What is it I am writing? Lexical frequency effects in spelling Russian prefixes: Uncertainty and competition in an apparently regular system*

VSEVOLOD KAPATSINSKII

Abstract

Whole-word frequency effects are shown to exist in what appears to be a completely regular system, the spelling of prefix-final /z/ in Russian. Russian prefixes that underlyingly end in /z/ (roz-, bez-, iz-) end in /s/ on the surface when followed by a voiceless consonant. According to the rules of Russian orthography, the surface form, rather than the underlying form, must be reflected in the spelling. However, spelling errors reflecting the underlying form often occur, especially for the prefix bez-. The present paper reports that the error rate, either in natural typing on the web or in a classroom dictation task, for a given word is negatively correlated with the frequency of the word, suggesting that Russian writers rely, to a significant extent, on memory of complete orthographic forms as opposed to the orthographic rule. The frequency effect holds even within the set of regular inflectional variants of a single lexeme, with more frequent wordforms showing lower error rates. The evidence demonstrates a high degree of reliance on whole-form lexical retrieval even in what appears to be a regular system that is explicitly taught to the writers throughout their schooling in a morphologically rich language and thus provides support for the use of lexical retrieval even when it is not necessary (Baayen et al. 2002, Butterworth 1983, Bybee 1985, 2000, vs. DiSciullo and Williams 1987, Pinker 1991). However, reliance on retrieval is argued to be especially strong when there is a relatively long period of temporary uncertainty regarding which rule is applicable during processing (see also Albright 2009, Barca et al. 2007, Burani et al. 2006). The importance of temporary uncertainty and resulting rule competition suggests that the regular|irregular distinction needs to be reconsidered as even fully “regular” systems may feature rule competition due to temporary uncertainties about rule applicability. Reliance on retrieval may go largely undetected in Russian during schooling because teaching and test materials focus on the spelling of frequent words, which can be either produced by rule or retrieved as wholes. The largely complementary methodological challenges in studying lexical frequency effects in corpus and experimental data are discussed.
Keywords: word frequency; lexical retrieval; grammatical computation; regularity; productivity; morphology; orthography; language production; errors; Russian

1. Introduction

1.1 Theoretical background

The balance between memory retrieval of complex expressions and online computation in language production has been a topic of much discussion and debate in both linguistics and psycholinguistics. One influential approach has held that retrieval is a backup plan, used only if the to-be-produced form is grammatically irregular and thus cannot be derived by the productive rules of the grammar (e.g., DiSciullo and Williams 1987, Pinker 1991). As DiSciullo and Williams (1987: 3) put it, “the lexicon is like a prison – it contains only the lawless”. An alternative hypothesis is that high-frequency expressions, including morphologically complex words and entire phrases, are retrieved from the lexicon as wholes even when they could also be generated by the regular rules of the grammar (e.g., Albright and Hayes 2003, Baayen et al. 2002, Barca et al. 2007, Butterworth 1983, Bybee 1985, 2000, Schreuder and Baayen 1995).

Linguistic theory draws a distinction between regular and irregular systems of generalizations. In rule-based theory, a regular system is one in which rules predicting different outputs do not compete with each other for application to any (class of) inputs. Thus, if rule A states that an input X corresponds to an output Y, there can be no rule B stating that X corresponds to a different output, Z (e.g., Plag 2003). In Optimality Theory (Prince and Smolensky 1993/2004), a regular system is one that can be described by a single strict constraint ranking, i.e., where the choice regarding which constraint to obey is deterministic, rather than stochastic (see also Coetzee and Pater in press).

Retrieval of complex expressions from the lexicon appears almost inevitable in a grammatical system that is not regular, i.e., where grammatical generalizations compete with each other for inputs and the competition is resolved stochastically (e.g., English past tense formation in Albright and Hayes 2003, velar palatalization in Russian in Kapatsinski 2010b, in press). In such a system, the grammar provides little evidence regarding the correct form of the output, which may lead the speaker to rely on lexical retrieval, since lexical retrieval is guaranteed to provide an accurate output, or at least an output that has been experienced often enough to be easily retrieved (cf. Zuraw 2000). The reliance on lexical retrieval is observed in that the same speaker can be very uncertain about how to inflect unfamiliar words and very certain about the inflections of words s/he knows. The grammar must account for a speaker’s behavior with novel lexical items, since these items by definition cannot be
retrieved from memory. The variable behavior can be explained by positing stochastic choice between competing generalizations (e.g., Albright and Hayes 2003, Boersma 1997, Coetzee and Pater in press, Kapatsinski 2010b). That is, rather than obeying the best-supported rule for each lexical item 100% of the time, which could result in deterministic behavior for each lexical item, the speaker obeys each competing rule in proportion to the amount of statistical support it has accumulated.

A stochastic grammar, which features rule competition, cannot account for lexically-specific deterministic behavior exhibited with familiar lexical items (e.g., Frisch et al. 2004: 220). If the speaker is uncertain about how to inflect unfamiliar words, his/her certainty about the inflection of a familiar word must come from retrieving the information about the familiar word from the lexicon. On the other hand, lexical retrieval appears unnecessary in a fully regular system, where grammatical generalizations do not compete with each other. In such a system, the grammar provides perfect information on the correct form to output, which allows for accurate production without lexical retrieval of derivable forms. Thus, if lexical retrieval were observed to be used to produce derivable forms in a fully regular system, this finding would provide strong evidence for a bias towards retrieval rather than computation in production. The problem is that, as we shall argue below, there are no fully regular systems. In an explicitly taught rule system, like the one examined in the present study, competition between the rules being taught is minimized, such that only one rule per lexical item is taught. Thus for the teacher the system of rules being taught appears to be completely regular. Nonetheless, for the learner trying to apply the system, competition during production still exists due to temporary uncertainties regarding which rules are applicable.

The hallmark of a retrieval process is the surface frequency effect. The higher the frequency of a word, the easier it is to retrieve from memory (e.g., Oldfield and Wingfield 1965). On the other hand, a grammatical rule, such as (1), applies equally well to all inputs that are eligible to undergo the change specified by the rule (i.e., all inputs that satisfy the rule’s structural description, cf. Albright and Hayes 2003, Bybee 2000, Clahsen 1999, Marcus et al. 1999, Pinker 1991), regardless of the frequency of the word in which the input is embedded. The frequency of the word containing the input structure is not part of the rule’s structural description, and in fact cannot be part of the rule’s structural description in a fully regular system because all instances of the input structure are equally eligible to undergo the rule (i.e., the rule is an operation over variables, as defined by Clahsen 1999, Marcus et al. 1999, and Pinker and Prince 1988).

(1) \([\alpha\text{voiced}] \rightarrow [\beta\text{voiced}]/\____[\beta\text{voiced}]\)
\([-\text{son}][-\text{son}]\)
Orthography is one part of the language system whose regularity varies widely across languages. Some languages, like Italian or Turkish, have a very regular orthographic system, where the spelling of a word is largely predictable from its phonology. Others, like English, have a very irregular phonology. Russian, the language examined in the present paper, falls somewhere in between (e.g., Glazkov 1999, Kuz’mina 1981, Sproat 2000). Nonetheless, even in languages where orthography as a whole is not regular it is possible to find parts of the grammar (henceforth, *systems*) that are regular and thus allow unambiguous derivation of a word’s spelling by retrieving individual morphemes and then applying the relevant morphological, phonological and/or orthographic rules. For instance, in English, the regular progressive suffix -ing has a regular spelling, thus in principle orthographic representations of whole words ending in -ing do not have to be stored to be spelled correctly. The spelling could be generated by retrieving and concatenating individual morphemes and then applying the relevant orthographic rules (as argued by Badecker et al. 1990, among others).

1.2 Previous work on whole-word frequency effects in production

In contrast to the voluminous literature on the retrieval of regularly inflected wordforms in perception, the corresponding literature in production is extremely sparse, and some researchers have argued that production might be less reliant on whole-form retrieval than perception (e.g., Baayen et al. 2002, Levelt et al. 1999). In regular systems, surface frequency effects in production have been reported for experienced writers only by Largy et al. (1996) and Sandra et al. (1999). For inexperienced writers, see Defior et al. (2008), Frisson and Sandra (2002), Largy et al. (2007) and Sandra and van Abbenyen (2009). Defior et al. (2008) and Largy et al. (1996, 2007) examine the spelling of silent inflections (the plural -s in Andalusian Spanish, a dialect where final /s/ is not pronounced, in the case of Defior et al., and plural markers in French nouns and verbs in the case of Largy et al.). Frisson and Sandra (2002) and Sandra et al. (1999) examine the spelling of homophonous first-person vs. third-person singular Dutch verbs ending in -d vs. -dt, both pronounced [t] due to final devoicing.

In spoken language production, Stemberger and MacWhinney (1986) and Bybee (2000) reported that English speakers omit the final /t/ or /d/ in past tense forms of high-frequency regular verbs more than in the past-tense forms of low-frequency regular verbs. Patterson and Connine (2001) reported English speakers exhibiting more flapping of intervocalic stops in high-frequency morphologically complex words than low-frequency ones. Ellis and Schmidt (1997) found word frequency effects on errors and latencies of plural production in the early stages of learning a miniature artificial language even for the
dominant prefix. However, the results are potentially explained by stem frequency rather than whole-word frequency, which would be consistent with retrieving individual morphemes and concatenating them by rule. Stem frequency and surface frequency are either not compared or, in the case of Ellis and Schmidt (1997), are perfectly confounded during training. So far, the only study that has reported a clear effect of whole-word frequency on regular morphologically complex words in speech production while controlling for base frequency is Tabak et al. (in press), who examined picture naming latencies for Dutch inflected past tense verb forms. Although all effects reported for speech involve irregular morphological systems, whole-word frequency effects appear to be observed in Tabak et al.’s study even for words bearing the most common competitor affix, which would be attached by a default ‘elsewhere’ rule according to Pinker (1991) and Pinker and Prince (1988).

The present study extends the previous work in several ways. First, this is the first study to examine words bearing purely contextual inflections (agreement affixes on adjectives, cf. Booij 1993 for the contextual/inherent distinction). Baayen et al. (1997) and Bertram et al. (2000) do not find whole-word frequency effects for verbal inflections representing person agreement in visual word recognition and hypothesize that, unlike words bearing inherent inflections, words bearing contextual inflections may be assembled online. This hypothesis remains consistent with data gathered so far. Frisson and Sandra (2002) and Sandra et al. (1999) find whole-word frequency effects for Dutch verbs bearing suffixes realizing person agreement. However, the relevant suffixes also realize tense, which is an inherent inflectional category for verbs. Largy et al. (1996) investigate the spelling of French verbs that are homophonous with nouns in a dictation task, finding that the error rate on the verb is highest when the verb is less frequent than its noun counterpart. The effect of noun (i.e., intruding form) frequency remains when the frequency of the verb (i.e., the target form) is controlled. However, the authors do not test whether the effect of verb frequency remains when noun frequency is controlled, thus the frequency of the noun homophone could be responsible for the observed “relative” frequency effect. As Largy et al. (1996: 247) point out, the nouns they investigate bear (silent) inherent inflection markers (plural suffixes). Thus, the results are potentially consistent with Baayen et al.’s (1997) and Bertram et al.’s (2000) hypothesis that words bearing contextual inflections are not stored. The present study tests this hypothesis by comparing whole-word frequency effects in the spelling of adjectives bearing purely contextual agreement suffixes vs. verbs bearing partially-inherent person/tense suffixes.

Second, the affixes examined in the present study exist in spoken language as well as written language and the orthographic rules we examine are phone(me)-grapheme correspondences rather than mappings between semantic and orthographic representations. It is widely agreed that there is an
orthographic output lexicon that is separate from the phonological lexicon and directly connected to semantics since the two lexicons can be damaged independently (Hillis 2001) and orthographic representations do not appear to be automatically activated during speech production, unlike during speech perception (Alario et al. 2007, Bi et al. 2009, Damian and Bowers 2009, Roelofs 2006). The orthographic output lexicon appears to contain (possibly among other things) representations for single morphemes (Allen and Badecker 2001, Badecker et al. 1990). Thus semantic-orthographic mappings examined in previous studies can be argued to be part of the lexicon. Thus, if the lexicon is assumed to be distinct from the grammar (Pinker 1991, Pinker and Prince 1988, Ullman 2005), semantic-orthographic mappings could be argued to be more likely to enter into competition against whole-word representations than clearly grammatical rules. Finally, the words examined in the present study do not have homonyms (although morphemes within the words do). Thus the errors examined in the present study cannot be described as homophone intrusion errors, since the erroneous output does not match any word or phrase target in Russian, and thus they may be argued to be less likely to result from lexical competition and more likely to reflect misapplication of a grammatical rule.

The last two characteristics of the present study are shared with Sandra and Van Abbenyen (2009), who examine the spelling of Dutch past tense forms by 12-year-old children in a fill-in-the-verb task. To be spelled correctly, the past tense form of a Dutch verb sometimes requires the writer to pay attention to its morphological structure, e.g., *braad+d* would be spelled *braadde* while a monomorphemic word with the same pronunciation would be spelled *brade*. Sandra and Van Abbenyen (2009) find that the Dutch children make most errors for verbs ending in *VC+d* (e.g., *bruldde*, corresponding to the infinitive *meld+en*) rather than *V+d* hypothetically because the orthographic sequence *VC+d* is quite common (e.g., occurring in *belde*, infinitive *bell+en*) whereas *V+d* sequences do not occur in the spelling. Sandra and Van Abbenyen (2009) argue that Dutch spelling involves whole-word representations and either competing sublexical orthographic patterns or lexical analogy.

Finally, on the methodological side, this is the first study of spelling to treat word frequency as a continuous variable, rather than looking at differences between “low-frequency” and “high-frequency” words (see Cohen 1983 for statistical arguments against dichotomization and Tabak et al., in press, for word frequency as a continuous variable in speech production). Finally, this is also the first study to use corpus data in addition to experimental data. Corpus data provide evidence regarding the existence and prevalence of lexical frequency effects in natural writing outside of the context of a dictation task and, as we shall see, also shed light on why lack of rule knowledge may go undetected in schooling.
1.3 The relation of writing to speech

To what extent can we expect results from written language production to generalize to spoken language production? One reason to be optimistic is that writing and speech clearly share many processing stages, which in most languages include morphological derivation. Furthermore, morphological rules that can be argued to involve only orthographic representations (as in French silent inflection, cf. Largy et al. 1996) do not appear to differ in kind from morphological rules that operate on phonological representations, which is not surprising given that the writing-specific morphological rules are rules that used to (also) apply in speech. Thus, the kinds of morphological rules found in writing are essentially the same as the kinds of morphological rules found in speech.

Morphological derivation may compete with lexical retrieval in both spoken and written language production. In perception, there is extensive evidence that, for literate speakers, orthographic representations are automatically activated when a word is heard, e.g., words prime homophones that are not homographs less than they prime homophones that are homographs (Pattamadilok et al. 2009, Perre et al. 2009, Taft et al. 2008, Ziegler et al. 2004, Ziegler and Ferrand 1998) suggesting that orthographic and phonological lexical representations form a single integrated lexicon. Furthermore, orthographic and phonological representations both appear to be directly connected to semantic representations, indicating that orthographic representations are not simply parasitic on phonological ones (e.g., Coltheart et al. 2001, Plaut et al. 1996). In production, parallel experiments testing for an influence of orthographic representations on phonological ones have tended to yield null results (cf. Damian and Bowers 2003 vs. Alario et al. 2007, Bi et al. 2009, Damian and Bowers 2009, Roelofs 2006). Thus orthographic and phonological output representations may not be part of the same lexicon, although conclusions based on null results are necessarily preliminary in nature. Nonetheless, we can expect the orthographic lexicon to compete with the grammar in similar ways to the phonological lexicon because lexical access in writing is subject to the same kinds of influences as lexical access in speech, including, crucially for our purposes, the effect of word frequency (Jescheniak and Levelt 1994, Kapatsinski 2010a, Oldfield and Wingfield 1965).

Morphological rules feed into phonological rules in speech production. Morphological and/or phonological rules feed into sound-to-spelling correspondence rules and orthotactic constraints in writing (e.g., Neijt 2002, Sproat 2002). Sound-to-spelling correspondence rules and orthotactic constraints are not shared with spoken language. However, they are formally parallel to the rules and constraints of phonology. If they were not, linguistic theory could not be used to successfully describe orthography, while in fact orthographies are
describable in the same kinds of formal notation as other areas of grammar (see, e.g., Bierwisch 1975, Nunberg 1990, Nunn 1998, Venezky 1970 for rule-based accounts, Geilfuß-Wolfgang 2002, Neef 2003, Primus 2004, Smith 2004, Sternefeld 2000, and Wiese 2004 for Optimality-Theoretic approaches, and Sproat 2000 for an approach based on Head-driven Phrase-Structure Grammar). In fact, some have argued that formal phonology is a better description of writing than of spoken language (e.g., Aronoff 1992, Faber 1992, Port 2007, Port and Leary 2005). In particular, the representations to which the rules or constraints of phonological theory refer consist of a sequence of discrete static targets (what Hayes 2009: 301 calls the “static feature hypothesis”), which is a much less controversial description of print than it is of speech (Port 2007, Port and Leary 2005). While modern phonological theory assumes multiple parallel sequences of static targets, i.e., tiers, the same arguments that have motivated the positing of tiers in phonology also apply to written language. For instance, Badecker (1996) and Caramazza and Miceli (1990) show compensatory lengthening phenomena in orthographic errors parallel to those described by Stemberger (1984) for speech errors and argue for a timing tier in orthography (see also Neijt 2002, Nunn 1998), thus the parallel between phonology and orthography still holds.

Finally, the existence of spelling pronunciations (e.g., Comrie et al. 1978/2003 for Russian, Giegerich 1992 for English, Warner et al. 2004 for Dutch) suggests that orthographic representations of words can influence phonology. Hall (2004) shows that orthographic constraints implementing phoneme-grapheme correspondences and phonological markedness constraints can interact in the derivation of a pronunciation for a novel loanword in German, which he models by placing orthographic and phonological constraints into a single constraint ranking with both the underlying phonological form and the orthographic form serving as inputs (see also Neijt 2002 for a similar proposal). Thus, for literate speakers, phonology and spelling rules form an integrated system such that pronunciation may be repaired if it disobeys dominant grapheme-phoneme correspondence relations (Hall 2004, Sproat 2002).

Despite all the similarities, there are also important differences between phonological and orthographic rules that may limit the generalizability of results obtained with writing to speech. First, orthographic rules (including the ones in the present study, according to Grigorenko 2003) are often learned through explicit instruction (though not all orthographic rules possessed by a writer are explicitly taught, e.g., Bryant et al. 2000, Geilfuß-Wolfgang 2002, Nunberg 1990, Pacton et al. 2001, 2005). Robinson (1997) and Williams (1999) tested the extent to which learners of a miniature second language relied on lexical retrieval following explicit and implicit training. They observed greater reliance on lexical retrieval when learning was implicit than when rules were explicitly taught. In addition, whereas humans learning the rules of a
language implicitly do not show a tendency to induce the system of rules that minimizes rule competition (Kapatsinski 2010b, in press), designers of orthographic rules certainly do attempt to create a regular grammar and teachers of orthography attempt not to teach multiple conflicting rules for a given word. Thus we may observe less reliance on lexical retrieval with explicitly taught orthographic rules than with implicitly acquired phonological rules.

On the other hand, orthographic rules are acquired relatively late compared to phonological rules. Ullman (2005) suggests that procedural memory declines throughout life while declarative memory improves. He further suggests that words are learned using declarative memory while rules are learned using procedural memory, at least if they are learned early under implicit conditions. Thus the age-related decline of procedural memory leads to sensitive-period effects on grammar acquisition. If rule learning is more difficult in late childhood than in early childhood, while word learning is equally easy, we may expect late-acquired rules to be more likely to be outcompeted by lexical retrieval than early-acquired rules. Thus, we may observe more reliance on lexical retrieval with orthographic rules, which are learned late, than with phonological rules, which are learned early. Some evidence for this hypothesis is provided by Bowden et al. (2010), who find stronger whole-word frequency effects on inflected Spanish verbs for non-native speakers of Spanish than for native ones.

It can certainly be argued that late-learned explicitly-taught rules in and of themselves (aside from implications for more ‘natural’ rules) are of no importance to linguistic theory because linguistic theory is concerned only with rules that are acquired using the innate domain-specific language acquisition device (LAD), which either is turned off at this point in development or specializes in implicit learning from natural linguistic input with the dynamic characteristics of speech or sign language. Only rules acquired using LAD are assumed to be part of the language faculty, a neural module used to process rules that are consistent with Universal Grammar (UG) and linguistic theory is assumed to be responsible only for representations that are part of the language faculty. The only candidate for a brain area hosting the part of the language faculty specialized in learning and processing natural grammatical rules that has been proposed so far is Broca’s area. In an intriguing study, Musso et al. (2003) found that activation in Broca’s area is observed when subjects are presented with sentences that violate previously learned rules that are consistent with UG but Broca’s area is not activated when a sentence violates previously learned rules that violate UG (e.g., “move the third word in the sentence to the front to form a question”, see also Smith et al. 1993). Musso et al. argue that Broca’s area is the seat of the mechanism for learning and processing syntax. Importantly, the subjects in the study were adults who were taught the (natural or unnatural) rules explicitly, and the stimuli exemplifying the rules were always
presented in print. One could also argue that the rules were orthotactic rather than syntactic in nature, since no meanings were presented. Thus, either explicit, late-learned rules can be processed using the language faculty or we must still await any evidence for a language faculty that only processes ‘natural’ linguistic rules. I suggest that late-acquired rules are of interest to linguistic theory if linguistic theory is to be truly ‘biolinguistic’ and describe the properties of grammatical computation carried out by the “language areas” of the brain (see also Friederici et al. 2002 and Newman-Norlund et al. 2006).

1.4 Aspects of Russian grammar

Russian orthography is morphophonological (Glazkov 1999, Kuz’mina 1981, Sproat 2000): there are default phoneme-grapheme correspondences, which specify how a phoneme is realized orthographically when in a strong position, e.g., /z/ realized as [z] is spelled ‘з’ while /s/ realized as [s] is spelled ‘с’. I will use ‘з’ for the Cyrillic letter ‘з’ and ‘с’ for the Cyrillic letter ‘с’ to make the default phoneme-grapheme correspondences obvious for readers not familiar with Cyrillic. Russian largely obeys the morphological principle (Avanesov 1964, Sandra and Fayol 2003), thus a morpheme usually has a constant spelling, which means that context-sensitive phonological processes, such as unstressed vowel reduction and word-final devoicing, are not represented in writing. This is also generally true for voicing assimilation, which is described by the rule in (1). However, for prefixes ending in /z/, the output of the assimilation rule is prescribed to be reflected in the orthography. Thus, one can think of Russian orthography as featuring the rule in (2) or, alternatively, the surface-phonology-to-orthography mapping in (3). In either case, the mapping is completely regular and is not lexically-specific.6

(2) ‘з’ → ‘с’/_ Prefix ‘-voiced’7
(3) [s] → ‘с’/_ Prefix ‘-voiced’/

There are a number of prefixes ending in /z/, the most common and productive of which is the adjectival caritive prefix bez- ‘-less’, shown in (4). Other /z/-final prefixes are verbal prefixes that indicate perfective aspect as well as add various partially unpredictable nuances of meaning, including iz- and roz-. There are also a number of verbal prefixes that do not undergo orthographic voicing assimilation, including s-, shown in (5), nad-, pod- and or-, although they do undergo phonetic voicing assimilation. In addition to voicing assimilation, all prefixes ending in /z/ or /s/ undergo additional place assimilation before /ʒ/ and /ʃ/, which is not reflected in the spelling.
(4) The adjectival caritive prefix bez-

a. Before sonorants (underlying form)
   lik ‘face’ beazel-lik-ij ‘faceless’ [z] ‘z’
   ux-o ‘ear’ beazel-ux-ij ‘earless’ [z] ‘z’
   mozg ‘brain’ beazel-mozg-lyj ‘brainless’ [z] ‘z’

b. Before voiceless obstruents (assimilated form)
   strax ‘fear’ beazel-stras-nyj ‘fearless’ [s] ‘s’
   prefix ‘prefix’ beazel-prefix-nyj ‘prefixless’ [s] ‘s’
   tolk ‘purpose’ beazel-tolk-ovyj ‘purposeless’ [s] ‘s’

c. Before voiced obstruents (underlying form)
   dom ‘home’ beazel-dom-nyj ‘homeless’ [z] ‘z’
   golos ‘voice’ beazel-golos-ylj ‘voiceless’ [z] ‘z’
   zub ‘tooth’ beazel-zub-yj ‘toothless’ [z] ‘z’

d. Before /ʒ/ and /ʃ/ (place assimilation not reflected orthographically)
   žiznj ‘life’ beazel-žizn-ennyj ‘lifeless’ [ʒ] ‘ʒ’
   štan-y ‘pants’ beazel-štan-nyj ‘pantless’ [ʃ] ‘ʃ’

(5) The verbal perfective s-

a. Before sonorants (underlying form)
   igratj ‘to play IMPFV’ s-ygratj ‘to play PFV’ [s] ‘s’
   lovij ‘to catch IMPFV’ s-lovij ‘to catch PFV’ [s] ‘s’
   molchatj ‘to keep quiet IMPV’ s-molchatj ‘to keep quiet PFV’

b. Before voiceless obstruents (underlying form)
   selitj ‘to house IMPFV’ s-selitj ‘to rehouse together PFV’ [s] ‘s’
   pitj ‘to drink IMPFV’ s-pitj ‘to drink off PFV’ [s] ‘s’
   taščitj ‘to drag IMPFV’ s-taščitj ‘to drag off PFV’ [s] ‘s’

c. Before voiced obstruents (assimilated form)
   datj ‘to give PFV’ s-datj ‘to give away PFV’ [z] ‘z’
   bitj ‘to hit IMPFV’ s-bitj ‘to hit down/together PFV’ [z] ‘z’

d. Before /ʒ/ and /ʃ/ (place assimilation not reflected orthographically)
   žečj ‘to burn IMPFV’ s-žečj ‘to burn PFV’ [ʒ] ‘ʒ’
   šitj ‘to knit IMPFV’ s-šitj ‘to knit PFV’ [ʃ] ‘ʃ’

All Russian adjectives (including those beginning with bez-) agree with the noun they modify in number and case. If the noun is singular, then the adjective also agrees with it in gender. Case, number and gender are reflected in the inflectional suffix. Adjectival inflection is regular in Russian and is illustrated in (6). Any form of an adjective is regularly derivable from any other form. (The choice of -oj vs. -yj in the masculine singular nominative and the identical masculine singular accusative inanimate is predictable given other forms if stress on the other forms is taken into account: final stress → -oj; stem
stress → -yj). It is therefore interesting to test whether Russian writers rely on their memory of particular inflectional forms in spelling the prefix. One reason to doubt this possibility is that the inflectional suffix follows the to-be-spelled prefix and can be separated from it by as many as a dozen letters, e.g., bes-
podborodoč-n-yj ‘chinless’. Thus, if error rate is influenced by the frequency of a particular wordform, this would provide strong evidence for the involvement of representations of individual wordforms in spelling.

Verbal inflection, unlike adjectival inflection, is irregular: given any particular form of a verb, it is not possible to predict other forms of the verb with complete certainty (see Shvedova et al. 1980: 646). The most predictive base forms are the non-past 3rd person plural and the past tense feminine singular but a large proportion of verbs require listing of multiple wordforms per verb (see Shvedova et al. 1980: 646–664 for a full description). Thus, if regularity of the relevant grammatical rules is the main factor influencing the extent to which the writer relies on lexical retrieval vs. grammatical computation (Pinker 1991, Pinker and Prince 1988, Ullman 2005), we may expect stronger frequency effects for the verbal prefixes than for bez-, at least within lexemes.9

The citation form for a Russian adjective is the masculine singular nominative, and the citation form for a verb is the infinitive. These are the forms that always appear in dictionaries and grammars, and the forms that dominate textbooks. Albright (2008) proposes that the learner of a language decides to memorize forms s/he encounters only if they cannot be derived from forms s/he already knows. Citation forms are then especially likely to be memorized because they, being encountered more often, have more opportunities to be encountered early during the learning process when other forms of the word have not been encountered yet and thus cannot be used to derive the citation forms. A less interesting factor that makes the same prediction (stronger frequency effects for citation forms) is that the proportion of hits that come from dictionaries and grammars, and are therefore error-free, may be higher for a rare

<table>
<thead>
<tr>
<th></th>
<th>Nominative</th>
<th>Genitive</th>
<th>Dative</th>
<th>Accusative</th>
<th>Instrumental</th>
<th>Prepositional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Masculine</td>
<td>-yjloj</td>
<td>-ogo</td>
<td>-om</td>
<td>-ogo</td>
<td>-yjloj</td>
<td>-ym</td>
</tr>
<tr>
<td>Neuter</td>
<td>-oje</td>
<td></td>
<td></td>
<td></td>
<td>-oje</td>
<td>-ym</td>
</tr>
<tr>
<td>Feminine</td>
<td>-aja</td>
<td>-oj</td>
<td>-uju</td>
<td>-oj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plural</td>
<td>-ije</td>
<td>-yx</td>
<td>-ym</td>
<td>-yx</td>
<td>-ije</td>
<td>-ymi</td>
</tr>
</tbody>
</table>
word than for a frequent word. Alternatively, word frequency may not be as important a predictor of accuracy for the masculine singular as for other forms if orthography training received by Russian speakers results in development of orthographic rules triggered by the presence of the masculine singular nominative suffixes -yj or -oj, since these suffixes usually occur in the forms Russian spellers are trained on. Thus nominative singular feminine and nominative plural forms were included in the present study for adjectives, and the third person singular non-past forms were included for verbs. These particular forms were chosen because they appeared to be the most common in pilot searching, allowing for the largest sample size and most reliable word frequency estimates.

While the regularity of the relevant phonology, orthography, and morphology makes storage and retrieval of words bearing the prefix bez- strictly unnecessary, there is also one fact about Russian morphology that may increase its likelihood (cf. Bertram et al. 2000) and cause the writer to pay attention to the inflectional suffix of the adjective beginning with bez-. This fact is that the prefix bez- ‘-less’ is similar in both form and meaning to the preposition bez ‘without’, the prefix being grammaticalized from the preposition. An adjective can in fact be productively formed from a prepositional phrase headed by bez, as long as the prepositional phrase has no adjective intervening between the preposition and the noun, as shown in (7).

(7) bez (svojevo) doma ‘without (one’s own) home’  bez-dom-n-yj
    ‘homeless’
bez (kakix-libo) problem ‘without (any) problems’  bez-problem-n-ij
    ‘problemless’

I would argue that the structure of an adjective like bez-dom-n-yj is more often [[bez[dom]]ppnyj]Adj than [bez[[dom][nyj],Adj]]Adj because there are many cases where the bez-less form of the adjective does not exist while the bez-initial prepositional phrase does exist and no cases where the adjective exists but the prepositional phrase does not. Some examples are shown in (8) where the corresponding bez-less adjectives are (from top to bottom) krylatyj, pečalnyj, preryvistyj, pušistyj, šerstistyj or volosatyj, and češujčatyj.

(8) bes-krylyj ‘wingless’                  bez kryljey ‘without wings’    *krylyj
  bes-kručinnyj ‘careless’                bez kručiny ‘without care’    *kručinnyj
bes-pereboj-nyj ‘free of interruption’  bez perebov ‘without interruptions’
bes-šerstnyj ‘hairless’ [of an animal]   bez šersti ‘without hair’    *šerstnyj
bes-češujnyj ‘scaleless [of fish]’      bez češui ‘without scales’    *češujnyj
bes-kljuvyj ‘beakless’                   bez kljuva ‘without a beak’    *kljuvyj
bes-polos-nyj ‘stripeless’               bez polos ‘without stripes’    *polosyj
Furthermore, it is always straightforward to predict the meaning of the derived adjective from the meaning of the prepositional phrase but the bez-initial adjective and the bez-less adjective are not so straightforwardly related. For instance, in (9a) the bez-less adjective is not a direct antonym of the bez-initial adjective, since the two occur with non-overlapping sets of nouns. Rather, the antonym is krovavyyj ‘bloody’, which has the same root but a different suffix. For instance, a revolt can be bez-krovnyj or krovavyyj and but not krovnyj. In contrast, the bez-less adjective occurs in krovnyj vrag ‘blood enemy’ and krovnje den’gi ‘hard-earned money’, but the two nouns cannot occur with beskrovnnyj. In (9b), a ‘foamless’ substance or process does not produce foam, while a foamy substance has foam, e.g., in my corpus, the pouring of beer (razliv piva) can be bespennyj but not pennyj whereas the opposite is true of beer (pivo) itself. In (9c), the difference in meanings is even more extreme with no obvious synchronic connection.

(9)  a. bes-krov-nyj ‘bloodless’ bes krovi krovnyj ‘hard-earned, blood’
    b. bes-pen-nyj ‘foamless’ bez peny pennyj ‘foamy’
    c. bes-cel'nyj ‘aimless’ bez celi ‘without aim’ cel'nyj ‘whole’

Like the prefix-final /z/, the preposition-final /z/ is devoiced by the following voiceless obstruent but, unlike for the prefix, the devoicing is not reflected in the spelling for the preposition, which is the case for all prepositions (and all separate words) in Russian. If bez-initial adjectives are (at least sometimes) derived from prepositional phrases that already have bez in language production, then, at some stage during language production, bez-the-preposition and bez-the-prefix are the same unit. This may make application of the spelling rules for the morpheme especially difficult by introducing uncertainty regarding whether the rules for prepositions or the rules for prefixes are applicable.10

More controversially, uncertainty will also arise when the writer does not know which /bez/ s/he is writing in advance in a dictation task, where on some occasions it could be resolved only by the inflectional suffix at the end of the bez-initial word or phrase.

Importantly, the difference in behavior between prefixes and prepositions is not grounds for considering Russian spelling irregular in the classical linguistic sense, since the entire system can be described using a single strict constraint ranking, illustrated in (10). The UNIFORMEXPOSITION constraint militates against having multiple spellings for the same morpheme (see Kenstowicz 1996 for the same family of constraints in phonology; this is also known as the morphological principle in orthography, e.g., Avanesov 1964, Glazkov 1999, Kuz’mina 1981, Sandra and Fayol 2003). This constraint must be ranked above Ident-[‘] (‘spell what you hear’) to ensure that assimilation is not reflected in the
What is it I am writing?

spelling of morphemes that do not end in /z/, as shown in (10a). The fact that prefixes ending in /z/ realized as [s] are spelled with ‘s’ rather than ‘z’ requires a high-ranked *‘z’: without it, ‘spell what you hear’ would lose to uniform exponence for prefixes ending in /z/ just like for all other prefixes (compare 10a vs. 10b–c). Finally, a high-ranked UNIFORMEXPOENCE constraint that applies only to stems is needed to ensure that all prepositions, even those ending in /z/, have a constant spelling, as shown in (10d).

(10)

<table>
<thead>
<tr>
<th></th>
<th>UNIFORMEXPOENCE_STEM‘</th>
<th>‘z’</th>
<th>UNIFORMEXPOENCE‘*</th>
<th>Ident-[‘]</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>/ot-d/ [odd]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>‘odd’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>‘otd’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b.</td>
<td>/bez-p/ [besp]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>‘bezp’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>‘besp’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c.</td>
<td>/bez-d/ [bezd]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>‘bezd’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d.</td>
<td>/bez p/ [besp]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>‘bez p’</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The most confusable prefix after bez- is iz-, which, in its semantically transparent uses (as in iz-ojti ‘come out’), is similar in both meaning and form to the preposition iz ‘from/out of’, although the former attaches to a verb while the latter attaches to a noun phrase. Other Russian prefixes are not as similar to a differently-spelled separate word. The prefix roz-, ‘apart’ in its semantically transparent uses, is similar in form to the noun raz ‘time’ as in ‘one time’, ‘two times’, etc.: the two sound the same when unstressed; the prefix is always unstressed in verbs while the noun may receive phrasal or contrastive stress. However, the two morphemes are not similar in meaning.

For prefixes ending in /z/, the presence of a similar word may increase the likelihood of an error, since the error results in the production of an existing
morpheme. This is not the case for the prefixes whose voicing assimilation is not reflected in the spelling. With such prefixes, producing an error could not result from the prefix being replaced by the orthographic form of another morpheme, since the prefix and the homophonous morpheme have the same spelling. The perfective prefix s- shares the phonetics and the orthography with the preposition s but is not at all semantically similar to s, which means ‘with’. The prefixes ot-, pod-, and nad- are like iz- except, just like for the corresponding prepositions ‘away from’, ‘under’ and ‘over’, their voicing assimilation is not reflected in the spelling, thus the similar words favor the same spelling as the prefix and are not increasing uncertainty about which rules to apply. The differences between the prefixes are summarized in (11).  

<table>
<thead>
<tr>
<th>Final C is underlyingly</th>
<th>Orthography mapped onto surface form or underlying form?</th>
<th>The homophonous word</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spelled the same</td>
<td>Has a similar meaning</td>
</tr>
<tr>
<td>bez-</td>
<td>Voiced</td>
<td>Surface</td>
</tr>
<tr>
<td>iz-</td>
<td>Voiced</td>
<td>Surface</td>
</tr>
<tr>
<td>roz-</td>
<td>Voiced</td>
<td>Surface</td>
</tr>
<tr>
<td>s-</td>
<td>Voiceless</td>
<td>Underlying</td>
</tr>
<tr>
<td>ot-</td>
<td>Voiceless</td>
<td>Underlying</td>
</tr>
<tr>
<td>pod-</td>
<td>Voiced</td>
<td>Underlying</td>
</tr>
</tbody>
</table>

Bertram et al. (2000), suggest, based on differences in sensitivity to full-form lexical frequency among Finnish affixes, that full-form lexical storage may be triggered by the existence of affixes that are homophonous with the affix borne by a word (see also Hotopf 1980, 1983, Morton 1980, Largy et al. 1996 for data showing that forms that are homophonous but not homographic are particularly susceptible to spelling errors due to their orthographic forms being confused for each other). All of the Russian prefixes we examine in the present study are homophonous with separate words in at least some prosodic positions, and thus words bearing them may be expected to be stored. However, the prefixes vary widely in how semantically similar they are to the homophonous words and how likely they are to be followed by the same stem as the homophonous word. Thus, by comparing the prefixes’ sensitivities to word frequency, we can evaluate how semantic and distributional similarity among homophonous morphemes may influence the balance between storage and computation. I shall argue that it is not homonymy itself that is important but the attendant uncertainty regarding which rules are applicable (see also Albright 2009). The importance of degree of uncertainty regarding rule applicability forces a reconsideration of the notion “regular system” as there is always
some uncertainty regarding which rules are applicable during the course of natural language processing.

2. The corpus study

2.1 Methods

2.1.1 Measuring frequency and error rate.
Since existing corpora of Russian do not contain appreciable numbers of spelling errors, the present sample of errors is based solely on web data, obtained by searching for all relevant words found in the 240,000-word Elsevier’s Russian-English Dictionary (Macura 1999) using the Google search engine. Convergent estimates of word frequencies were obtained from Google (http://google.com) and the Russian National Corpus (Institute for Russian Language 2005, available at http://corpora.yandex.ru/search-main.html). The great advantage of Google frequency estimates is the sheer size of the data sample. While many words in the dictionary could not be found using either Google or the Russian National Corpus, only 45% of the words found on Google could be found in the Russian National Corpus. The validity of web-based frequency counts is supported by the results of Keller and Lapata (2003), who observe that plausibility judgments for bigrams that are found only on the Web (and not in the British National Corpus) are reliably predicted by Google frequencies, indicating that Google counts are capturing psychologically relevant variation on the low end of the phrase frequency continuum that the corpus counts are not. Furthermore, even for bigrams found both in the corpus and on Google, correlations with plausibility judgments are higher for web-based frequency counts than for corpus-based ones in Keller and Lapata’s (2003) data (see also Kapatsinski and Radicke 2009).

Nonetheless, the use of a secondary source of word frequencies beyond Google is called for. First, the observed error rates, which can only be obtained from the web are not independent of word frequency on the web by definition (error rate = number of errors divided by word frequency on the web). Second, there are certain limitations of frequency estimates based on data from the web derived via a commercial search engine, which call for a secondary source of word frequencies to correlate with observed error rates. First, web data do not provide a representative sample of topics and communicative contexts encountered by a speaker throughout life, and are disproportionately biased towards advertisements. For instance, the second most frequent word bearing the prefix bez- ‘less’ in web data is the adjective besprovodnoj ‘wireless’, which is hardly used in spoken Russian. Second, the number of hits returned by a search engine is influenced by whether the search term has been recently used: the result
of a search is cached and reused in future searches until the cache is cleared (Eu 2008). Third, search engines, such as Google, do not count all hits found on webpages but rather extrapolate from a (non-representative) sample of the first X pages retrieved (Eu 2008).

The estimate of the number of errors is also affected by the issues above, although perhaps to a lesser extent since the number of errors per word is usually relatively low. In addition, it is also affected by the following issues:

1) Slavic languages other than Russian have different spelling conventions. In particular, voicing assimilation for iz- and roz- is not reflected in the spelling in Bulgarian. Bulgarian shares many verb forms with Russian. Fortunately, the prefix bez- is free from this problem since (non-predicative) adjective forms differ between Russian and Bulgarian (cf. http://bg.wiktionary.org/). For the verbal prefixes, I used the infinitive and the third person singular indicative non-past forms, which are not shared with Bulgarian. Nonetheless, there remains the possibility of interference from other languages’ orthographic norms with the writers’ Russian as well as transfer of Russian morphology into non-Russian texts. To reduce the latter possibility, all words where the first page of results contained an entry that was not written in Russian were excluded from the sample.

2) Voicing assimilation is consistently represented in writing in Russian only since Pokrovskij and Bonch-Bruevich (1918). The spelling was first standardized by Grot in 1874–1886 (Grot 1886) based on informal observation of how words were already being spelled (Grot 1873: 12). According to Grot (1886: 46–48) all /z/-final prefixes were to be spelled with ‘z’ before ‘s’, and with ‘s’ before other voiceless obstruents, except for bez-, which was to be always spelled with ‘z’. This means that some ‘errors’ found by Google may be correct spellings from pre-1918 documents. It is impractical to check every retrieved webpage, given that the number of retrieved webpages containing potentially erroneous spellings for the most frequent words is in the hundreds of thousands. Furthermore, checking retrieved webpages only for rare words would artificially decrease the error rate for such words relative to high-frequency words. Thus it was decided to exclude all words for which the first page of results contains at least two hits coming from pre-spelling-reform documents. See also the general discussion for post-hoc analyses designed to estimate the possible contribution of pre-revolutionary spellings to the results.

3) There is an active community of Russian Internet users who playfully swap default graphemes for phonemes that neutralize to a single phone in certain environments (падонкоффский сленг, ‘bottom-dwellers’ slang’, cf. uaff.org for some prime examples from the originator of the trend and others). Thus, since /s/ and /z/ neutralize before obstruents and word-
finally, the prefix *s-* is spelled *z-* by such writers, *ot-* is spelled *od-*., and *pod-* is spelled *pot-* . Fortunately, such intentional misspellings do not appear to affect the prefixes whose spelling reflects voicing assimilation (*bez-* , *roz-* , and *iz-* ).

2.1.2 Statistical issues in correlating frequency and error rate.

Error rate is defined as the number of errors divided by the total number of hits, whether correctly spelled or erroneous. Thus, for the word *bezprovodnoj*, the error rate is the frequency of ‘besprovodnoj’ divided by the sum of the frequencies of ‘bezprovodnoj’ and ‘besprovodnoj’. The number of errors is, of course, an integer, which introduces a possibility for a spurious positive correlation between word frequency and error rate if words with no errors are included in the analysis. Suppose the error rate is constant across word frequency, for instance, 1%. In this case, words with a frequency below 1/errorRate (in our case 100) would be expected to have no errors. Thus, the error rate for low-frequency words is likely to be underestimated relative to high-frequency words.

It became apparent from early on in the project that the relationship between error rate and word frequency is best fit by a power function, i.e., that the logarithm of the error rate varies linearly with the logarithm of word frequency. That is, since $\log(x) = -\log(1/x)$, the number of errors stays relatively constant while the number of correct spellings varies. Small changes in the number of errors produce large differences in log-transformed error rate when the number of errors is small. This led me to exclude all words that have few errors. The limit on the number of errors was set somewhat arbitrarily at 10 hits.

Given the lower limit on the number of errors that the word must have to be included in the study, i.e., 10, the minimum error rate for the present study is 10/word frequency on Google. This increases the possibility of a spurious negative correlation. This possibility was controlled statistically via two methods. The first, rather crude method is to subdivide the data, with a separate correlation analysis run on each subset. A word has to have a frequency of at least X and an error rate of at least 10/X to be included in the subset. For instance, if an analysis includes words that have a frequency of at least 100, then we must exclude all words with an error rate of .1 or lower because a word with a frequency of 100 and an error rate of .1 would have 10 errors and thus would be excluded from the study. The subsets were arbitrarily defined to be separated from each other by a multiple of 10, starting with 100/0.1 until all data are included in some subset analysis. For *bez-* , this procedure resulted in the following subsets: 100/0.1, 1000/0.01, and 10000/0.001. The disadvantages of this method are that 1) some data are not used in any analysis, 2) some data are used in multiple analyses, requiring correction for multiple comparisons, 3) the subset boundaries are defined arbitrarily, and 4) there is no way to estimate...
how much variance in error rate is accounted for by word frequency over and above the artifacts of the sampling procedure.

In order to overcome these limitations, I replicated the sampling procedure in a Monte Carlo simulation, using the same procedure used to obtain the sample from the actual population to draw similar samples from a null population, in which the correlation between word frequency and error rate is zero whereas the marginal distributions of word frequency and error rate are the same as in the actual population. To do this, we define log(error rate) for each word sampled from the null population to be the mean log(error rate) in the actual sample plus a random deviation based on a normal distribution with the same variance as the noise observed around the correlation line between log(word frequency) and log(error rate) in the actual sample. Any word whose error rate is less than ten over its word frequency is excluded from each null sample. The correlation between log(word frequency) and log(error rate) is calculated for each resulting null sample, producing a distribution of correlation coefficients for the null hypothesis. If the correlation coefficient in our actual sample falls outside the distribution of correlation coefficients derived from the null population, where the true correlation is zero, then we can say that the probability $p$ of obtaining our results by chance given our sampling procedure is less than $1/N$, where $N$ is the number of samples we drew from the null population. See Appendix 1 for a link to the R script.

2.2 Results

2.2.1 Adjectives with bez-: Between-lexeme analysis.

Figure 1 shows the across-lexeme correlation between word frequency and error rate for the prefix bez-. For each wordform, there is a very strong correlation between word frequency and error rate, such that the more frequent the word, the lower the error rate. Since $\log(x) = -\log(1/x)$, the fact that the relationship is roughly linear in log-log coordinates indicates that across the range of word frequencies the number of errors stays relatively constant while the number of correctly spelled instances of a word grows, accounting for its increased overall frequency. When the sampling procedure is applied to a null population, the observed correlations are replicated 0/100000 times (as shown in Appendix 1), hence $p < .00001$. Given the null hypothesis accounting for at most $0.22 = 4\%$ of the variance in error rate as a function of frequency, we can be very confident that more than 50\% of the variance in error rate is a genuine effect of word frequency rather than an artifact of sampling. The full distribution for the proportion of variance accounted for by the effect of frequency on error rate but not by the null hypothesis is shown in Appendix 1. In addition, error rate on the web significantly correlates with word frequency estimated
What is it I am writing? 177

using an independent source, the Russian National Corpus: with rank transformed variables, \( r = -0.65 (R^2 = 42\%) \), \( t(291) = -14.61, p < .00001 \) for masculine singulars; \( r = -0.63 (R^2 = 40\%) \), \( t(179) = 10.80, p < .00001 \) for feminine singulars; \( r = -0.67 (R^2 = 45\%) \), \( t(173) = -11.79, p < .00001 \). Figure 1 also shows that there is no appreciable effect of gender/number: we observe the same relationship between word frequency and error rate for all forms of the word, and not just for the citation form. Thus it appears that lexical retrieval plays a role in the spelling of all forms of adjectives bearing bez and that the fact that the masculine singular is the citation form affects neither the error rate, not its relationship with word frequency.

A limitation of typed data for the purposes of analyzing factors influencing error rate is the possibility of automatic spellchecking. If spellchecking is done on the basis of a dictionary of stored wordforms, it could produce a lexical frequency effect in the data, since frequent words are more likely to be found in the spellchecker’s dictionary than rare words. In order to test the influence of the spellchecker, I inputted all correct spellings of the words of interest into Microsoft Word and checked their spellings. While almost all spellings of verbs were in the spellchecker’s dictionary, this was true for only about 2/3 of the bez-bearing adjectives. Error rate was significantly influenced by whether or not the word could be spellchecked: mean error rate for words that could be spellchecked is 8% for masculine forms, 12% for feminine forms, and 10% for plural forms while mean error rate for words that could not be spellchecked is 31% for masculine or feminine forms and 35% for plural forms. The differences between words that could and could not be spellchecked are significant after partialling out Google word frequency \( (t(289) = 2.65, p = .009) \), suggesting that the effect of spellchecker is not fully attributable to differences in

Figure 1. Correlation between word frequency on Google (in thousands of hits) and error rate for masculine singular (\( N = 366 \)), feminine singular (\( N = 177 \)) and plural (\( N = 183 \)) nominative forms of adjectives with bez-. The graph shows regression lines (\( R^2 \) values are on the lower right) and the results of significance tests for each subset.\(^{18}\)
Google frequency between words that could and could not be spellchecked. However, it is important to note that words that could not have been spellchecked are also, in their great majority (84%) words that are not found in the Russian National Corpus, whereas the great majority (89%) of words that could be spellchecked are found in the corpus. Thus, the “spellchecker effect” could also be seen as a lexicality effect distinguishing novel words from words that have been experienced (see Pagliuca et al. 2008 for experimental evidence of the lexicality effect in reading the regular Italian orthography). Table 1 shows that there also appears to be an effect of Google frequency that is not captured by whether the word could have been spellchecked, at least for words that could have been spellchecked (corpus frequency estimates yield $r = -.56$, $t(164) = 8.72$, $p < .000001$ for masculine singulars; $r = -.53$, $t(116) = 6.76$, $p < .000001$ for feminine singulars; $r = -.52$, $t(113) = 6.53$, $p < .000001$ for plurals). Given that very few words that could not have been spellchecked occur in the corpus, it is, perhaps, no wonder that the frequency estimates for those words are unreliable (the correlation between corpus frequency and Google frequency is only $r = .06$ for singular, $r = .01$ for feminine, and $r = .3$ for plural forms) and not very predictive of error rate (n.s., except for Google frequency with masculine forms, as shown in Table 1).

2.2.2 Adjectives with bez-: Within-lexeme analysis.

Given that the masculine singular is the citation form, and the most commonly occurring form in the paradigm for most adjectives, we might have expected Russian speakers to memorize masculine singulars and derive other adjective forms from them (e.g., Albright 2008). If this is true, then error rates for all adjective forms should be best predicted by the frequency of the corresponding masculine singular. The relevant data are shown in Table 2 and are clearly inconsistent with this hypothesis: 1) the error rate for a wordform is best predicted by the frequency of that wordform, not the frequency of the inflectional base, and 2) the error rate for the base form (top row) is just as predictable given other wordform frequencies as the error rate for other wordforms.

There are a number of words for which data from multiple forms are available. Forms of the same word often differ in frequency. For instance, the adjective besčislenn- ‘countless’ much more commonly modifies plural nouns than singular nouns. If writers rely on retrieval of individual wordforms, we should see that even within a lexeme the more frequent wordforms should have a lower error rate. This hypothesis is tested in Figure 2. Once again, we see a negative correlation between wordform frequency and error rate ($t(180) = -8.5$, $p < .000001$ for the masculine vs. feminine comparison; $t(174) = -6.9$, $p < .000001$ for the masculine vs. plural comparison), even for wordforms differing only in the inflectional ending, suggesting that writers rely on retrieval of morphologically complex, regularly inflected wordforms for accurate spelling.
What is it I am writing?

Table 1. Results of significance of tests of rank-order correlations between word frequency and error rate within subsets of the data defined by presence/absence in the spellchecker's dictionary, minimum error rate and frequency.

<table>
<thead>
<tr>
<th></th>
<th>Minimum frequency (hits)</th>
<th>Minimum error rate</th>
<th>Could have been spellchecked</th>
<th>Could not have been spellchecked</th>
<th>Frequency by spellcheck interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Masculine</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>.1</td>
<td>t(40) = 1.99, p = .06</td>
<td></td>
<td>t(100) = 2.79, p = .006</td>
<td>n.s.</td>
</tr>
<tr>
<td>1000</td>
<td>.01</td>
<td>t(132) = 9.52, p &lt; .00001</td>
<td></td>
<td>t(84) = 4.17, p = .00008</td>
<td>n.s.</td>
</tr>
<tr>
<td>10000</td>
<td>.001</td>
<td>t(92) = 7.57, p &lt; .00001</td>
<td>N/A (N = 0)</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td><strong>Feminine</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>.1</td>
<td>t(40) = 3.63, p = .0008</td>
<td>t(51) = 1.04, n.s.</td>
<td>N/A</td>
<td>n.s.</td>
</tr>
<tr>
<td>1000</td>
<td>.01</td>
<td>t(92) = 6.39, p &lt; .00001</td>
<td>N/A (N = 9)</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>10000</td>
<td>.001</td>
<td>t(68) = 5.56, p &lt; .00001</td>
<td>N/A (N = 0)</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td><strong>Plural</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>.1</td>
<td>t(31) = 1.09, n.s.</td>
<td>t(52) = 1.34, n.s.</td>
<td>N/A</td>
<td>n.s.</td>
</tr>
<tr>
<td>1000</td>
<td>.01</td>
<td>t(84) = 5.76, p &lt; .00001</td>
<td>t(18) = 1.44, n.s.</td>
<td>N/A</td>
<td>n.s.</td>
</tr>
<tr>
<td>10000</td>
<td>.001</td>
<td>t(69) = 5.26, p &lt; .00001</td>
<td>N/A (N = 0)</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>
Finally, it is important to note that if one form of a word is in MS Word’s spellchecker’s dictionary, then all forms of that word are, thus within-lexeme frequency effects cannot be due to the operation of the spellchecker.

2.2.3 Verbs with iz-

Figure 3 shows that the spelling of the prefix -iz, like the spelling of bez-, appears to be dependent in part on lexical retrieval. Simulations confirm the significance of the frequency-error rate correlation (p < .00001 for the infinitive, p = .0003 for the past tense, p = .009 for the present tense, p = .03 for future tense, see Appendix 1). The correlations are also significant with corpus-based frequency estimates (for present tense forms, r = -.67, t(23) = -4.32, p = .0003; for future forms, r = -.53, t(23) = -2.98, p = .006). There is, in addition, a (barely) significant correlation between the difference in frequencies between the infinitive and future forms and the difference in error rates (r = -.4, p = .042, Figure 4).

Table 2. $R^2$ values for correlations between word frequency (columns) and error rate (rows) for masculine singular, feminine singular, and plural nominative adjective forms.20

<table>
<thead>
<tr>
<th></th>
<th>Masculine</th>
<th>Feminine</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masculine</td>
<td>.63</td>
<td>.53</td>
<td>.45</td>
</tr>
<tr>
<td>Feminine</td>
<td>.40</td>
<td>.61</td>
<td>.42</td>
</tr>
<tr>
<td>Plural</td>
<td>.49</td>
<td>.51</td>
<td>.65</td>
</tr>
</tbody>
</table>
2.2.4 Verbs with roz-

Pilot searching showed that a large proportion of apparent errors retrieved by searching for past tense forms of roz- verbs came from Bulgarian texts. The infinitive is the citation form, hence data on infinitives may be contaminated by dictionary and textbook sources. Thus, we concentrate on present and future 3rd person singular forms, which are not found in dictionaries and are not shared with Bulgarian. The data are shown in Figure 5. The figure shows apparent negative correlations between word frequency and error rate. The correlation between error rate and word frequency for present tense forms is not significant according to the simulations (p = .35, see Figure 5). For future
forms, the correlation is highly significant with simulations (p = .00002, see Appendix 1) or with corpus-based frequency estimates, r = −.24, t(135) = .004).

Furthermore, Figure 6 shows that, albeit weakly, the difference in frequency between present and past tense forms correlates with a difference in error rates (t(90) = −2.47, p = .013); this correlation is significantly weaker than the comparable correlations for bez- (t(267) = 1.85, p = .01 for the feminine vs. masculine bez- adjectives, t(263) = 1.88, p = .06 for the masculine vs. plural bez- adjectives).

2.2.5 Verbs with pod-
For the verbs prefixed with pod-, a sizeable correlation between word frequency and error rate (shown in Figure 7) is obtained only for the past singular mascu-
What is it I am writing?

line form, however simulations show that the correlation could have resulted from the sampling procedure (p = .17). No significant correlation is obtained for either present or future third person singular (R² < .1 for the future form, R² < .025 for the present form).

2.2.6 Verbs with ot-
There are very few verbs with ot- for which errors are found. Among these verbs, there is an apparent correlation between word frequency and error rate for the present tense third person singular form but it is not significant (Figure 8, p = .68 according to the simulations).

2.2.7 Verbs with s-
There is no correlation between word frequency and error rate for the prefix s- (Figure 9, R²(54) = .02).

2.2.8 Comparing affixes
The prefixes iz-, roz-, and bez-, which have corresponding differently-spelled homophonous words, show statistically reliable effects of word frequency on the accuracy of spelling the prefix-final consonant. In this section, we examine whether the strength of the observed relationship between word frequency and
error rate differs across affixes. There are two ways in which the strength of the relationship can be measured. First, the stronger a correlation, the more the slope of the corresponding regression line deviates from zero. Second, the stronger a correlation, the more variance in the dependent variable is accounted for by the independent variable. Table 3 compares the relationships between frequency and error rate observed for the prefixes bez- and iz- in terms of regression line slopes (tested using interactions in ANOVA). The slope of the relationship is significantly shallower for iz- than for bez- indicating that the spelling of iz- is less sensitive to word frequency than the spelling of bez-. The amount of variance in error rate accounted for by frequency alone is also higher for bez- than for iz- (p < .00001 based on t-tests between samples of $R^2_{\text{observed}} - R^2_{\text{null}}$ shown in Appendix 1. In addition, bez- consistently shows a higher error rate when frequency is controlled (being a covariate in the ANOVA).

Table 4 compares the relationships between frequency and error rate observed for iz- to the relationship observed for roz-. The slope of the relationship is significantly shallower for the present tense forms of verbs prefixed with roz- than for verbs prefixed with iz-. The future forms of verbs prefixed with roz- show a frequency effect that is as robust as the one found for iz- based on slope. The effect is weaker than the effect for past, present and infinitive forms of verbs bearing iz- based on variance in error rate accounted for by frequency but is stronger than the frequency effect observed with future forms of iz-.

Table 5 shows that the effect of word frequency on the spelling of future forms of verbs prefixed with roz-, while not being consistently less robust as the effect found for verbs prefixed with iz-, is significantly less robust than the effect of frequency on the spelling of adjectives prefixed with bez-. The effect of frequency on spelling accuracy observed for roz- is also significantly weaker in terms of the variance in error rate accounted for by frequency (p < .00001). Furthermore, the error rate for bez- is higher than for roz- controlling for frequency.

Finally, the within-lexeme frequency effect observed for bez- is significantly stronger than the effect for roz- in terms of amount of variance in the difference
Table 3. Statistical tests of main effects of affix identity on error rate and interactions between word frequency and affix identity: the two affixes exhibiting the strongest frequency effects, bez- (rows) vs. iz- (columns). All interactions are in the direction of the word frequency effect on error rate being weaker for iz-. All main effects are in the direction of error rate being lower for iz- than for bez-. Since all words bearing iz- are in Microsoft Word’s spellchecker’s dictionary, only bez-bearing words found in Microsoft Word’s dictionary are included in this analysis.

<table>
<thead>
<tr>
<th>Affix</th>
<th>Frequency*Affix</th>
<th>Infinitive</th>
<th>Past</th>
<th>Present</th>
<th>Future</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masculine</td>
<td>Affix</td>
<td>t(256) = 3.36, p = .0009</td>
<td>t(213) = 1.9, p = .06</td>
<td>t(187) = 3.12, p = .002</td>
<td>t(188) = 1.25, n.s.</td>
</tr>
<tr>
<td>Feminine</td>
<td>Affix</td>
<td>t(256) = 5.12, p &lt; .00001</td>
<td>t(213) = 2.5, p = .01</td>
<td>t(187) = 1.24, n.s.</td>
<td>t(188) = 3.22, p = .0015</td>
</tr>
<tr>
<td>Plural</td>
<td>Affix</td>
<td>t(208) = 4.69, p &lt; .00001</td>
<td>t(165) = 2.98, p = .003</td>
<td>t(139) = 3.95, p = .0001</td>
<td>t(140) = 2.1, p = .037</td>
</tr>
</tbody>
</table>

Table 4. Statistical tests of main effects of affix identity on error rate and interactions between word frequency and affix identity: roz- (rows) vs. iz- (columns). Positive interactions are in the direction of the word frequency effect on error rate being stronger for iz-. All main effects are in the direction of error rate being lower for roz- than for iz-.

<table>
<thead>
<tr>
<th>Affix</th>
<th>Frequency*Affix</th>
<th>Infinitive</th>
<th>Past</th>
<th>Present</th>
<th>Future</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>Affix</td>
<td>t(241) = 3.65, p = .0003</td>
<td>t(198) = 2.94, p = .004</td>
<td>t(172) &lt; 1, n.s.</td>
<td>t(173) = 2.3, p = .022</td>
</tr>
<tr>
<td>Future</td>
<td>Affix</td>
<td>t(241) = 4.38, p = .00002</td>
<td>t(198) = 2.4, p = .014</td>
<td>t(172) = 1.36, n.s.</td>
<td>t(173) = 2.7, p = .008</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Affix</th>
<th>Frequency*Affix</th>
<th>Infinitive</th>
<th>Past</th>
<th>Present</th>
<th>Future</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>Affix</td>
<td>t(231) &lt; 1, n.s.</td>
<td>t(188) &lt; 1, n.s.</td>
<td>t(162) &lt; 1, n.s.</td>
<td>t(163) &lt; 1, n.s.</td>
</tr>
<tr>
<td>Future</td>
<td>Affix</td>
<td>t(231) &lt; 1, n.s.</td>
<td>t(188) &lt; 1, n.s.</td>
<td>t(162) &lt; 1, n.s.</td>
<td>t(163) &lt; 1, n.s.</td>
</tr>
</tbody>
</table>
in error rate between wordforms accounted for by differences in frequency between the same wordforms (p < .001).

To summarize, both between-lexeme and within-lexeme frequency effects on spelling accuracy are especially strong for the prefix bez- and the error rate for a word of a given frequency is higher if the word bears bez- than if it bears one of the other prefixes.

3. The experiment

While it would be interesting to test whether the form-specific frequency effects seen with verbs can be replicated experimentally, the next section presents only results for the prefix bez-. This is due to the very low error rate seen with other prefixes. While the vast dataset of the web allows us to distinguish between an error rate of .01 and .001, this can only be accomplished experimentally with thousands of subjects, even assuming the same error rate in an experimental task and informal, anonymous writing on the Internet.

3.1 Methods

The data come from a classroom dictation test of orthographic knowledge in which a sample of 25 words in the feminine singular nominative form bearing bez- was presented to 51 first-year native Russian university students in Nizhny Novgorod, Russia, in a sentence context. The students were asked to write down the sentences by hand so there was no possibility of automatic spellchecking. The task was speeded, with sentences presented one after the other, which resulted in a low incidence of corrections on the relevant trials (<2%). All sentences were read by the same female native speaker of Russian. The students were graded on their spelling performance and thus had high motivation to perform accurately.

The feminine form was chosen so as to avoid using the citation form. Twenty-four of the 25 words were selected by random sampling of feminine
What is it I am writing? 187

wordforms bearing bez- found on the web. The 25th word is the word besplatnaja ‘free’, which was the most frequent lexeme in web data. The words were presented in sentences taken from the first page of erroneous hits retrieved from Google, which were interspersed with filler sentences that did not contain the prefix bez- and tested other orthographic rules, including 25 sentences containing words with the prefix s- before a voiced consonant, which had a very low error rate (<1%), the spelling of pre- and pri- and words with silent consonants (100 sentences total). The sample of sentences containing bez- before a voiceless consonant is shown in Appendix 2. There were no words where bez- occurred in an environment where errors are not expected. Figure 10 shows the correlation between error rate and word frequency observed for these words in Google data. Sentences were presented in random order. There was no correlation between order in the stimulus sequence and frequency (–.09 < r < −.04, .6 < p = .85 depending on frequency estimate).

3.2 Results

Figure 11 shows the correlation between error rate in the experimental sample and frequency on the web. The frequency estimates from the base (masculine) forms appear to correlate with the experimental data more strongly than the frequency estimates for the forms that were actually presented, although the 95% confidence intervals for the correlation coefficients based on masculine and feminine frequencies overlap, so the difference in correlation strengths is not statistically significant. Regardless of whether the estimate of frequency is based on masculine or feminine forms, and whether it is based on Google or corpus data, the correlation between frequency and error rate is significant (adjusting for non-normality and making the Google data comparable to the corpus data, which contains zeroes, using the rank transformation, r = −.68 ± .28, t(24) = −4.58, p = .0001 for masculine Google estimates; r = −.49 ± .35, t(24) =
−2.75, \( p = .01 \) for feminine Google estimates; \( r = −.69 \pm .24, t(24) = −4.68, p < .00001 \) for corpus-based masculine estimates; and \( r = −.79 \pm .21, t(24) = −6.4, p < .00001 \) for corpus feminine estimates). The higher the frequency of the word, the fewer the errors made in spelling the prefix even in a high-stakes classroom-dictation dictation task explicitly testing knowledge of orthographic rules. There was no significant correlation of error rate with order of presentation (\( r = −.08, t(24) = −.4, p = .7 \)).

4. General discussion

To summarize, we observe strong frequency effects in the spelling of Russian /z/-final prefixes despite prefix spelling being regular and taught explicitly as a regular rule. The frequency effects are stronger for the adjectival prefix bez-than for verbal prefixes. This result is counter to what one would expect if the major predictor of reliance on retrieval were regularity (Pinker 1991) or contextuality (Bertram et al. 2000) of inflection: adjectival inflection is regular and purely contextual while verbal inflection is irregular and partly inherent. At least for bez-, there are frequency effects even within lexemes, indicating retrieval of regularly inflected wordforms. These results raise two questions:

(12) Why are there whole-word frequency effects in this regular system, and can we generalize from these findings to other regular systems?
(13) Why are the frequency effects especially strong for bez-, the most regular of all prefixes? What makes Russian speakers rely on retrieval of inflected wordforms more when they are spelling bez- than when they are spelling other prefixes that obey the same rule?

I argue that the answer to (12) is that all apparently regular systems actually feature competition between generalizations because processing always
What is it I am writing? 189

involves temporary uncertainty regarding which rules are applicable. If there is much uncertainty, the learner may give up on the rule(s) and rely heavily on lexical retrieval. This hypothesis provides a natural answer to (13): writers rely on lexical retrieval in spelling bez- more than in spelling other prefixes because there is more uncertainty regarding which spelling rules are applicable during the spelling of bez- than during the spelling of any other prefix.

4.1 How many bez’s are there in Russian?

The main reason for reliance on lexical retrieval in spelling Russian /z/-final prefixes (bez-, iz-, roz-) may lie in that the rules that apply to a prefix-final consonant are different from the rules that apply to a word-final consonant. Thus in order to apply the rules correctly, the writer must know that s/he is dealing with a prefix. This may be particularly difficult for the prefix bez-, which appears to have a much higher error rate than the other (verbal) prefixes, despite adjectival inflection being regular and verbal inflection being irregular, which could have encouraged whole-word storage for verbs and discouraged it for adjectives. Application of the orthographic rule may be difficult for bez- because bez-the-prefix, is hard to distinguish from bez-the-preposition. In fact, when a listener hears [bes], s/he may not know whether the [bes] is a prefix or a preposition until after the following stem is processed. Likewise, when bez- is being produced, the writer or speaker may not be aware of whether they are producing the prefix or the preposition: since bez-initial adjectives are often derived from bez-initial prepositional phrases, bez- and bez may be the same morpheme for speakers of Russian, at least before literacy training begins.

The possibility that bez- and bez are the same morpheme for illiterate Russian speakers raises the fascinating possibility that a difference in separability and therefore wordhood does not necessarily translate into a difference in lexical representations. Rather than having bez- and bez, Russian may possess a rule that states that adjectives can only be derived from prepositional phrases in which the noun immediately follows the adjective, which explains why words like *bessineborodyj ‘blue-beard-less’ do not exist. The inseparability of bez- when it is followed by an adjective is then a simple consequence of this rule. Other rule-based or constraint-based solutions are, of course, possible. The crucial point is that bez-initial adjectives appear to be derived from bez-initial prepositional phrases, as indicated by the fact that a bez-initial adjective often has a corresponding bez-initial prepositional phrase while lacking a corresponding bez-less adjective, as shown by the examples in (8) and (9), while the reverse pattern is not observed. The fact that bez- can be separated from the noun in the prepositional phrase but not in the adjective does not necessarily mean that the bez in the former is different from the bez- in the latter, and in
fact Russian speakers appear not to reliably distinguish the two despite years of explicit training.

There are at least two advantages of assuming that bez- and bez are the same unit. From a theoretical standpoint, it is more parsimonious to claim that bez- is the same morpheme throughout the derivation of an adjective form than to claim that it starts out as a separate stem and then changes status to a prefix. From the standpoint of explaining the spelling data, assuming that bez- and bez are the same lexical entry would mean that Russian speakers can be uncertain regarding which bez they are producing in normal, endogenously-driven writing, and not just writing from dictation.

4.2 The nature of competition and the absence of true regularity

There are two possible ways in which the competition from bez-the-preposition can interfere with correct spelling of bez-the-prefix. One possibility is that during lexical retrieval, the prepositional phrase competes with the corresponding adjective (Largy et al. 1996, Morton 1980, Sandra and Fayol 2003, Sandra et al. 1999, 2004). Then, incorrect spellings of bez- would result from the phrase temporarily dominating the competition against the adjective (the dominance has to be temporary to explain why the adjective is spelled as a single word despite having a ‘z’). Alternatively, the existence of bez-the-preposition could establish a competition between morphemes or rules of the grammar, resulting in a shift in favor of lexical retrieval for all words containing bez-rather than just words that have a strong phrasal competitor.

In order to see whether errors in the spelling of bez- could be explained by lexical competition between the adjective and the corresponding prepositional phrase, I correlated the frequency of the prepositional phrase relative to the frequency of the adjective with error rate. Relative frequency had only a trend towards significance in the expected direction (with rank-transformed variables, $t(269) = -1.74, p = .08$) and accounted for only 2% of variance in error rate above the frequency of the adjective, which accounts for over 40% of the variance. For the experimental (dictation) data, the frequency of the prepositional phrase is significantly correlated with error rate on the adjective but in the unexpected direction: the higher the frequency of the phrase, the lower the error rate (after controlling for adjective frequency, $-2.153 > t(24) > -2.604, .015 < p < .045$ depending on the adjective frequency estimate). Thus, it does not appear that the errors are due to the phrasal parse winning over the adjectival parse. Rather it appears that the existence of the preposition establishes an alternative orthographic rule or an alternative morpheme spelling that acts as a default, being used when lexical retrieval fails (Clahsen 1999, Pinker 1991, Pinker and Prince 1988).
An additional piece of evidence for blaming a competing rule or morpheme spelling for the high sensitivity of the spelling of bez- to lexical frequency is that a strong correlation between word frequency and error rate is found even for those words bearing bez- that have a low error rate that is similar to the error rate for other prefixes, which do not exhibit the strong correlation bez-does. Thus it would appear that the shift in favor of full-form retrieval is made for all words containing bez-, rather than at the level of individual problematic words.

The fact that it is the ‘z’ spelling that acts as a default, and, unlike the prefix, the preposition is almost never erroneously spelled with ‘s’ is consistent with the idea that adjectives are derived from prepositional phrases (which are always spelled with ‘z’), while prepositional phrases are not derived from adjectives. Thus, the ‘z’ associated with the preposition could erroneously persist and surface in the derivation of an adjective while an ‘s’ from an adjective is never activated during the derivation of a prepositional phrase and thus has little to no chance of being produced. Alternatively, ‘z’ could be the default due to the strength of the morphological principle (orthographic paradigm uniformity constraints) in Russian spelling, where morphemes tend to have a constant spelling reflecting the underlying form.

One could argue that the spelling of Russian prefix-final consonants is irregular because there is temporary uncertainty regarding which rule (the one for /z/-final prefixes or the one for stems) should apply, which in the worst case scenario of no preceding context is resolved by the suffix. It is important to note that under this definition there are no regular grammatical systems, as temporary uncertainty about the shape and structure of the input, and thus about which rules can apply to the input, is inevitable. Thus, rather than contrasting regular and irregular systems it may be more productive to discuss and ideally measure the degree and temporal extent of uncertainty the speaker/writer has regarding the applicability of various grammatical generalizations during language production (see Grosjean 1980 and Tanenhaus 2007 for promising methodologies).

The importance of temporary uncertainty regarding rule applicability is also consistent with the findings of Barca, Burani and Ellis (Barca et al. 2007, Burani et al. 2006), who examined frequency effects in the reading aloud of words spelled in the highly regular Italian orthography. While the spelling-to-sound correspondence rules were regular in all words, some words contained letters whose pronunciation could be determined only with reference to the orthographic context, e.g., the letter ‘c’ is pronounced [tʃ] before front vowels but [k] elsewhere. Naming times for words that contained letters with context-sensitive spelling were influenced by word frequency more than naming times for words containing only letters with context-insensitive spellings. A context-sensitive rule involves a period of uncertainty during which it is not clear.
whether the rule should apply, thus it may be less likely to be applied productively. In the spoken domain, this hypothesis is also supported by the results of Warker and Dell (2006), who found that context-insensitive phonotactic restrictions were easier to learn than context-sensitive ones, and by Newport and Aslin (2004), who found that dependencies between non-adjacent elements were harder to learn than dependencies between adjacent elements. In addition, Albright (2009) observed that paradigm gaps tend to occur with stems to which multiple rules can apply with approximately equal probability, resulting in high uncertainty, which may cause both rules to fail.

High uncertainty regarding rule applicability taxes working memory, resulting in higher probability of working memory overload and error due to either not noticing the presence of a required part of the rule’s structural description or not noticing something about the input that contradicts the rule’s structural description. The involvement of working memory in spelling errors is also suggested by Frisson and Sandra (2002) and Sandra et al. (1999), who found that errors in spelling homophonous subject agreement morphemes on verbs become more likely as distance between the verb and the subject increases. If applying the rule often leads to incorrect results due to working memory failures, while lexical retrieval usually works, reliance on lexical retrieval can be seen as rational.

The observed absence of an effect of prepositional phrase frequency contrasts with the data reported by Largy et al. (1996), who observed that the spelling of silent inflections of French verbs is strongly influenced by the frequency of the homophonous noun. Largy et al.’s data strongly suggest that the erroneous noun spelling surfaces as a result of the noun representation winning the lexical competition over the verb representation whereas the data in the present study suggest that the error results from application of a grammatical generalization or the use of a default spelling for a morpheme.24 Thus it appears plausible that spelling errors can result from competition between either lexical representations, grammatical generalizations, or lexical representations competing with grammatical generalizations (see also Frisson and Sandra 2002, Largy et al. 2007, Sandra et al. 1999). One difference between the two studies that may influence the results is that the errors examined in the present study cannot result from an erroneous orthographic lexical representation achieving complete victory over the correct one because the adjectives spelled with ‘z’ are nonetheless spelled as single words. For these errors to be due to dominance in lexical competition, the dominance would have to happen just in time to influence the letter choice but not to result in spelling the prefix as a separate word. This just-in-time dominance may only happen relatively rarely. Thus, if errors can result from either grammatical or lexical competition, errors that do not match any lexical form may be more likely to have a grammatical trigger. Results suggesting the involvement of both grammatical generalizations and
lexical retrieval in spelling are also reported by Frisson and Sandra (2002) and Sandra et al. (1999), who find that the effect of distance between subject and verb on error rate in selecting the right silent subject-verb agreement morpheme to spell is independent of the frequency of the verb form. Alternatively, if the errors are the result of competition between morpheme spellings rather than orthographic rules associated with different morphemes, the discrepancy between the results of Largy et al. (1996) and the present study could be attributed to silent morphemes being poorer competitors than morphemes that are actually pronounced.

4.3 Alternative explanations for frequency effects

We will now consider the characteristics of the Russian spelling rules examined in the present study that distinguish them from other kinds of regular rules found in linguistic systems and ask whether some of the peculiarities of Russian spelling might be responsible for 1) the fact that whole-word frequency effects are present at all in what looks like a regular system (this section) and 2) the fact that bez- is more susceptible to frequency effects (section 4.4). If both of the effects can be explained by other factors, the status of uncertainty regarding rule applicability as an explanatory variable would be unconfirmed.

4.3.1 Age of exposure

A potential limitation on the generalizability of the present results is the fact that the rules examined here are learned late, which may make them especially susceptible to word frequency effects (Ullman 2005). The late acquisition is due in part to the rules being orthographic rules, thus the relevant exposure tends to start shortly before children enter school at the earliest. The late exposure, however, does not explain why acquisition of the rule appears to be easier for verbal prefixes compared to the adjectival prefix bez-, leading to reliance on retrieval of complete inflected wordforms for the spelling of bez-initial adjectives to an extent unparalleled during the spelling of verbs despite verbal inflection being irregular and adjectival inflection being regular. This difference between prefixes we suggest is due to the higher uncertainty about the identity of the to-be-spelled morpheme during production of bez- than during the production of any other prefix. We hypothesize that the extent of temporary uncertainty is relevant to all kinds of rules, whether learned late or early, and whether used in writing or in speech.

4.3.2 The nature of training

While textbooks of Russian orthography represent prefix spelling as a regular system of exceptionless rules and rules are explicitly taught, which should discourage reliance on lexical retrieval (Robinson 1997, Williams 1999), the way the knowledge of that system is taught and tested may not discourage reliance
on lexical retrieval as much as it could have. Figure 12 shows which words with the prefix bez- are presented to students of Russian orthography in exercises from a convenience sample of popular workbooks used for review exercises in high school and as part of college entrance exam preparation (Bushueva 1994, Fridman et al. 2005, Golub and Rozental’ 1990, Rozental’ 1998, Rozental’ and Golub 1999). As can be seen in the Figure, the words on which the learners are trained and tested, particularly in the case of bez-, tend to be the high-frequency words (the frequencies of tested and untested words are significantly different: with rank frequency, t(80) = 11.4, p < .00001 for corpus-based; t(100) = 11.96, p < .00001 for web-based frequency estimates).

High-frequency words can be spelled correctly by either applying the orthographic rule or retrieving the word’s orthographic representation. Thus, knowledge of the orthographic rules is not optimally tested for bez- by the examined workbooks, allowing students to complete the exercises by relying on lexical retrieval. While bez- is found in many words that are less frequent than the words Russian writers are usually taught to spell, this is not the case for roz- because roz- is less productive than bez-. Importantly, the number of tested bez-bearing words is even higher than the number of tested roz-bearing words and the frequencies of tested bez-bearing words are not significantly higher.
than the frequencies of tested roz-bearing words. Therefore, differences in testing do not appear to explain why the bez- error rate is an order of magnitude higher than the roz- error rate even after (log) word frequency is partialled out (compare error rates for the same frequencies in the two graphs in Figure 25) and why the relationship between word frequency and error rate is stronger for bez-.25

It is important to note that the correlation between word frequency and error rate for bez- holds for both words found in workbooks and words not found in workbooks: for words found in workbooks, $r = -.78$, $t(46) = -8.37$, $p < .00001$ for web-based estimates, $r = -.61$, $t(46) = -5.27$, $p < .0001$ for corpus-based estimates; for words not found in workbooks, $r = -.75$, $t(244) = -17.78$, $p < .00001$ for web-based estimates, $r = -.57$, $t(244) = -11.03$, $p < .0001$ for corpus-based estimates). Thus, the effect of word frequency is not an artifact of the frequent words being more likely to be trained and tested in school.

In the implicit learning situation experienced by a child acquiring their first language, even in the absence of a correlation between word frequency and the likelihood of the word undergoing a certain rule, the rule will also tend to be exemplified by a few high-frequency words due to word frequencies being distributed in accordance with Zipf's law (Yang 2010). Thus, in this respect the orthographic rules examined in the present study are like the rules of spoken language. Thus it is not that the orthographic rules of Russian are taught in a way that makes them more prone to being outcompeted by lexical retrieval than rules acquired implicitly. They are taught in a way that makes reliance of lexical retrieval less likely than if they were learned implicitly (see also Robinson 1997, Williams 1999). However, they could be taught in a way that would discourage reliance on lexical retrieval even more.

4.3.3 Morphologization

The rules examined here are also clearly morphologically conditioned, and morphological conditioning tends to correlate with low productivity to such an extent that Bybee (2008) calls it a universal of language change. This low productivity may lead the rules to lose the competition to lexical retrieval (Bybee 2008). However, the rules are morphologically conditioned for all of the prefixes included in the present study to the same extent but they are much less productive for bez- than for the other prefixes. Thus morphological conditioning does not explain away the effect of temporary uncertainty. On the other hand, the extent of temporary uncertainty during rule application may help explain why rules lose productivity when they acquire morphological conditioning. Morphological structure is one more thing that needs to be evaluated before the decision on whether to apply the rule can be made, increasing temporary uncertainty, working memory load, and the likelihood of error. Furthermore, morphological conditioning is likely to be especially fatal to rule
productivity because a morphologically conditioned rule can only be applied after lexical retrieval has already been to some extent completed. Thus, morphological conditioning gives lexical retrieval a headstart over the rule, making it especially likely that the rule will lose the competition.

4.3.4 General tendencies in Russian spelling
Another possible reason for the examined rules to be relatively unproductive is that they violate the general tendency within Russian to assign each morpheme a consistent spelling (the morphological principle, Avanesov 1964, Sproat 2002). Avanesov (1964), arguing for spelling reform, explicitly mentions that /z/-final prefixes are an exception to the morphological principle, which may make their spelling harder to learn. Sproat (2002) suggests that languages tend towards having phonological-to-orthographic mappings that involve a consistent level in the phonological derivation and that when multiple levels are involved, the system is likely to change (either in terms of orthography or pronunciation). Again, however, this factor does not explain the differences in productivity of the rule across prefixes.

While the spelling of prefixes in Russian is regular, it is not the case that all of Russian orthography is regular or transparent. This may also lead to a relatively high degree of reliance on lexical retrieval, although see Barca et al. (2007), Burani et al. (2006), Pagliuca et al. (2008) and Raman (2003) for evidence of lexical frequency effects in regular transparent orthographies: Italian and Turkish (albeit in reading rather than spelling). This factor too does not account for differences across prefixes.

To summarize this section, there are some factors that may increase the susceptibility of the spelling of /z/-final prefixes to frequency effects: 1) late age of exposure, 2) training that mimics natural language in focusing on high-frequency words, 3) the presence of morphological conditioning, 4) the fact that not all of Russian orthography is regular, and 5) the fact that the rules in question disobey the morphological principle (high ranking of paradigm uniformity constraints) that other rules of Russian orthography obey. Factors (2) and (4) are shared with all or almost all other rules found in languages, while factor (1) may be largely restricted to orthographic rules. While important to explore further, these factors do not explain differences between prefixes that are explained by uncertainty regarding rule applicability and thus do not threaten uncertainty as an explanatory principle.

4.4 Alternative explanations for differences between prefixes

4.4.1 Parseability
One potential explanation for the differences between prefixes is that not representing the voicing assimilation across the prefix-stem boundary
orthographically serves as a boundary cue (see Hay 2003 for the same claim in phonology). The prefix bez- is the most productive prefix examined in the present study, as shown both by the fact that it forms more words and by the fact that the words in which it occurs tend to be less frequent than the words bearing the other prefixes (Baayen 1992, Hay 2003). Thus, the prefix-final consonant in bez- might be especially likely not to be devoiced in orthography (and, perhaps, also in pronunciation, although we have no evidence on that) because it is perceived as being a separate unit and thus is marked as such. However, if prefixes are spelled with ‘z’ when they are most parseable, we should see this effect within prefixes as well. Hay (2003) shows that affixes are most likely to be parsed out when the word without the affix is more frequent than the word with the affix, i.e., when the affixed word has a low frequency relative to its unaffixed counterpart. However, relative frequency does not explain any variance in spelling beyond what is explained by the frequency of the whole word for any of the prefixes in the present study and neither does unaffixed form frequency (all F < 1). Relative frequency is, of course, very strongly correlated with derived form frequency, thus one might argue that derived form frequency simply gobbles up all the variance relative frequency could have accounted for. However, the effect of derived form frequency is not in the direction predicted by the ‘z’-as-a-boundary-cue hypothesis: high-frequency words, which are presumably less likely to be parsed and more fused into an automated production unit (Kapatsinski 2010a), exhibit more ‘s’ spellings.

4.4.2 History of Russian spelling
Another alternative explanation for the difference between prefixes is offered by pre-1918 spelling rules. Prior to the spelling reform (Pokrovskij and Bonch-Bruevich 1918), which resulted in the current rules, bez- was exceptional among the /z/-final prefixes in always being spelled with ‘z’ (Grot 1886: 46–48). The others were (starting in 1874 when spelling rules were standardized) to be spelled with ‘s’ before all voiceless consonants except ‘s’ where it was spelled ‘z’ (‘zs’ perhaps really being a boundary cue at that point in the history of the language). There are two reasons to doubt that this can account for the observed differences in error rate and sensitivity to word frequency between prefixes. First, texts typed before the 1918 reform are almost never seen by modern readers. Pre-revolutionary spellings might have snuck into the Google counts but not into the dictation data. Nonetheless, we observe very similar correlations between error rate and frequency in both datasets and very similar error rates as well. Second, the correlation between error rate and frequency for roz- before ‘s’, where roz- was always spelled with ‘z’ just like bez-, is still much lower than the correlation for bez- before ‘s’ (R² = .2 vs. .7, p < .004), thus the difference in susceptibility to frequency effects between bez- and roz- cannot be due to pre-revolutionary spellings sneaking into the dataset.
Furthermore, there is no effect of the following segment being ‘s’ on the spelling of roz- (F < 1), which would have been (near-)categorical before 1918. There is also very little error and no correlation between error rate and frequency for s- despite the fact that s- too was spelled with a ‘z’ before voiced obstruents at one (albeit earlier) time (Grot 1873: 14). Thus, it does not appear that pre-revolutionary spelling rules have influenced the present results. In fact, Grot (1874–1886) standardized Russian spelling in part based on observations of what writers were already doing. Thus, Grot (1873: 89) notes that, unlike other /z/-final prefixes, bez- was often being spelled with ‘z’ even before voiceless consonants, which is itself in need of explanation. The most plausible explanations to me are that 1) at that stage in the development of the language, ‘z’ was a boundary cue (which is why it was often spelled before ‘s’ for all prefixes), and bez- was especially parseable due to its high productivity and semantic compositionality of the resulting forms, and/or 2) bez-the-prefix and bez-the-preposition were perceived as being the same unit and constituted a single lexical entry, as they may be still.

4.4.3 Inflected words or prefabs?

At least for the prefixes bez-, iz- and roz-, there are frequency effects within lexemes, such that the more frequent wordforms within an inflectional paradigm are produced more accurately than the less frequent wordforms. It is important to note that differences in frequency between different forms of the same word are often driven by the existence of frequent collocations. This may be especially true for forms of the same word that differ by a contextual inflectional affix, like the adjectives in the present study. For instance, the feminine form of ‘sleepless’ is much more common than the masculine form because the word for ‘night’ is feminine in Russian, and ‘sleepless’ most commonly modifies ‘night’. Thus, the observed wordform frequency effects could reflect either reliance on retrieval of specific inflected wordforms or on retrieval of prefabricated multiword expressions, such as ‘sleepless night’ (see Arnon and Snider 2010, Bybee and Scheibman 1999, Kapatsinski and Radicke 2009, Sosa and MacFarlane 2002, and Tremblay 2009 for empirical evidence for storage and retrieval of multiword expressions). If different forms of the same adjective are especially likely to differ in frequency due to belonging to different collocations, this may explain why the within-lexeme frequency effects for bez- are especially strong (though see Bertram et al. 2000 for evidence that contextually-inflected wordforms are less likely to be stored and retrieved as wholes in perception). However, this does not explain why the error rate for bez- is especially high. Presumably the existence of the collocations would help the spelling of bez- if the writers could rely on the spelling rules equally easily for bez- and the other prefixes. Thus, while collocational frequency may account for the within-lexeme frequency effect, it does not explain why the
What is it I am writing?

spelling rules are less productive for bez- than for other prefixes. Uncertainty appears to still be needed to explain the data.

4.5 Methodological implications

On the methodological side the present paper reveals a number of challenges in studying the relationship between frequency and error rate, or the rate of occurrence of any (non-standard) feature that either does or does not occur in each token of a word. First, corpus data, and especially web data, come from a variety of speakers; thus the effects of lexical frequency on error rate could potentially be due to different speakers using high-frequency and low-frequency words. While low-frequency words may be especially likely to be used by well-educated writers with large vocabularies, who make few errors, low-frequency words are also especially likely to come from texts obeying different norms, including pre-revolutionary texts and texts produced by non-native speakers. Second, the target word in corpus data is not known, e.g., a ‘pot’ in the spelling could be either a misspelling of ‘pod’ [pot] or a misspelling of ‘po’ [po]. Third, typed data could be influenced by spell-checker software, which, if operating on the basis of lexical retrieval, could produce a word frequency effect on error rate (although the present effects are unlikely to be solely due to spellchecker behavior because they hold when presence in the spellchecker’s dictionary is controlled, at least for the most popular spellchecker program). These difficulties are circumvented by using experimental data in conjunction with corpus data.

On the other hand, experimental/elicited data are limited in that an elicitation task, and especially a classroom dictation task, can make the respondent focus on the orthography to an atypical extent, causing the writer to try to rely on prescriptive orthographic rules much more than s/he does in everyday life. (Although in the present study the dependence between word frequency and error rate is just as strong in the elicited data as in the corpus data.) Experimental data are also limited in their resolution at very low error probabilities, requiring unfeasible numbers of subjects to be run to distinguish between, for instance, an error rate of .001 and an error rate of .0001, which is a particular concern given the power-law relationship between error rate and frequency. Finally, the uncertainty about whether the letter string being written is an adjective or a prepositional phrase is necessarily greater with exogenously-driven writing (dictation) than with endogenously-driven writing. Thus corpus and experimental data complement each other and are most informative in combination (see also Arppe and Järviči 2007, Gilquin and Gries 2009 and other contributions to Corpus Linguistics and Linguistic Theory 5(1)).

Issues of spurious correlations arise when error rate is correlated with frequency in the same corpus because the number of errors is divided by
frequency to derive the error rate. We developed a Monte Carlo simulation approach to this issue that allows for adequate statistical correction for the spurious correlation. The non-spurious nature of the observed frequency/error-rate correlations is also reinforced by obtaining independent frequency estimates from a different corpus and independent error rate estimates from an elicitation task.26

5. Conclusion

Lexical frequency effects were found in a fully regular system that is taught as a regular rule system, the spelling of prefix-final consonants in Russian (see also Largy et al. 1996, Sandra et al. 1999, Sandra and van Abbenyen 2009 and Tabak et al. in press). Furthermore, regularly inflected wordforms are stored in memory and retrieved in production even when the inflectional suffix is a pure agreement suffix (contra Baayen et al. 1997 and Bertram et al. 2000). The present data indicate that forms that can be derived by rule in a fully regular system are nonetheless stored and retrieved in production (contra DiSciullo and Williams 1987, Pinker 1991).

The reliance on retrieval is greatest for the rule whose structural conditions are most difficult for the writer to evaluate: the rule that says that requires the input consonant at the end of [bes] to be identified as prefix-final rather than preposition-final. It is suggested that the prefix and the preposition may be the same unit / lexical entry for Russian speakers prior to acquisition of literacy (the bez-initial adjectives, in which it appears to be a prefix based on the criterion of inseparability, are derived from bez-initial prepositional phrases, in which bez can be separated from the noun). Thus in order to acquire orthography Russian writers have to learn to split one lexical entry in two and are often uncertain regarding whether they are producing bez-the-prefix or bez-the-preposition even after years of training. In addition, an analysis of workbooks shows that Russian students appear to be mostly tested on their knowledge of spellings of frequent words, which allows the students to pass the tests by relying on lexical retrieval rather than application of orthographic rules, just like speakers of a language do not have to learn a phonological or morphological rule if they can simply retrieve the words exemplifying the rule from the lexicon (which Bybee 1985 suggests to be what leads rules exemplified by frequent words to lose productivity).

The present results suggest that a distinction between regular and irregular systems may be a poor description of the variance in grammatical systems and the degree to which a grammatical system can compete in efficiency against lexical retrieval. Rather, we need to develop ways to quantify uncertainty speakers have about which rules can or should apply.
Appendix 1: Results of Monte Carlo simulations

The code is available at http://uoregon.edu/~vkapatsi/FreqVsErrorRate.R

The distributions of correlations in 100,000 samples from a population where the error rate is constant across word frequency while word frequency is distributed as in the actual samples. The dashed line shows the correlation observed in the actual data.
The amount of variance accounted for by word frequency but not by the null hypothesis.

For *bez*-

![Histograms for *bez*](image)

For *iz*-

![Histograms for *iz*](image)

For *roz*-

![Histograms for *roz*](image)

**Appendix 2: Experimental stimuli / Top erroneous hits on Google for feminine adjectives with *bez***

Cluster: [sts]
Мазь, безцветная при обычном освещении, светится ярко красным в ультрафиолетовом свете.
The cream, colorless under normal lighting, glows bright red in ultraviolet light.

Cluster: [sk]
После четырех девяток играется безкозырная игра: по девять карт без козыря, остальные правила те же.
After four nines, the trumpless game starts: [every one gets] nine cards without a trump, the other rules [stay] the same.

Я скромная, умная, но безквартирная девушка из жуткой Сибирской дыры.
I am a modest, smart but apartmentless girl from a frightful Siberian hole.

Сегодня в продаже имеется безкаркасная мягкая мебель.
Today for sale we have frameworkless soft furniture.

Машинку, за сутки ставшую почти родной, тут же продали, я осталась безколесная и вернулась на работу в сильном расстройстве.
The car, which became so dear during the day [I had it], was immediately sold, and I became wheelless and returned to work in great distress.

Больному выполняется безконтрастная компьютерная томограмма области желчного пузыря, т. е. изображение, идентичное обзорному рентгеновскому снимку.
The patient undergoes contrastless computer tomography of the gall-bladder area, i.e., an image equivalent to an overview x-ray image.

Облигация безкuponная – это облигация, по которой не выплачиваются проценты, но которая изначально продаётся по цене существенно ниже номинальной.
A couponless bond is a bond that does not pay interest but which is initially sold at a price significantly below its denomination.

У страны уникальная экономика – безкредитная.
The country has a unique economy – [a] creditless one.

Безконтрольная застройка уничтожит курорты Закарпатья.
Control-free construction will destroy the resorts of the Carpathian mountains.

Cluster: [ss]

Во время войны эта безстрашная старушка, представьте, была фронтовой разведчицей.
During the war, this fearless old lady, imagine, was a frontline scout.

Сплошная безсистемная говорильня о языке и жизни, с низким результатом.
It’s all systemless chatter about language and life, with low results.

В ключевом моменте была безсловная пауза продолжительностью почти минуту.
At the key moment, [there] was a wordless pause almost a minute long.
Cluster [st]:
Изнутри толкает сила неизвестная, безтелесная.
From the inside pushes a force, unknown, bodyless.

Какая, на ваш взгляд, самая безтолковая собака?
What, in your opinion, is the most useless dog?

Следовательно потребовалось чуть больше месяца, чтобы установить с точностью, что сделка была безтоварная, и директор «Посейдона» имел умыслен присвоить крупную сумму денег.
The investigator needed a little more than a month to determine precisely that the deal was commodityless and the director of “Poseidon” aimed to steal a large sum of money.

Решить проблему детского курения может только безтабачная среда, считают ведущие пульмонологи России.
To solve the problem of child smoking can only tobaccoless environment, think Russia’s leading lung specialists.

Cluster [sx]:
Хорошая, но безхарактерная оптика за сумасшедшие деньги.
Good but characterless optics for insane money.

Диета была вегетарианская, овощная, местами сыроедческая, почти безхлебная и безкашная.
The diet was vegetarian, sometimes based on raw food, almost breadless and kashiless.

Виктор, а мышь у Вас безхвостая? Если «да», то поможет только замена на хвостатую.
Victor, is your mouse tailless? If yes, then only replacement with a tailed one will help.

Cluster [sp]:
Безповторная выборка может быть заменена повторной, в результате чего получается следующее уравнение.
Replacementless drawing can be replaced by a replaceful one, resulting in the following equation.

Предлагается безпыльная алмазная резка проемов в стенах.
Offering dustless diamond cutting of holes in walls.

Компания заявляет, что на ее компьютерах возможна безпроблемная установка оригинальной операционной системы.
The company claims that its computers offer problemless installation of the operating system.
What is it I am writing?  205

У нардепов будет безпределная зарплата.
The peoples' representatives will have limitless wages.

Предлагается безплатная реклама на нашем сайте.
Offering payless advertisement on our site.

В продаже имеется безпамятная видеокамера с кодеком.
For sale we have a memoryless videocamera with a codec.

Acknowledgement

Many thanks to two anonymous reviewers for helpful comments on the manuscript, to Eric Pederson for suggesting that I check for spellchecker effects, and to Varvara Kapatsinskaya for administering the dictation task.

Bionote

The author received his M.A. in Linguistics at the University of New Mexico in 2005 and his PhD in Linguistics and Cognitive Science at Indiana University in 2009. He is currently an assistant professor at the Department of Linguistics at the University of Oregon. The author’s research interests lie in experimental and corpus linguistics. He is mainly concerned with describing and explaining which statistical patterns in linguistic data are tracked by language learners and what units and generalizations they acquire and use in processing.

E-mail: vkapatsi@uoregon.edu

Notes

1. Throughout this paper, I use slashes (e.g., /X/) for the underlying form, no marking or square brackets (e.g., [X]) for the surface phonetic/phonological form and single quotes (e.g., ‘X’) for the orthographic form. ‘y’ = ‘ы’, /ɨ/; ‘šč’ = ‘щ’, /ʃʃ/ or /ʃʃ/ depending on dialect; ‘ś’ = ‘ш’, /ʃ/; ‘ž’ = ‘ж’, /ʒ/.

2. Note that here I am restricting myself to rule-based computation vs. any kind of retrieval of complete words, not necessarily one-step full-form retrieval. As Taft (2004) shows, the presence of whole-word frequency effects and the absence of base frequency words are not definitive evidence against morpheme-by-morpheme lexical retrieval because of the possibility of morpheme co-occurrence effects. Nonetheless, the morpheme co-occurrence effects emerge only if whole words are retrieved (whether in one step or morpheme-by-morpheme) rather than composed via application of general, non-lexically-specific grammatical rules (Frisson and Sandra 2002: 552–553, Sandra and Fayol 2003). See Arnon and Snider (2010), Kapatsinski and Radicke (2009), and Tremblay (2009) for methods to distinguish between the two retrieval scenarios.

3. This is a particular concern for Optimality-Theoretic approaches to orthography because Optimality Theory (Prince and Smolensky 1993/2004) classically assumes that all
constraints are universal and innate. At least some of the orthographic constraints are clearly language-specific (e.g., Hall 2004) and all are unlikely to be innate. However, in practice, constraints that refer to specific morphemes and thus are unlikely to be innate also appear to be necessary to describe spoken-language morphophonology (e.g., Prince and Smolensky [1993/2004] propose Align-um-Left for Tagalog um-infixation, a constraint that would be either useless or detrimental for children acquiring any other language).

4. There is, of course, clear precedent for holding linguistic theory responsible for describing late-acquired and explicitly-acquired rules, e.g., the work on meter (e.g., Fabb and Halle 2008, Hayes 1983, Kiparsky 1975, 2005), language games (e.g., Laycock 1972, Treiman 1986), miniature artificial languages (e.g., Ettlinger 2009, Finley 2008, Kapatsinski 2010b, Schane et al. 1974/1975), and second language learning (e.g., White 1990), which would have to be considered misguided under this view.

5. See also Opitz and Friederici (2004) for evidence of Broca’s area specifically associated with rule-based processing, and Maess et al. (2001) for evidence that the rules of musical “syntax” are also processed in Broca’s area.

6. As pointed out by a reviewer, there is one possible lexical exception to the rule that voicing assimilation is not to be reflected in the spelling: svat ‘male matchmaker’ vs. svad’ba ‘marriage’ (Rozental’ and Golub 1998: 23). However, it is likely that svad’ba is not svat+b+a for speakers of modern Russian: forms derived from svad’ba, e.g., svadeb ‘of marriages’, svadebnij ‘marriage-ADJ’ retain [d] despite the absence of a voiced obstruent after it, whereas the verb svatat’ ‘to (try to) match’ has a [t] in the same environment. The derivational suffix -b is unproductive and occurs in only a handful of words, and the word svat is rare compared to svad’ba. In any case, even if this is a genuine exception, it would not make the spelling of prefix-final consonants irregular since it involves the spelling of a root-final consonant in a noun and therefore does not introduce competition into the spelling of adjectival and verbal prefixes.

7. For prefixes ending in other obstruents ([s], [d], [t], or [v]), it is the underlying, unassimilated form that is reflected in the orthography. The ‘-voiced’ specification in (2)–(3) holds at either the orthographic or underlying phonological level but not at the surface phonetic level (Kuz’mina [1981]; for other examples of orthographic rules that parallel phonological rules but do not match them exactly, see Badecker [1996], Geißfuß-Wolfgang [2002] and Nunn [1998]; see Neijt [2002] vs. Sproat [2002] for different positions on whether the context is orthographic or underlying-phonological). The prefix-final [z] can precede another prefix, [v], which is devoiced phonetically but not orthographically ([f] but ‘v’). In this case, the orthography treats [v] as voiced, thus the preceding [z] is devoiced phonologically but not orthographically ([s] but ‘z’), e.g., bez-v-kus-nyj ‘tasteless’ and bes-flor-nyj ‘fluorine-free’ are both pronounced with [sf] but are spelled differently.

8. While I include a limited number of examples of each type, this should not be taken as an indication that all classes of forms are equally common. In particular, there are relatively few stems that begin with voiced obstruents.

9. The adjectival suffix (n- vs. ov- vs. -enn vs. 0 vs. -al/n) and stress placement appear to default to -n and stem stress (this combination appears in 85% [311/368] of adjectives found in my Google sample), but are not completely predictable. Thus one could argue that the uninflected stem does need to be stored, at least for adjectives not formed using the default method, but the inflected forms need not be.

10. Note that I am not claiming that all such adjectives are always derived from prepositional phrases. It can well be argued that the adjective is often derived from the noun directly without the intermediate phrasal stage, since the corresponding prepositional phrases, while interpretable, are sometimes lower in token frequency than the adjectives and occasionally sound somewhat awkward combined with the same head noun.
What is it I am writing?

11. Russian also has the prefixes nad- and v(o)z-, which we do not consider because very few verbs with more than 10 errors were found for these prefixes (3 for nad-, 6 for voz-, 5 for vz-) due in part to very low error rates with these prefixes (which is expected given that they have no competing differently-spelled forms) and in part to low type frequencies of the prefixes.

12. For other prefixes, intentional misspellings often can be detected by the systematicity of misspelling, which occurs throughout the text, as well as the presence of misspellings that are not motivated by pronunciation or the existence of a related confusable morpheme, e.g., famously, красавец ‘krasavčeg’ for красавец ‘krasavčik’ (‘cutie’), or превед ‘preved’ for прият ‘privet’ (‘hi’); final obstruents are devoiced in spoken Russian but are instead orthographically voiced in bottom-dwellers’ slang, resulting in forms that are faithful to neither the surface form nor the underlying form (see also Glazkov [1999] for evidence that phonemes in Russian have default orthographic realizations, and that Russian writers are sensitive to this). There are also characteristic double-letter spellings that are indicative of bottom-dwellers’ slang, including ‘фф’ (‘ff’) for devoiced /v/, and ‘цц’ for ‘тс’ (/ts/) straddling the boundary of the reflexive verbal suffix -sja following [1].

13. The only erroneous form searched for was the form deviating from the correct form only in the consonant of interest.

14. Of course, this is especially true for the difference between zero errors and one error, since the logarithm of zero is negative infinity.

15. I say ‘somewhat arbitrary’ because 10 is the number of hits Google displays on a single page of search results. Frequency (i.e., number of pages found), is calculated by Google differently depending on whether the number of hits exceeds 10 and thus must be split into multiple pages. When the last page of retrieved results is reached, the reported number of pages found is (sometimes drastically) reduced. Given that it is impossible (not allowed by Google) to look at / click through more than 10000 hits per search term, it was decided to use the frequency as reported when one is not on the last page of search results, which meant excluding words for which the number of errors or correct productions found was 10 or lower.

16. Since the appropriateness of using the mean log error rate as a measure of central tendency may be controversial given that the distribution of error rates is skewed, I have also tried a number of other manually inputted error rates. The smaller the error rate, the stronger the negative correlation observed under the null hypothesis after deleting the words with few errors or few correct forms. The null hypothesis reproduces the observed correlation only if its predicted (average) error rate is below .0001, which is definitely not the case in the data. Fewer than 5% of words found on Google were excluded because of having too few erroneous hits compared to an average rate of 58% predicted by the null hypothesis of a constant error rate of .0001. The distribution of numbers of excluded words under the null hypothesis of a constant error rate equal to .0001 does not include the observed value (varying between 56% and 59% in the 100,000 samples from the null), hence the observed data have p < .00001 under this null hypothesis).

17. Note that the rank transformation, though the standard way to deal with non-normal distributions in non-parametric statistics, does not remove the skew in the word frequency distribution where many words have the same frequency (0) or (1) in the corpus, and thus the same rank. If words with frequencies below 2 are removed, which removes the skew following rank transformation, the correlations remain significant: for masculine singulars, r = -.52, t(123) = -6.83, p < .00001; for feminine singulars, r = -.47, t(88) = -5.02, p < .00001; for plurals, r = -.51, t(90) = -5.65, p < .00001).

18. Note that the axes are logarithmically scaled. The significance tests are based on correlations between rank-transformed variables. Exclusion of words with few errors mean that points cannot occur to the left of the line connecting lower left corners of the subset regions: (.1;1), (1;01), (10;001).
19. This is perhaps unsurprising given the high productivity of converting bez+NP phrases into adjectives.

20. This comparison excludes adjectives where masculine nominative and feminine instrumental forms are identical (N = 10).

21. Let us define $F_{\{m;p\}}$ as the frequency of observing the masculine or plural form spelled with a ‘z’ or ‘s’ or, if unspecified, either ‘z’ or ‘s’. The null hypothesis is that $F_{m}/F_{m} = F_{p}/F_{p}$ for each word, i.e., the error rates depend only on the frequency of the masculine form. Then, the difference in frequencies $F_{m}(F_{m} + F_{p})$ would be correlated not with $(F_{m}/F_{m}) / ((F_{m} + F_{p}) / (F_{m} + F_{p}))$ but with $(F_{m}/F_{m}) / ((F_{m} + F_{p}) / (F_{m} + F_{p})) = 1/2$. Thus, the correlation expected under the null hypothesis is zero.

22. This is perhaps unfair to the Google estimates since the relationship between Google frequency and error rate is better fit by a logarithmic function than a linear function on ranks ($R^2 = 46\%$ vs. 67\% for the masculine Google estimates). However, many (5–6) corpus frequencies are zero, thus the logarithmic function cannot be fit to them. To make the statistics comparable, rank correlations are used for both corpus and Google data.

23. A necessary caveat is that the semantically-related phrasal competitor may not be the only phrasal competitor for the adjective during retrieval. The present analyses do not rule out an effect of competition from the entire cohort of phrases that have the same beginning as the target adjective. To fully take into account cohort competitors, we would need to weight each competitor appropriately for degree of phonetic overlap with the target adjective and other bez-initial adjectives whose roots begin with a voiceless consonant. The weighting is not informed by the present data but could, perhaps, be developed empirically through the use of eyetracking in the visual world paradigm with auditory input triggering shifts of attention to orthographic representations of competitors as the adjective or prepositional phrase unfolds (see Tanenhaus [2007] for a review of the methodology and similar applications).

24. Although, given that Largy et al. (1996) do not test for an influence of verb frequency when noun frequency is controlled, it is possible that the correct form is produced by rule, while the error results from lexical retrieval winning over grammatical computation. This is particularly plausible for their data given that sentential context favoring the noun appears to have an effect (relative to a neutral context) while sentential context favoring the verb does not (Largy et al. 1996: 246–247, 2007: 432–435).

25. To some extent the fact that the workbooks test knowledge of the spellings of relatively common words may be due to sentences in exercises and tests being drawn from (classical) fiction. Prescriptively incorrect spellings of low-frequency/unfamiliar words and words that are characteristic of other genres, such as business documents, are thus likely to go undetected during school training and testing. The lack of error detection can potentially lead to conventionalization of exceptional spellings of low-frequency and professional terms. This is not to say that the practice of focusing training and testing on high-frequency words is unjustified (cf. Sandra and Fayol 2003): after all, it is most important that the student learns how to spell words s/he often uses and the words for which everyone has a prescriptive norm. Problems may arise when a student enters a new field, whose vocabulary is not yet familiar to him/her.

26. The frequency values from a different corpus can also be modeled as the values from the original corpus plus some random noise reflecting the imperfect nature of the correlation between the frequency values for the same words in the two corpora. This extra noise decreases the likelihood of reproducing the observed correlation due to artifacts of sampling if samples from the null use frequency estimates from an independent corpus: $(\text{errorRate}_{\text{Corpus1}} = \text{errorCount}_{\text{Corpus1}} / \text{frequency}_{\text{Corpus1}}$ but frequency$_{\text{Corpus2}} = a + b * \text{frequency}_{\text{Corpus1}}$ ± randomNumber). Thus it appears that using only Corpus1 in the simulations is justified as it is maximally favorable towards the null.
What is it I am writing?

References


What is it I am writing? 213


