,0	643,4
2	41,1	43,5	632,3
3	41,1	139,0	607,6

Table 2a. Average RTs for verb stimuli in Exp.1.

group	*av. F (ipm)*	*av. summed F of prefixed words*	*av. RT (ms)*
1	33,1	60,3	637,8
2	31,9	220,6	635,4

Table 2b. Average RTs for noun stimuli in Exp.1.

We believe that these results can be explained as follows. The majority of Russian prefixed verbs and nouns are likely to be stored as a whole because even relatively transparent ones tend to have some aspects of meaning that cannot be predicted compositionally. Still, prefixed verbs have close connections with their unprefixed counterpart in the mental lexicon due to direct derivational links and therefore influence lexical access to it. Prefixed deverbal nouns are not connected to their unprefixed counterpart in a similar way due to the lack of derivational links, so the summed frequency of such nouns does not influence lexical access to it.

However, an alternative explanation can also be suggested: prefixed verbs are decomposed (and thus boost the frequency of their unprefixed counterpart), while the results for nouns are inconclusive. We chose deverbal nouns for our experiment to find enough relatively transparent prefixed and unprefixed ones, and, if prefixed ones are decomposed, the system should go to the prefixed verb by stripping the suffix rather than to the unprefixed noun by stripping the prefix (*rodit'(v)* → *porodit'(v)* → *porožd<u>enie</u>(n)*). To refute this alternative explanation, we designed a follow-up experiment.

2.2 Experiment 2

Method. The method was the same as in Experiment 1. Participants were 24 speakers of Russian (age: 18-55 years, 18 female). Materials included 60 prefixed verb and noun stimuli and 60 nonce words. Real words were chosen from the pool of prefixed verbs and nouns whose unprefixed counterparts were analyzed in Experiment 1. This time both verbs and nouns were grouped in pairs. They were matched in length, CV structure and the frequency of their unprefixed counterparts, but differed in whole word frequency. An example is given in Table 3.

word	*letters (in Cyrillic)*	*word F (ipm)*	*unprefixed word F*	*group*
podyšat' to breath a little	8	7,7	90,8	1
otplatit' to pay back	9	1,7	89,0	2
porožd<u>e</u>nie production	10	5,1	98,5	1
projavl<u>e</u>nie manifestation	10	45,3	94,3	2

Table 3. An example of stimuli for Exp.2.

Moreover, we took care of the following. If verbs like *podyšat'* 'to breath a little' and *otplatit'* 'to pay back' from Table 3 are accessed as a whole, their word frequency should matter, and *podyšat'* (group 1) will be accessed faster. Now let us assume that they are decomposed, and so are many other prefixed verbs. Then not the word frequency of *dyšat'* 'to breath' and *platit'* 'to pay' will predict the speed of the lexical access, but the frequency of these unprefixed verbs plus the summed frequency of their decomposed derivates. As Table 1 shows, this value is greater for *platit'* than for *dyšat'*, so *otplatit'* (group 2) will be accessed faster. This was true for all other prefixed verb pairs in Experiment 2, so the whole word access and decomposition scenarios always gave different predictions.

We could not find prefixed noun pairs with a similar distribution of frequencies in our materials. However, no approach would predict that they could be decomposed by stripping off their prefix first anyway. So noun stimuli were included mainly to make experimental materials more diverse, they will not let us tease apart different lexical access scenarios.

Results and discussion. We analyzed participants' question-answering accuracy and reaction times. All participants gave at least 85% of correct answers (92,0% on average); trials with incorrect answers were excluded from further analysis. We also discarded all RTs that exceeded 1,5 s. In total, 0,4% of reactions to real stimuli were discarded.

We demonstrated that this time, RTs for verbs and nouns differed depending on their whole word frequencies. The difference was statistically significant both for prefixed verbs (RM ANOVA, $F1(1,23) = 17,87$, $p < 0,001$, $F2(1,17) = 5,98$, $p = 0,026$) and for prefixed nouns ($F1(1,23) = 21,27$, $p < 0,001$, $F2(1,11) = 7,88$, $p = 0,017$). Average RTs for different groups are shown in Tables 4a and 4b.

group	*av. F*	*corresp. unpref. verb from Exp.1*	*av. RT*
1	16,3	low summed F	707,3
2	2,0	high summed F	746,0

Table 4a. Average RTs for verb stimuli in Exp.2.

group	*av. F*	*corresp. unpref. noun from Exp.1*	*av. RT*
1	12,4	low summed F	688,4
2	76,4	high summed F	657,5

Table 4b. Average RTs for noun stimuli in Exp.2.

The results are indicative of the whole word lexical access. We can conclude that prefixed verbs influence lexical access to their unprefixed counterpart not through decomposition, but because they are closely connected in the mental lexicon due to direct derivational links.

3 Conclusion

Using Russian prefixed and unprefixed verbs, we demonstrated that a group of semantically transparent derivates influence the recognition of the word they are derived from. The higher is summed frequency of derivates, the faster is the lexical access. One could argue that this is due to decomposition.We showed that this is not the case.

In two lexical decision experiments we conducted, reactions times to prefixed verbs and deverbal nouns depended on their own frequencies, which points to whole word storage. At the same time, reaction times to unprefixed verbs were influenced by the summed frequency of their derivates (created by prefixation and postfixation). We conclude that this effect is explained not by decomposition of the derivates during lexical access, but by their strong connection to the word they are derived from.

Our conclusion is confirmed by the data from deverbal nouns. On the surface (i.e. phonologically), the overlap between unprefixed and prefixed verbs on the one hand and unprefixed and prefixed nouns on the other hand is the same: as examples from Tables 1 and 3 show, they coincide once the prefix is stripped. If this factor played a role, the results for unprefixed verbs and nouns would be the same.

However, reactions times to unprefixed nouns are not influenced by the summed frequency of their prefixed counterparts. This proves that connections through derivational links matter. Prefixed deverbal nouns are derived from prefixed verbs, not from unprefixed nouns (*porodit'(v)* 'to give birth, to generate' → *porožděnie(n)* 'production', not *rožděnie(n)* 'birth' → *porožděnie(n)* 'production'). Phonologically, prefixed nouns resemble unprefixed ones much more than prefixed verbs, but this does not play a role.

In total, our results can be taken as a piece of evidence for a new type of frequency information to be taken into account. Somewhat similar conclusions were reached by Moscoso del Prado Martín et al. (2004) who studied morphological family size effects in Finnish compared to Dutch and Hebrew.

Of course, many things remain to be explored. As we noted earlier, we did not look at suffixation. We did not specify the mechanisms by which derivationally related forms are connected in the mental lexicon and how these connections are formed. In the connectionist approach where no decomposition is assumed, regular connections between words' phonological forms and meanings should matter. In dual route models, it can be suggested that decomposition normally does not win in some cases like derived verbs and nouns we analyzed, but still takes place. Then only the existence of a direct derivational link and, probably, semantic transparency should really matter.

To solve these and other problems, many crucial questions need to be answered. Which deri-

vates ‘boost’ the frequency of a base word and to what extent? What is the role of semantic transparency and phonological similarity between a derivate and its base form? How important is it for their connection whether they belong to one part of speech or to one inflectional class? Would stress shifts and alternations influence our results? We hope to address some of these questions in our further research.

References

Alegre, M., & Gordon, P. 1999. Frequency effects and the representational status of regular inflections. Journal of Memory and Language, 40, 41-61.

Baayen, H., Wurm, L.H., & Aycock, J. 2007. Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. The Mental Lexicon, 2, 419-463.

Lyashevskaya, O.N., & Sharoff, S.A. 2009. Chastotnyj Slovar’ Sovremennogo Russkogo Jazyka (‘The Frequency Dictionary of Modern Russian Language’). Moscow: Azbukovnik.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R., & Baayen, R.H. 2004. Morphological family size in a morphologically rich language: The case of Finnish compared to Dutch and Hebrew. Journal of Experimental Psychology: Learning, Memory and Cognition, 30, 1271–1278.

Niswander-Klement, E., & Pollatsek, A. 2006. The effects of root frequency, word frequency, and length on the processing of prefixed English words during reading. Memory and Cognition, 34, 685-702.

Taft, M. 2004. Morphological decomposition and the reverse base frequency effect. Quarterly Journal of Experimental Psychology, 57A, 745-765.