EXPERIMENTAL STUDY OF SEVERAL CORE CONCEPTS OF THEORETICAL MORPHOLOGY (ON THE MATERIAL OF RUSSIAN): REGULARITY, SYNCRETISM, MARKEDNESS

Dissertation for the degree of
Doctor of Science in Philology and Linguistics HSE

Moscow 2018
Наталия Анатольевна Слюсарь

ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ РЯДА ОСНОВНЫХ ПОНЯТИЙ ТЕОРЕТИЧЕСКОЙ МОРФОЛОГИИ (НА МАТЕРИАЛЕ РУССКОГО ЯЗЫКА): РЕГУЛЯРНОСТЬ, СИНКРЕТИЗМ, МАРКИРОВАННОСТЬ

Специальность 10.02.19 «Теория языка»

Диссертация на соискание ученой степени доктора филологических наук НИУ ВШЭ

Москва 2018
The dissertation was prepared at the National Research University “Higher School of Economics.”

**Publications**

Seven publications were selected for the defense. The corresponding author is marked by an asterisk.


The results of the present study have also been presented in the following papers:


Conference presentations and grants

The main results and conclusions of the study have been presented in 2011–2018 in 65 oral and poster presentations at 46 international conferences, including:

- CUNY Conference on Human Sentence Processing (2014, 2015, 2016, 2017);
- International Morphological Processing Conference (2013, 2015, 2017);
- International Conference on the Mental Lexicon (2014, 2016);
- Annual Meeting of North East Linguistic Society (NELS) (2016);
- International Morphology Meeting (IMM) (2014, 2016, 2018);
- American International Morphology Meeting (AIMM) (2012);
- “Night Whites” Workshop on Experimental Studies of Speech and Language (2011, 2014, 2018);
- International Organization of Psychophysiology (IOP) World Congress (2016, 2018);
- International Computer Linguistics Conference “Dialogue” (2013, 2015, 2018);

The studies presented in section 2 were partially supported by the grant 16-18-00041 from the Russian Science Foundation. Section 3 includes studies partially supported by the grant 16-18-02071 from the Russian Science Foundation. The project discussed in section 4 was funded by the grant 14-04-12034 from the Russian Foundation for Humanities.
# Table of contents

1. Introduction ......................................................................................................................... 6
2. Production and processing of isolated word forms .............................................................. 8
3. Production and processing of word forms in a sentence .................................................... 11
4. Creating resources for experimental research ................................................................. 18
5. Conclusions .......................................................................................................................... 19

References .................................................................................................................................... 20

Appendix A. Paper “An ER-fMRI study of Russian inflectional morphology” .................... 25
Appendix B. Paper “Changes in functional connectivity within the fronto-temporal brain network induced by regular and irregular Russian verb production” ..................... 26
Appendix C. Paper “Forms and features: the role of syncretism in number agreement attraction” .................................................................................................................................................. 37
Appendix D. Paper “Gender agreement attraction in Russian: production and comprehension evidence” .................................................................................................................................................. 38
Appendix E. Paper “Gender, declension and stem-final consonants: an experimental study of gender agreement in Russian” .................................................................................................................................................. 59
Appendix F. Paper “Processing of a free word order language: The role of syntax and discourse context” .................................................................................................................................................. 73
Appendix G. Paper “StimulStat: a lexical database for Russian” ............................................. 74
1. Introduction

The papers collected in this dissertation are dedicated to the experimental study of inflectional morphology. In six papers, we aim to study the relevance of several core morphological concepts for production and processing, including morphological regularity, productivity, feature markedness, different types of syncretism. As a result, we also learn something new about these concepts, being able to tease apart different approaches to them. This determines the theoretical significance of the study. Thus, the larger goal of the dissertation is to show how theory and experiment can interact fruitfully.

The seventh paper presents a lexical database that contains more than 50 thousand Russian words with more than 1.7 million word forms characterized for more than 70 different parameters. It can be used to facilitate stimulus selection for experimental research, which determines the practical significance of the study. We relied on this database in our experimental research, so it was essential to achieve the goals set in the present study.

Experimental papers study production and processing of isolated word forms and word forms in a sentence (primarily on the example of number and gender agreement) and in a wider discourse context. They are presented in more detail in subsequent sections. As we show below, these papers are dedicated to several topics that are a matter of ongoing debate in the field and make substantial contributions to these discussions, which defines the relevance of the study.

It should also be noted that the pool of languages used in experimental morphology is not very wide so far. All experiments presented in this dissertation were conducted on the material of Russian. This gave us an opportunity to tease apart several relevant factors and to introduce new questions, which was crucial for our contribution to the debates addressed in this dissertation.

The present study uses behavioral and neuroimaging methods; they are presented in subsequent sections. The validity of the results is guaranteed by the use of rigorous experimental design including material selection and experimental procedures. Data analysis always relies on the appropriate statistical methods.

In three papers included in this dissertation, Natalia Slioussar is the only author, in three more papers, she is the first and/or the corresponding author. Her contribution to coauthored papers can be described as follows. In the papers describing fMRI experiments (Slioussar et al. 2014; Kireev et al. 2015), she was responsible for selecting linguistic stimuli, interpreting the
results from the linguistic point of view and to a large extent for writing the papers. In (Slioussar and Malko 2016), both Natalia Slioussar and Anton Malko, her former MA student at Saint-Petersburg State University, took part in conducting experiments, analyzing the data and preparing the draft of paper. However, Natalia Slioussar played the leading role in planning the study and interpreting the results, as well as in revising the paper for publication.

The work on the StimulStat database was done together with Svetlana Alexeeva, who is the first and the corresponding author of the paper (Alexeeva, Slioussar, and Chernova, 2017), and Daria Chernova. Svetlana Alexeeva played a major role in creating and developing the database, Daria Chernova solved many technical problems to make it functional. Natalia Slioussar was the head of the project responsible for planning and making linguistic decisions (and also the head of the grant from the Russian Foundation for Humanities awarded to the team to fund the project). She also did most work preparing the paper for publication.

The main results of the project can be summarized as follows:

1. Morphologically regular and irregular word forms are processed differently in the brain, and the notion of regularity is associated with the so-called defaultness of the inflectional class, rather than with its type frequency or productivity.

2. Different methods of functional magnetic resonance imaging (fMRI) data analysis reveal different aspects of linguistic processing.

3. Case syncretism, not only systematic, but also accidental, plays a major role in production and processing of agreement attraction errors and therefore in the computation of grammatical agreement in general.

4. Feature markedness manifests itself in production and processing of grammatical agreement. However, experimental data do not point to a single markedness hierarchy, and suggest that representational markedness is relevant for production, while frequency-based hierarchy is relevant for comprehension.

5. Agreement processing is influenced by the prototypicality of inflectional affixes.

6. Word form processing depends on their information structure status and their syntactic position. Experimental data demonstrate the interplay of these two factors: discourse context problems are processed slower for word forms in non-canonical (scrambled) positions.
(7) The lexical database *StimulStat* with a web interface created for Russian allows searching for lemmas and word forms using more than 70 phonologic, orthographic, grammatical and semantic parameters that were demonstrated to be relevant for experimental studies. It can be used in experimental and other linguistic research on Russian.

### 2. Production and processing of isolated word forms

**Papers selected for the defense:** (Slioussar et al. 2014; Kireev et al. 2015).

Inflectional morphology is at the center of an important debate in cognitive science, concerning the general principles according to which the mental lexicon is organized and the foundational properties of our cognitive architecture. This debate focuses on the distinction between morphologically regular and irregular word forms. The so-called “dual route” approach assumes that the former are generated and processed by symbolic rules, while the latter stored in the lexicon, from where they can be retrieved through associative memory mechanisms (e.g. Clahsen 1999; Marslen-Wilson and Tyler 1997; Orsolini and Marslen-Wilson 1997; Pinker 1991, 1999; Pinker and Prince 1988; Ullman 2004). According to the “single route” approach, our cognitive architecture does not rely on symbolic rules and all word forms are computed by a single integrated system (e.g. MacWhinney and Leinbach 1991; McClelland and Patterson 2002; Plunkett and Marchman 1993; Ragnasdóttir et al. 1999; Rumelhart and McClelland 1986).

Behavioral studies testing DR and SR approaches analyze a variety of languages, but neuroimaging studies rely primarily on English and German data (e.g. Beretta et al. 2003; de Diego-Balaguer et al. 2006; Desai et al. 2006; Indefrey et al. 1997; Jaeger et al. 1996; Joanisse and Seidenberg 2005; Marslen-Wilson and Tyler 1998; Münte et al. 1999; Newman et al. 2007; Oh et al. 2011; Sach et al. 2004; Sahin et al. 2006; Ullman et al. 1997). Inflectional morphology in morphologically richer languages like Finnish, Polish and Arabic was examined in a number of neuroimaging studies, but these studies did not focus on the regular vs. irregular distinction. This is why we decided to address the problem of regularity in an fMRI study of Russian, a language with rich and diverse morphology.

The Russian verb system is very complex, and there are several approaches to dividing verbs into inflectional classes. These classes differ in type frequency, and four or five of them (depending on the chosen classification) are productive. The most frequent productive AJ class was observed to behave as the default class in several previous behavioral studies (Gor and
Thus, there is no obvious division into regular and irregular verbs, and, if any distinctions between different classes are found, we may be able to explore the nature of regularity and to find out which properties (type frequency, productivity or defaultness) are relevant for production and processing.

In our first experiment, we decided to look at the two poles of the verb class system, comparing verbs from the most frequent and productive AJ class to verbs from small unproductive classes (we reasoned that if any differences between these two groups were found, we could compare them to other verbs in subsequent studies). Participants were asked to generate present tense forms from different visually presented real and nonce verbs and to pluralize real and nonce nouns. First we performed a subtractive analysis of the data (Slioussar et al. 2014), then a ROI – whole brain voxel-wise analysis of context dependent changes in functional connectivity (PPI analysis) (Kireev et al. 2015).

We demonstrated that morphological regularity and processing difficulty effects can be teased apart. Activity of the left inferior frontal gyrus (LIFG) was greater for production of irregular verbs (compared to regular ones) and real verbs (compared to nonce ones), so this pattern was explained by processing difficulty (Slioussar et al. 2014). But the functional connectivity of the LIFG with temporal lobe was relatively increased when regular verbs were produced (compared to irregular ones). Nothing similar was found for the real/nonce distinction, so this was concluded to be a genuine regularity effect (Kireev et al. 2015).

Let us discuss the PPI results in more detail. Firstly, we found that functional connectivity between the LIFG and bilaterally distributed clusters in the superior temporal gyri was significantly greater in regular real verb trials than in irregular ones. No other comparisons gave significant results. Secondly, we observed a significant positive covariance between the number of mistakes in irregular real verb trials and the increase in functional connectivity between LIFG and the right anterior cingulate cortex in these trails as compared to regular ones. Thus, regularity and processing difficulty effects could be dissociated not only using different methods of analysis, but also by the PPI analysis alone.

We found only one previous published PPI study of inflectional morphology (Stamatakis et al. 2005). In this study, functional connectivity between functionally predefined ROIs was assessed during the same/different judgment task. Stimuli were orally presented pairs of English words and nonce words (e.g. jumped – jump, thought – think, jade – jay). Thus, the method and materials were very different from ours. Our first finding was similar to what
Stamatakis et al. reported, which shows that the observed regularity effect is very robust, being valid crosslinguistically both for production and comprehension.

As for the second finding, Stamatakis et al. have similar results going in the opposite direction. This is also true for the subtractive analysis of their data reported in (Tyler et al. 2005). We hypothesized that the processing difficulty went in the opposite directions in the two studies. Tyler et al. and Stamatakis et al. looked at stimulus pairs like *stayed* – *stay* vs. *taught* – *teach*. In regular pairs, the first stimulus was morphologically complex and the second was not, while in irregular pairs, both stimuli were morphologically simple. Thus, regular verb trials induced more processing load. In our study, all verb forms participants read or produced were morphologically complex. But irregular verbs involved various alternations in the stems etc., so irregular verb trials induced more processing load.

The two verb groups used in the first experiment differed by many properties (type frequency, productivity, defaultness), so it was impossible to determine which one was responsible for the observed regularity effect. Therefore we conducted a second fMRI study adding a third verb class — the I class (very frequent, but less frequent than the AJ class, productive, but not default). The results have already been presented at several conferences, but have not been published yet.

In this experiment, we also wanted to find out whether the observed effect would be replicated in comprehension. However, we were worried that in case of isolated word forms, processing of inflectional morphology would be shallow because it cannot be meaningfully interpreted or used for syntactic parsing. The only published fMRI study where regular vs. irregular verbs were compared in comprehension was discussed above (Stamatakis et al. 2005; Tyler et al. 2005), and it is notable that an ingenious design was used to draw participants’ attention to the morphological features of stimuli.

This design could not be replicated in Russian, so we came up with the following one. In every trial, participants first saw a real or nonce verb in the infinitive form and a pronoun *ja* ‘I’ or *on* ‘he’ below it (600 ms). After an interval, two present tense forms of the previously shown verb appeared on the left and on the right of the screen (1500 ms). One of them agreed with the pronoun, the other did not. Participants were asked to select the correct form. We preferred this design to showing only one form (agreeing or not) because this task would involve agreement violations and would focus participants’ attention on error detection.
We analyzed BOLD signal changes associated both with the 1st stimulus (an infinitive and a pronoun) and the 2nd stimulus (two present tense forms). Based on these analyses we selected ROIs for the PPI analysis. All non-trivial results were associated with the 2nd stimulus. We showed that the effects of interest are the same in production and in comprehension. The subtractive analysis showed that the activity of the LIFG gradually increased from the AJ class to I class and then to irregular verbs. The effect was analogous to the processing difficulty effect from (Slioussar et al. 2014).

The PPI analysis revealed a connectivity pattern that was very similar to the ones reported in (Kireev et al. 2015) and (Stamatakis et al. 2005). The fact that it was found in the two languages with relatively poor and relatively rich inflectional morphology and in the studies using three different tasks proves that this effect is reliable. Moreover, we found out whether it can be associated with type frequency, productivity or defaultness. The latter was true: an increase in functional connectivity of the LIFG was observed for the AJ class (as opposed to the I class and irregular verbs). Notably, this can be explained only in the dual route approach to inflectional morphology postulating a categorical distinction between the default class and the other classes. In the other approaches, regularity effects, if present at all, are expected to correlate with type frequency and productivity.

3. Production and processing of word forms in a sentence

3.1. Number agreement attraction

Paper selected for the defense: (Slioussar 2018a).

Three papers presented in this section are dedicated to production and processing of predicative agreement. For this domain, the phenomenon of agreement attraction plays an important role. An example of an attraction error is given (1a): the verb agrees not with the head of the subject NP but with a dependent NP (an attractor).

(1) a. *The key to the cabinets were rusty.
   b. *The key to the cabinet were rusty.

Across languages, such errors in number agreement have been shown to arise more frequently than errors of the type exhibited in (1b), where no attraction is possible (e.g. Bock and Miller 1991; Eberhard et al. 2005; Franck et al. 2002, 2006; Hartsuiker et al. 2003; Solomon and Pearlmutter 2004; Staub 2009, 2010; Vigliocco et al. 1995, 1996). In comprehension experiments, attraction errors have been demonstrated to trigger more grammaticality judgment
mistakes and to provoke less pronounced effects in reading time and EEG studies than other agreement errors (e.g. Clifton et al. 1999; Dillon et al. 2013; Pearlmutter et al. 1999; Tanner et al. 2014; Wagers et al. 2009).

Two major approaches to agreement attraction can be identified in the literature: representational and retrieval approaches. According to the representational approach (e.g. Brehm and Bock 2013; Eberhard et al. 2005; Franck et al. 2002; Nicol et al. 1997; Staub 2010), agreement attraction takes place because the mental representation of the subject NP’s number feature is faulty or ambiguous. Some authors assume that the number feature can “percolate” from the embedded NP to the subject NP. Others, relying primarily on the Marking and Morphing model (Eberhard et al. 2005), argue that the number value of the subject NP is a continuum. The more plural the subject NP, the higher the possibility of choosing a plural verb. This plurality depends on properties of the subject NP as a whole and of its head, such as collectivity, distributivity etc. The retrieval approach (e.g. Badecker and Kuminiak 2007; Dillon et al. 2013; Solomon and Pearlmutter 2004; Wagers et al. 2009) claims that the number feature on the subject NP is always represented unambiguously and correctly and attraction errors arise when the subject NP is accessed via cue-based retrieval to determine the number of the agreeing verb because several nouns are simultaneously active.

Experiments on agreement attraction have been useful to study many important questions: how agreement is produced and processed, whether this morphosyntactic operation can be influenced by semantics and phonology etc. Here, we will focus on the ones directly relevant for our studies. Firstly, in all examined languages, significant attraction effects were found only in the sentences with singular heads and plural attractors, but not in the opposite configuration. We will come back to this asymmetry in section 2.2.

Secondly, data from languages with morphologically marked cases show that syncretism is an important factor. For example, in an experiment on German (Hartsuiker et al. 2003), subjects like (2a), in which the form of the attractor is ambiguous between accusative and nominative, provoked significantly more errors than subjects like (2b), in which the attractor is unambiguously dative.

(2) a. die Stellungnahme gegen die Demonstrationen
    theF,NOM,SG position against theACC,PL demonstratons

    b. die Stellungnahme zu den Demonstrationen
    theF,NOM,SG position on theDAT,PL demonstratons
Notably, the syncretism factor has been examined only in production and never been examined separately from number: in (2a), the attractor is plural and morphologically ambiguous with nominative plural, and in (2b), it is plural and not ambiguous. This is possible to do in Russian, which has forms like večerinki ‘party<sub>GEN.SG/NOM.PL/ACC.PL</sub>’. In one production and two comprehension experiments (a speeded grammaticality judgement task and a self-paced reading task), we compared number agreement attraction effects with eight types of subject NPs exemplified in (3)–(4).

(3) a. trassa / trassy ėerez pole
    highway<sub>NOM.SG/NOM.PL</sub> across field<sub>ACC.SG(=NOM.SG)</sub> — so-called systematic syncretism
    b. trassa / trassy ėerez polja
    highway<sub>NOM.SG/NOM.PL</sub> across field<sub>ACC.PL(=NOM.PL)</sub> — so-called systematic syncretism

(4) a. komnata / komnaty dlja večerinki
    room<sub>NOM.SG/NOM.PL</sub> for party<sub>GEN.SG(=NOM.PL)</sub> — so-called accidental syncretism
    b. komnata / komnaty dlja večerinok
    room<sub>NOM.SG/NOM.PL</sub> for party<sub>GEN.PL</sub>

No effects were found with plural heads, as it was expected based on the previous findings. With singular heads, both in production and in comprehension syncretic genitive singular forms were found to trigger larger attraction effects than morphologically unambiguous genitive plural forms. Accusative plural forms coinciding with the nominative plural were shown to be the most effective attractors.

These results have implications for different models of attraction and for other discussions in morphology concerning ambiguity processing, different approaches to syncretism and the problem of lexical insertion. In particular, they show that syncretism does not simply boost the effect of the plural feature — it is an independent factor. The paper argues that the results cannot be explained as local coherence phenomena and suggests an explanation in terms of the cue-based retrieval approach to agreement attraction, with two notable modifications. First, the study shows that retrieval cues can tap into alternative feature sets of syncretic forms. Notably, unlike in the previously discussed cases, this is possible although no parallel syntactic parses are considered and no reanalysis is attempted. Second, it demonstrates that retrieval should rely on compound cues such as {nominative+plural} instead of a set of individual cues.

3.2. Gender agreement attraction

Paper selected for the defense: (Slioussar, Malko 2016).
As we noted in the previous section, in all studied languages, attraction effects were found to be asymmetric. They can be observed when the head is singular, and the attractor is plural, but are much weaker or virtually non-existent in the opposite configuration. In the majority of agreement attraction studies, this asymmetry is explained in terms of feature markedness. Plural is assumed to be the marked value of number feature, and the asymmetry is attributed to the fact that attractors with a marked feature are more disruptive.

However, many approaches to markedness have been proposed in the literature: the marked value of a certain feature can be determined based on frequency, presence of a non-zero affix, default use of a form (e.g., in impersonal sentences), various semantic tests etc. An overview can be found in (Haspelmath 2006). It is impossible to evaluate which approach is psycholinguistically relevant looking only at singular vs. plural — other features should be studied. We conducted one production and three self-paced reading experiments on gender agreement attraction in Russian (Sloussar, Malko 2016). There were very few production experiments (Badecker and Kuminiak 2007; Franck et al. 2008; Martin et al. 2014; Vigliocco, Franck, 1999) and no previous comprehension studies looking at gender.

We compared sentences with subject NPs with different gender combinations, as in (5a-b). Badecker and Kuminiak’s (2007) study on Slovak was especially relevant for us. Both Slovak and Russian have three genders. Masculine is the most frequent, while neuter is used as grammatical default (in impersonal sentences). Badecker and Kuminiak found that in production, feminine behaves as more marked than masculine and neuter, and masculine as more marked than neuter. Our production experiments observed the same pattern in Russian.

(5) a. okno v pole / vo dvor
    windowN to fieldN to yardM

b. vyxod v les / v pole
    exitM to forestM to fieldN

Number agreement attraction studies found parallel patterns in production and comprehension, but our comprehension experiments revealed a different pattern: masculine, the most frequent gender, triggered weakest effects. Thus, our study shows that different approaches to markedness are psycholinguistically relevant. There was another important finding: in the comprehension experiments attraction was observed for all dependent noun genders, but only for a subset of head noun genders. This goes against the traditional assumption that the features of the dependent noun are crucial for attraction, showing the features of the head are more
important. We demonstrated that this approach could be extended to previous findings on attraction and that there existed other evidence for it.

3.3. Prototypicality of inflections in agreement processing

Paper selected for the defense: (Slioussar 2018b).

The experiment presented in this section (Slioussar 2018b) is also dedicated to subject-predicate gender agreement processing, but studies it without attraction, focusing on two factors: the gender of the noun and its inflectional class, or declension. Russian nouns are inflected for case and number, and, depending on the set of their inflections, are divided into several declensions.

We will rely on the most widely accepted system presented in Table 1. As Table 1 shows, most consonant-final nominative singular forms (having zero inflection) are masculine, and most feminine nominative singular forms end in -a/ja, with 3rd declension consonant-final feminine nouns (3D-F) being less usual, or non-typical.

![Table 1](image)

Many experimental studies found differences between nouns with more and less typical inflections in a variety of languages (e.g. Andonova et al. 2004; Bates et al. 1995; Gollan and Frost 2001; Spalek et al. 2008). However, these studies usually looked at the processing of isolated nouns. Among the few sentence-processing studies, Caffarra et al. (2015) looked at Italian nouns with more and less typical endings presented in the same sentences. They were preceded by articles, which carried gender information. Nouns from the two groups elicited different ERP responses. Franck et al. (2008) and Vigliocco and Zilli (1999) demonstrated for Italian, Spanish, and French that heads with regular inflections are more resistant to gender agreement attraction.

---

1 Slioussar and Samoilova (2015) identified the frequency of nouns with different grammatical characteristics in the grammatically disambiguated subcorpus of the Russian National Corpus (http://www.ruscorpora.ru).
Only one paper (Taraban and Kempe, 1999) addressed this problem in Russian, and no differences between 2nd declension masculine and 3rd declension feminine nouns were found for native speakers. The task involved selecting a correct verb form for different subject nouns. We decided to come back to this question using the self-paced reading method, which could allow us tracing earlier processing stages. Materials were sentence sets in six experimental conditions like in (6).

(6) Xalat / kurtka / šinel’ + był / byla potrepannym / potrepannoj + ot mnogoletnej noski.
robe2D-M / jacket1D-F / overcoat3D-F + wasM/F shabbyM/F + from years-long wear

Firstly, the influence of the two factors of interest could be detected only in the sentences with agreement errors, i.e. no gender or declension is intrinsically more difficult to process (at least, in the sentence context). Secondly, declension played a role at a very early stage and its effect was very short-lived, while the role of gender became visible later and its effect was more pronounced. Namely, agreement errors were noticed significantly later with non-typical 3D-F nouns than with typical 1D-F or 2D-M ones. But error-related delay on subsequent words was significantly more pronounced for M subjects than for F ones (both 1D and 3D).

The latter result suggests that predictions we make about predicate gender are stronger for M subjects. This finding agrees with Slioussar and Malko’s (2016) study presented in the previous section, as well as with experimental studies of adjectival agreement in Russian (Akhutina et al. 1999, 2001; Romanova and Gor 2017). The former result can be explained by the fact that 3D-F nouns essentially look like masculine. But an alternative explanation is also possible: all agreement errors (in M or in N) may be harder to detect after 3D-F nouns, i.e. their gender can be in general harder to retrieve. A follow-up experiment is conducted to tease apart these two scenarios. Another current experiment looks at non-typical masculine nouns like papa ‘dad’.

3.4. Processing different word orders

Paper selected for the defense: (Slioussar 2011).

The paper presented in this section (Slioussar 2011) studies how different word forms are processed depending on the word order in the sentence and on its information structure (IS), i.e. what information is given or new. Russian has flexible constituent order. Available orders are used to encode IS, which means that they have different context requirements: for example, an OVS sentence is felicitous when the object is given and the subject is new, but not in zero context.
However, most experimental studies of flexible constituent order in different languages have focused on its syntactic properties, ignoring the context factor. Canonical and noncanonical orders were compared in zero context, and the latter were found to be more difficult to process (e.g., Bader and Meng, 1999; Erdocia, Laka, Mestres-Misse, and Rodriguez-Fornells, 2009; Frazier and Flores d’Arcais, 1989; Hyönä and Hujanen, 1997; Miyamoto and Takahashi, 2002, 2004; Stojanović, 1999; Vasishth, 2002). This was explained by the increased syntactic complexity of noncanonical orders. But an alternative explanation is also possible: noncanonical orders are infelicitous in zero context, which violates their IS requirements.

Context-sensitive studies of scrambling are very rare (Bornkessel and Schlesewsky 2006; Bornkessel et al. 2003; Kaiser and Trueswell 2004).\(^2\) Kaiser and Trueswell (2004) compared Finnish SVO and OVS structures in appropriate and inappropriate contexts. In Finnish, as in most other free word order languages, the vast majority of narrative sentences have a ‘given-new’ constituent order. So appropriate contexts presupposed such an order in target constructions, while inappropriate ones presupposed a ‘new-given’ order — i.e., essentially violating the IS requirements of target sentences. Kaiser and Trueswell showed that noncanonical constructions in appropriate contexts were processed faster than those in inappropriate contexts, but were still slower than the canonical order.

In our study on Russian (Slioussar 2011), we used the same methodology, but decided to take longer constructions and more extensive context sentences (making two out of three constituents in the target sentence given). In the first self-paced reading experiment, we compared S V IO DO (canonical), DO S V IO and DO IO V S constructions in appropriate vs. inappropriate contexts. Examples of context sentences are given in (7a-b), examples of target sentences are given in (8a-b). (7a) was used as appropriate context for (8a) and as inappropriate context for (8b) and for another target sentence with the DO S V IO order.

(7) a. Na 8 marta Danja Kašin pozdravil Mašu Smolinu.
   on March 8 [D.K.].NOM congratulated [M.S.].ACC
   [M.S.].NOM congrated [D.K.].ACC
   b. Na 8 marta Maša Smolina polučila pljuševogo slona i konfety.
   on March 8 [M.S.].NOM received [plush elephant].ACC and [candies].ACC
   [candies].ACC and [plush elephant].ACC

(8) a. Danja Kašin podaril Maše Smolinoj pljuševogo slona. S V IO DO
   [D.K.].NOM gave [M.S.].DAT [plush elephant].ACC
   [M.S.].DAT [plush elephant].ACC
   b. Pljuševogo slona Maše Smolinoj podaril Danja Kašin. DO IO V S
   [plush elephant].ACC [M.S.].DAT gave [D.K.].NOM
   [plush elephant].ACC [D.K.].NOM gave [M.S.].DAT

\(^2\) Sekerina (2003) also compared Russian PP S V IO DO PP and PP DO S V IO PP constructions in isolation and in the same context.
The context factor was significant, while the syntax factor was not. The less pronounced context effect evidenced in (Kaiser and Trueswell 2004) might be due to the use of shorter target and context sentences. We also demonstrated that the slow-down starts at the first contextually inappropriate constituent, which shows that the information about context requirements is taken into account immediately, but that it develops faster on preverbal subjects and postverbal indirect objects (occupying their canonical syntactic positions) than on preverbal indirect objects (occupying a noncanonical position, or scrambled). In the second self-paced reading experiment, these findings were replicated for IO S V DO and IO DO V S orders. S V IO DO orders with a continuation were used to show that there is no additional effect of inappropriate context at the end of the sentence. Thus, processing of word forms in a sentence depends both on their IS status and syntactic position.

4. Creating resources for experimental research

Paper selected for the defense: (Alexeeva, Slioussar, and Chernova 2017).

Experimental studies identified a large list of word properties that play a role for speech production and comprehension. They include lemma and form frequency, word length, the number of syllables, stress pattern, whether the word has homonyms, homographs or orthographical neighbors, whether it has multiple senses etc. Of course, various grammatical characteristics (part of speech, inflectional paradigm etc.) are also important. For several languages, databases with search tools were designed to take these properties into account during stimulus selection (e.g. Balota et al. 2007; Colheart 1981; Davis 2005; Davis, Perea 2005; Duchon et al. 2013; Heister et al. 2011; New et al. 2004).

As for Russian, some required characteristics were not represented in electronic dictionaries and databases at all, while the others were scattered across frequency lists, grammatical or explanatory dictionaries or others resources. Therefore, we created StimulStat – a lexical database for Russian in the form of a web application (http://stimul.cognitivestudies.ru). The database contains more than 50,000 most frequent Russian words (1.7 million word forms). These words and forms are characterized according to more than 70 properties, including frequency, length, phonological and grammatical properties, orthographic neighbourhood frequency and size, homonymy and polysemy. Some properties were retrieved from various dictionaries and presented collectively in a searchable form for the first time, the others were computed specifically for the database. In the paper describing the database (Alexeeva,
Slioussar, and Chernova 2017) we also present some interesting crosslinguistic differences that can be identified using such databases.

5. Conclusions

Experiments collected in this dissertation study how word forms are produced and processed in isolation and inside sentences on the material of Russian. Many experiments were the first or among the first to look at a certain phenomenon (e.g. at the factors of syncretism and declension in agreement processing studies, at gender agreement attraction in comprehension) or to apply a particular method (e.g. the PPI method of fMRI data analysis in a linguistic study). The study shows that several notions discussed in theoretical morphology (including morphological regularity, markedness, syncretism) are psycholinguistically relevant and that experimental data can shed new light on them, demonstrating how theory and experiment can interact fruitfully.

Firstly, the studies discussing how word forms are represented and processed in the mental lexicon debate whether there is a categorical distinction between morphologically regular and irregular forms. This is a repercussion of a major controversy in cognitive sciences whether our cognitive architecture uses symbolic rules or not. In our fMRI experiments, we could tease apart processing difficulty and morphological regularity effects and went on to ask how morphological regularity should be defined: whether it is related to type frequency, productivity or defaultness of the inflectional class.

Secondly, we looked at agreement — one of the basic grammatical operations that governs using word forms in a sentence. Many psycholinguistic works aim to find out how subject–predicate agreement is produced and processed relying on so-called attraction errors in number and gender agreement. Most studies are dedicated to number agreement attraction, while four our experiments focused on gender, showing that this feature is interestingly different from number. Number agreement attraction was examined in three other experiments, which tested the role of different types of syncretism for this phenomenon. Both groups of experiments provide arguments for one of the alternative approaches to agreement attraction, the retrieval approach, and call for certain modifications in it. We also examined how the prototypicality of inflections (i.e. whether a particular inflection is more or less typical for a particular grammatical feature, e.g. for a particular gender) affects agreement processing. This factor was
repeatedly demonstrated to play a role for processing of isolated forms, but only a couple of recent studies showed that it might also be important on the sentence level.

Finally, we studied how different word forms are processed depending on the word order in the sentence. This question was addressed in many experimental studies, but the majority of them looked at sentences in isolation thus ignoring the context factor — in most languages that allow for word order permutations they are used to convey different information structures and thus are felicitous only in particular contexts. Our experiments took this factor into account and demonstrated how it affects online processing.

References


Appendix A.

Paper “An ER-fMRI study of Russian inflectional morphology”


Abstract

The generation of regular and irregular past tense verbs has long been a testing ground for different models of inflection in the mental lexicon. Behavioral studies examined a variety of languages, but neuroimaging studies rely almost exclusively on English and German data. In our fMRI experiment, participants inflected Russian verbs and nouns of different types and corresponding nonce stimuli. Irregular real and nonce verbs activated inferior frontal and inferior parietal regions more than regular verbs did, while no areas were more activated in the opposite comparison. We explain this activation pattern by increasing processing load: a parametric contrast revealed that these regions are also more activated for nonce stimuli compared to real stimuli. A very similar pattern is found for nouns. Unlike most previously obtained results, our findings are more readily compatible with the single-system approach to inflection, which does not postulate a categorical difference between regular and irregular forms.

The link on the publisher’s website

This is not an open access paper, so according to the publisher’s policy it cannot be reproduced. It can be accessed on the publisher’s website:
Appendix B.

Paper “Changes in functional connectivity within the fronto-temporal brain network induced by regular and irregular Russian verb production”

Changes in functional connectivity within the fronto-temporal brain network induced by regular and irregular Russian verb production

Maxim Kireev1,2, Natalia Slioussar3,2 *, Alexander D. Korotkov1,4, Tatiana V. Chemigovskaya1,2 and Svyatoslav V. Medvedev1

1 N.P. Bechtereva Institute of the Human Brain, Russian Academy of Sciences, St. Petersburg, Russia
2 Faculty of Liberal Arts and Sciences, St. Petersburg State University, St. Petersburg, Russia
3 Faculty of Philology, Higher School of Economics, Moscow, Russia
4 Radiological Center of Tyumen Regional Oncology Center, Tyumen, Russia

INTRODUCTION

Numerous studies examine morphologically complex forms to compare different models of inflection in the mental lexicon. One of the crucial things they focus on is the distinction between regular and irregular forms. The so-called “dual route” (DR) approach assumes that the former are generated and processed by symbolic rules, while the latter stored in the lexicon, from where they can be retrieved through associative memory mechanisms (e.g., Pinker and Prince, 1988; Pinker, 1991; Marslen-Wilson and Tyler, 1997; Orsolini and Marslen-Wilson, 1997; Clahsen, 1999; Ullman, 2004). According to the “single route” (SR) approach, all forms are computed by a single integrated system that contains no symbolic rules (e.g., Rumelhart and McClelland, 1986; MacWhinney and Leinbach, 1991; Plunkett and Marchman, 1993; Ragnasdóttir et al., 1999; McClelland and Patterson, 2002).

Behavioral studies testing DR and SR approaches analyze a variety of languages, but neuroimaging studies rely primarily on English and German data (e.g., Jaeger et al., 1996; Indefrey et al., 1997; Ullman et al., 1997; Marslen-Wilson and Tyler, 1998; Münte et al., 1999; Newman et al., 1999, 2007; Beretta et al., 2003; Sach et al., 2004; Joanisse and Seidenberg, 2005; Desai et al., 2006; Oh et al., 2011). Inflectional morphology in morphologically richer languages like Finnish, Polish, and Arabic was examined in a number of neuroimaging studies (e.g., Lehtonen et al., 2006; Boudelaa et al., 2010; Leminen et al., 2011; Szlachta et al., 2012). However, these studies did not compare regular and irregular forms, focusing on other problems (the distinction between inflectional and derivational morphology, the role of general perceptual and specifically linguistic complexity, etc.).

In the present study, we turned to Russian, a language with rich and diverse morphology, and conducted an fMRI investigation where participants were asked to generate present tense forms from different real and nonce (nonword) verbs and to
pluralize real and nonce nouns. Addressing the problem of regularity in a morphologically rich language is important because one can tease apart several factors that are confounded in a language like English (while English definitely has its own advantages with its minimalist system and sharp contrasts between inflectional classes). To give one example, all regular past tense forms are morphologically complex in English, i.e., contain a stem and a suffix (−ed), while irregular forms are morphologically simple. In Russian, all past tense forms are morphologically complex, which gives us an opportunity to find out whether the effects observed in English were due to regularity or to morphological complexity. Other properties of Russian that may be relevant for the debate will be pointed out in Section “A Brief Description of the Russian Verb and Noun Systems.” We opted for a production task because it was used in the majority of neuroimaging studies focusing on regular vs. irregular inflectional morphology.

Experimental data reflecting the localization and the direction of the change in functional activity are reported in Slioussar et al. (2014). In this paper, we present a ROI-whole brain voxel-wise analysis of context dependent changes in functional connectivity [a psychophysiological interaction (PPI) analysis; Gitelman et al., 2003]. The first type of analysis makes it possible to reveal functionally segregated brain areas that change their activity in response to experimental manipulations, while PPI is a measure of functional connectivity, which provides complementary information showing how these segregated brain areas are integrated (Friston, 2011). Although PPI analysis does not make it possible to infer causal relationships, it gives an opportunity to observe how the functional interplay between involved brain regions is changed as a function of the psychological context.

Therefore, we saw PPI analysis as a valuable tool to approach the problem from a new angle, especially given the fact that we found only one previous PPI study of inflectional morphology (Stamatakis et al., 2005). Important similarities and differences between Stamatakis et al.’s (2005) findings and our results offer a novel perspective on our findings from Slioussar et al. (2014), the account proposed by Stamatakis et al. (2005) and a number of problems discussed in other studies.

A BRIEF DESCRIPTION OF THE RUSSIAN VERB AND NOUN SYSTEMS

The Russian verb system is very complex, and there are several approaches to dividing verbs into classes. According to the one developed in Jakobson (1948), Townsend (1975) and Davidson et al. (1996), Russian has 11 verb classes and several so-called anomalous verbs. Ten classes are identified by their suffixes, while the 11th class has a zero suffix, and is subdivided into subclasses depending on the quality of the root-final consonant [Jakobson (1948) and Townsend (1975) counted them as 13 separate classes].

All verbs have two stems: the present/future tense stem and the past tense stem. Depending on the class, the correlation between them may include truncations or additions of the final consonant or vowel, stress shifts, suffix alternations, alternations of stem vowels, and stem-final consonants. The verb class also determines which set of endings is used in the present and future tense (first and second conjugation types). Usually, the class is unrecoverable from a particular form. For example, деля́ть ‘to do’ belongs to the AJ class, and its third person plural present tense form is деля́ть-ут (-ут is added, first conjugation type).1 Писа́ть ‘to write’ belongs to the A class, and its third person plural present tense form is писа́ть (-а- suffix is truncated, first conjugation type, final consonant alternation, stress shift). Дерза́ть ‘to hold’ belongs to the ZHA class, and its third person plural present tense form is дерза́ть (-а- suffix is truncated, second conjugation type).

Verb classes dramatically differ in frequency, and five of them are productive. Thus, there is no single productive pattern that can be applied to any stem irrespective of its phonological characteristics, and no obvious division into regular verbs (RVs) and irregular verbs (IVs) in this system. In our fMRI experiment, we decided to look at the two poles of this system, comparing verbs from the most frequent and productive AJ class to verbs from small unproductive classes (we reasoned that if any differences between these two groups were found, we could compare them to other verbs in subsequent studies). For the sake of brevity, we will further call these groups regular and irregular.

Russian nouns are inflected for number and case and are classified into different declensions depending on the set of their number and case endings. In many ways, this system is simpler than the system of verb classes. There are only three declensions (plus a group of nouns with adjectival endings, several exceptional cases and a number of uninflected nouns). These declensions differ in frequency, but all three are productive. Usually, the declension can be unambiguously determined from the nominative singular form. Inside every declension there are small groups of nouns with minor irregularities: unusual endings in some forms or stem alternations. For our study we selected a group of nouns that lose the last vowel of the stem in many forms including the nominative plural form (e.g. koster ‘fire’ – kostry) and a group where the stem never changes, as in the majority of Russian nouns (e.g., sofer ‘driver’ – sofrey). We will further call the first group irregular, although this is a relatively minor irregularity.

PREVIOUS STUDIES TESTING THE SR AND DR APPROACHES ON RUSSIAN

Behavioral studies testing SR and DR approaches on Russian looked at adult native speakers, L1 and L2 learners and subjects with various neurological and developmental deficits (e.g., Gor and Chernigovskaya, 2001, 2003, 2005; Gor, 2003, 2010; Chernigovskaya et al., 2007; Svistunova, 2008; Gor et al., 2009; Gor and Jackson, 2013). Participants were provided with infinitives or past tense forms of real or nonce verbs and prompted to generate first person singular and third person plural present tense forms. The findings did not unambiguously support either DR or SR approach. For example, on one hand, adults were shown to use the most frequent AJ class pattern as the default one, although Russian has several highly frequent productive verb classes. In particular, it was often applied to nonce verbs irrespective of their morphological properties. On the other hand, children consecutively overgeneralize several conjugaional patterns in the

---

1There are several ways to transliterate Russian words from Cyrillic to Latin alphabet. In this paper, we use the so-called scholarly transliteration system.
course of acquisition. As a result, the group of authors working on Russian argued that Yang’s (2002) model relying on multiple rules of different status might be better suited to account for their findings. A similar model for Russian was developed by Gor (2003).

Subtractive analysis of the data from the present experiment we reported in Slioussar et al. (2014) is the only fMRI study of Russian inflectional morphology we are aware of. Previous neuroimaging studies arguing for the DR approach, as well as a number of studies that do not directly address the DR vs. SR debate (e.g., Marslen-Wilson and Tyler, 2007; Bozic et al., 2010, 2013; Szlachta et al., 2012), argue that rule-based processing is supported by the fronto-parietal network, particularly by Broca’s area. However, only two fMRI studies comparing regular vs. irregular form production found more activation in Broca’s area for regulars (Dhond et al., 2003; Oh et al., 2011). Increased left IFG activation for regulars was also observed in an fMRI study where the processing of spoken regular and irregular forms was compared in a same-different judgment task (Tyler et al., 2005).

Other fMRI studies report the opposite pattern: Broca’s area was activated more by irregulars (Beretta et al., 2003; de Diego-Balaguer et al., 2006; Desai et al., 2006; Sahin et al., 2006). Two alternative explanations are proposed. Proponents of the DR approach suggest that these results can be explained by conflict monitoring between the regular rule and irregular form or by inhibition of regular rule application (e.g., Sahin et al., 2006). Desai et al. (2006) argue for the SR approach: they conclude that the observed activation differences reflect the greater processing load posed by irregulars, which rely on less frequent inflection patterns than RVs and therefore have greater attentional and response selection demands.

In Slioussar et al. (2014), nonce verbs and nouns were added to the comparison. Participants silently read stimuli and produced aloud particular forms from them. We found that functional activity within the fronto-parietal network was influenced by regularity and lexicality: it was greater for IVs than for regular ones and for nonce verbs than for real ones. We demonstrated that the effects of regularity and lexicality were very similar and concluded that the observed BOLD changes were induced not by (ir)regularity as such, but by the increase of processing load from RV to irregular (IV) to regular nonce verb (RNV) to irregular nonce verbs (INV).

This conclusion was supported by the (RV > B) < (IV > B) < (RNV > B) < (INV > B) parametric contrast, where B is an implicitly modeled baseline, and by behavioral results: the number of mistakes increased from RV to IV to RNV to INV condition. The results for nouns were similar. Only the main effect of regularity did not reach significance in the factorial analysis of fMRI data – presumably, because the only irregular feature we could find for our noun stimuli was rather minor (see A Brief Description of the Russian Verb and Noun Systems).

A PREVIOUS PPI STUDY OF INFLECTIONAL MORPHOLOGY AND THE PRESENT STUDY

We were only able to find only one PPI study of inflectional morphology (Stamatakis et al., 2005). In this study, functional connectivity between functionally predefined regions of interest (ROIs) located in the left inferior frontal gyrus (LIFG), anterior cingulate cortex (ACC), superior temporal gyrus (STG), and middle temporal gyrus (MTG) was assessed during the same/different judgment task. Stimuli were aurally presented pairs of English words and nonce words, in particular, RV and IV pairs like jumped – jump and thought – think.

Stamatakis et al. (2005) report a positive influence of LIFG activity on the activity in the left STG/MTG and a modulatory influence of ACC activity on this fronto-temporal connectivity. The former effect did not depend on regularity per se, but we know from the subtractive analysis of the data reported in Tyler et al. (2005) that RVs activated the LIFG, bilateral STG and MTG significantly more than irregular ones in this study. The latter effect was significantly stronger for regulars than for irregulars. Stamatakis et al. (2005) believe that these findings indicate greater engagement of the fronto-temporal network in RV processing, with the ACC playing a monitoring role. They conclude: “this reflects the additional processing demands posed by regular inflected forms, requiring modulation of temporal lobe lexical access processes by morphological parsing functions supported by the LIFG” (p. 116).

 Undertaking a PPI analysis of our data, we were primarily interested in two things. Firstly, an advantage of this approach is that task-dependent connectivity changes may be detected even when the levels of functional brain activity are not affected by experimental manipulations. We aimed to reveal functional interactions underlying changes in functional activity observed within the LIFG during regular and irregular form production (Slioussar et al., 2014). As we noted above, the increase in LIFG activity in IV trials was explained by the difference in processing load between these two tasks in Slioussar et al. (2014). In principle, this difference could attenuate functional activity changes associated with regularity. Therefore we turned to PPI analysis to find out whether this was indeed the case and to tease apart connectivity changes associated with morphological properties and with cognitive demands.

Secondly, we were interested how our findings would compare to Stamatakis et al.’s (2005) given several important differences in our experiments. First of all, there are obvious differences in the experimental task and in the language used (morphologically poor English vs. morphologically rich Russian). Furthermore, subtractive analyses presented in Tyler et al. (2005) and Slioussar et al. (2014) revealed the opposite results, in particular, the LIFG was more activated by regulars in the first study and by irregulars in the second. Finally, the analyses of behavioral data (the number of mistakes in different conditions) showed that irregular trials were more difficult than regular ones for the participants of our study, while Tyler et al. (2005) reported very similar accuracy rates.

In general, we wanted to see whether the functional connectivity of LIFG would be substantially different during comprehension and production of regular vs. irregular forms (although our task definitely involves a silent reading stage as well). In particular, we expected that if the findings from Stamatakis et al. (2005) are genuine regularity effects, we might be able to replicate them despite all the differences, teasing them apart from processing difficulty effects identified in Slioussar et al. (2014). Foreshadowing the results, this is exactly what we did in the present study.
MATERIALS AND METHODS

PARTICIPANTS
Twenty-one healthy subjects participated in the study (13 females, 8 males). All participants were native speakers of Russian, 19–32 years of age, with no history of neurological or psychological disorders. All participants were right-handed, as assessed by the Edinburgh Handedness Inventory (Oldfield, 1971). Subjects were given no information about the specific purpose of the study. All subjects gave their written informed consent prior to the study and were paid for their participation. All procedures were in accordance with the Declaration of Helsinki and were approved by the Ethics Committee of the N.P. Bechtereva Institute of the Human Brain, Russian Academy of Sciences.

MATERIALS
Materials consisted of eight groups of real and nonce verbs and nouns, illustrated in Table 1 (a complete list is given in Supplementary Material). The first group of 35 real verbs belonged to the AJ class (RV); the second group contained 35 verbs from several small non-productive classes (IV). Only unprefixled imperfective verbs were used. Two matching groups of 35 nonce verbs (RNVs and INVs) mimicked the general characteristics of the corresponding real verb groups (length and phonological properties of the stem).

The first group of 35 real nouns had no stem changes (regular nouns, RN), while in the second group the last vowel of the stem was dropped in many forms including the nominative plural form (irregular nouns, IN): e.g., safer ‘driver’ – sovery vs. koster ‘fire’ – kostry. All nouns were masculine, belonged to the first declension and had the nominative plural form ending in -y. Two groups of 35 nonce nouns (regular nonce nouns, RNN, and irregular nonce nouns, INN) were created to match two real noun groups. Frequency was balanced for all real stimulus groups using The Frequency Dictionary of the Modern Russian Language (Lyashevskaya and Sharoff, 2009). Stimuli in all groups were matched for length (see Supplementary Material).

Vowels are dropped only in a subgroup of noun stems ending in particular vowel and consonant clusters (e.g., -er, -or, -el, -ol etc.). We selected stems with such clusters both for irregular and for RN groups so as not to make the former more phonologically homogenous than the latter. Final vowel dropping is usually predictable from the combination of consonants before this vowel and from the position of the stress. However, since stimuli were presented visually, no information about stress was available for nonce nouns, and different nominative plural forms could be licitly derived from them.

LANGUAGE PROTOCOL AND EXPERIMENTAL fMRI PARADIGM
In total, we had 280 stimuli. Each stimulus was visually presented for 700 ms. Fixation crosses (“xxxxx”) were displayed during inter-stimulus intervals, which varied between 3100 and 3500 ms with a 100 ms step. 140 “null-events” (fixation crosses) were pseudo-randomly intermixed with the stimuli (Friston et al., 1999). The experiment was divided into three consecutive runs with 2–5 min rest between them and was preceded by a short practice run. The first 10 dummy scans of each run were discarded. Stimulus delivery and synchronization with fMRI data acquisition were carried out via the Eloquence fMRI System (In vivo) and E-Prime software (version 1.1, Psychology Software Tools Inc., Pittsburgh, PA, USA).

Verbs were presented in the infinitive form, and nouns were presented in the nominative singular form. Subjects were instructed to generate aloud as fast as possible the first person singular present tense form if they saw a verb or the nominative plural form if they saw a noun. All responses were recorded simultaneously with fMRI data acquisition by means of the Persuio MRI Noise Cancellation System (Psychology Software Tools, Inc.). Their correctness was assessed offline. When a participant’s responses were no longer appropriate for the target’s category, the corresponding trials were discarded in the subsequent fMRI analyses.

MR IMAGING PROTOCOL
Magnetic resonance imaging was performed on a 3 Tesla Philips Achieva scanner. In addition to a scout sequence, participants underwent structural and functional imaging. Structural images were acquired applying a T1-weighted pulse sequence (TIW-3D-FFE; TR = 2.5 ms; TE = 3.1 ms; 30° flip angle) measuring 130 axial slices (field of view, FOV = 240 mm × 240 mm; 256 × 256 scan matrix) of 0.94 mm thickness. Functional images were obtained using an echo planar imaging (EPI) sequence (TE = 35 ms; 90° flip angle; FOV = 208 mm × 208 mm; 128 × 128 scan matrix). Thirty-two continuous 3.5 mm thick axial slices (voxel size = 3 mm × 3 mm × 3.5 mm) covering the entire cerebrum and most of the cerebellum were oriented with reference to the structural image. The images were acquired with a repetition time (TR) of 2000 ms. In order to avoid extensive head motions we used an MR-compatible soft cervical collar.

CONNECTIVITY ANALYSIS
fMRI data preprocessing included realignment, slice-time correction, spatial normalization, and 8 mm full-width/ half-maximum isotropic Gaussian smoothing (for details, see Slioussar et al., 2014). It was carried out using SPM8 software (Wellcome Department of Cognitive Neurology, London, UK). Artifact Detection Toolbox2 was used to remove fMRI outliers from the analysis. During the PPI analysis, ROIs were selected from the cluster in the LIFG, which exhibited greater BOLD values for the production of irregular forms (Slioussar et al., 2014). Three ROIs

Table 1 | Examples of stimuli in different conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Presented forms</th>
<th>Correct answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular verbs (RV)</td>
<td>kivat’ ‘to nod’</td>
<td>kivaju</td>
</tr>
<tr>
<td>Irregular verbs (IV)</td>
<td>kol’ ‘to cleave, to sing’</td>
<td>kolju</td>
</tr>
<tr>
<td>Regular nonce verbs (RNV)</td>
<td>vupat’</td>
<td>vupaju</td>
</tr>
<tr>
<td>Irregular nonce verbs (INV)</td>
<td>xorat’</td>
<td>xorju</td>
</tr>
<tr>
<td>Regular nouns (RN)</td>
<td>sokol ‘falcon’</td>
<td>sakoly</td>
</tr>
<tr>
<td>Irregular nouns (IN)</td>
<td>posol ‘ambassador’</td>
<td>posoly</td>
</tr>
<tr>
<td>Regular nonce nouns (RNN)</td>
<td>mokol’</td>
<td>mokoly (and mokly)</td>
</tr>
<tr>
<td>Irregular nonce nouns (INN)</td>
<td>fopat’</td>
<td>foply (and fopoly)</td>
</tr>
</tbody>
</table>
FIGURE 1 | Location of ROIs in the LIFG and results of PPI analysis. (A) Three ROIs overlaid on the areas Slioussar et al. (2014) identified as sensitive to the main effect of regularity (regular real and nonce verbs were compared with irregular real and nonce verbs). (B) Increase in functional connectivity induced by regular real verb production in the RV > IV comparison for the LIFGop3 seed region. (C) Covariance between the number of mistakes in irregular real verb production and functional connectivity induced by irregular real verbs in the IV > RV comparison for the LIFGop3 seed region.

were created by centering a 4 mm radius sphere in the corresponding local maxima in the opercular part of the LIFG (BA 44, see Figure 1A), as defined by the Anatomy toolbox 2.0 (Eickhoff et al., 2005). The analysis of functional connectivity changes was performed between each of the selected ROIs and the remaining voxels of the brain using the generalized PPI toolbox3 (McLaren et al., 2012) and included the following steps. First, neuronal activity underlying the observed BOLD changes in every ROI was mathematically estimated (Gitelman et al., 2003). Then the estimated neuronal activity was multiplied by the vectors of each condition’s ON times and convolved with the hemodynamic response function (McLaren et al., 2012; Cisler et al., 2013).

As a result, PPI-regressors corresponding to every experimental trial were created, and the PPI analysis was performed using the general linear model (GLM). Additionally, the GLM included the following nuisance variables: (1) regressors modeling the BOLD signal changes induced by eight experimental conditions and mistake trials (as in the conventional subtractive GLM analysis); (2) head motion parameters and the global mean fMRI outliers; (3) a regressor reflecting the time series of BOLD signal changes within the ROI to exclude context-dependent changes occurring at the hemodynamic level.

To be able to compare our results to those of Stamatakis et al. (2005), we focused on the contrast between regular and irregular real word trials in the connectivity analysis, as these authors did. The fronto-temporal connectivity observed by Stamatakis et al. (2005) is most reasonably described as a frontal modulation of lexical access processes, which is obviously not applicable to nonce stimuli. However, the findings from other comparisons are also reported. As in Slioussar et al. (2014), we analyzed verbs and nouns separately rather than putting them together and treating word category as the third factor, primarily because the type of irregularity we were able to find for nouns was very minor compared to what we had in the case of verbs. Thus, RV > IV and IV > RV contrasts of PPI-parameters were estimated with the use of one-sample t-tests. Additionally, PPI-parameters for all real and nonce verb trials were analyzed using the ANOVA with two repeated measure factors: lexicality (real vs. nonce) and regularity. The same was done for nouns.

Statistical parametric mappings were computed using the $p < 0.001$ voxel-wise uncorrected threshold. To avoid false positive findings, the FWE $p < 0.05$ correction for multiple comparisons was applied at the cluster level. Since two t-tests were calculated...

3http://www.nitrc.org/projects/gppi
for each of the three ROIs, an additional Bonferroni–Holm correction for multiple comparisons was used. The anatomical location of the functional connectivity changes revealed was identified by the Anatomy toolbox. The REX toolbox\(^4\) was used to demonstrate differences between beta values reflecting functional connectivity changes in the revealed clusters.

**RESULTS**

In the RV > IV comparisons, the PPI analysis revealed clusters bilaterally located in the anterior portion of the superior temporal gyri (STG, see Table 2; Figure 1B). This effect was observed only for the LIFGop3 ROI seed, RV > IV PPI-contrasts for the other two ROI seeds were not significant. Calculating the mean values of PPI-parameters within the obtained clusters pointed to a relative increase in connectivity in RV trials in comparison to IV trials.

In the IV > RV comparisons, no significant changes in functional connectivity were found for all selected ROI seeds. Since we had concluded in Slioussar et al. (2014) that IV trials were characterized by an increase in processing load in comparison to RV trials, and this conclusion relied not only on neuroimaging, but also on behavioral data (number of mistakes in different conditions), we undertook the following subsidiary analysis to reveal processing load effects. We took the number of mistakes committed by every participant in the IV trials as an individual measure of task difficulty. As we reported in Slioussar et al. (2014), participants made significantly more mistakes in the IV condition than in the RV condition (96 out of 735 vs. 22 out of 735 responses in total, or 13.1% vs. 3.0%, respectively). If we look at each participant separately, the number of mistakes in the IV trials varies from 1 out of 35, or 2.9%, to 9 out of 35, or 25.7%.

Then we submitted IV > RV contrasts of PPI parameters calculated for every participant to the second level group analysis. A one-sample \(^-\)test, as it is implemented in SPM8, was used with the percentage of mistakes committed by every participant as a variable of interest and estimates of individual IV > RV PPI contrasts as a dependent variable. The results were significant only for one ROI seed, LIFGop3, the same as in the RV > IV PPI analysis above. For this ROI seed, we observed a significant positive covariance between the number of mistakes in IV trials and the difference in functional connectivity between the LIFG and the right ACC (BA 32; see Table 3; Figure 1C). Notably, when error rates are low, the difference is negative, i.e., connectivity between the LIFG and the ACC is greater in RV trials. As error rates grow, the difference approaches zero and then becomes positive, i.e., for participants who made more errors than the others, connectivity between the LIFG and the ACC is greater in the IV trials.

Comparisons involving real noun stimuli (RN > IN and IN > RN), as well as factorial analyses for verb and noun conditions, did not yield significant results.

**DISCUSSION**

We believe that the most noteworthy outcome of the present study is that the connectivity analysis allowed us to dissociate regularity and processing difficulty effects and, as we hope to show below, gain a deeper understanding of their nature. Since we are going to compare our results to Stamatakis et al.’s (2005) and Tyler et al.’s (2005), let us start by highlighting some relevant differences between English and Russian verbs.

Stamatakis et al. (2005) and Tyler et al. (2005) looked at stimulus pairs like *stayed – stay vs. taught – teach*. In the regular pairs, the first stimulus was morphologically complex and the second was not, while in irregular pairs, both stimuli were morphologically simple. Obviously, the regular pattern also differs from irregular ones in terms of productivity and type frequency, and it is the morphological default (some authors argue that being the default pattern is a separate property that cannot be reduced to productivity and type frequency, e.g., Cihlens, 1999; Beretta et al., 2003). Behavioral results (error rates) were very similar for regular and irregular sets in this study: 5.1 and 4.3% respectively.

Due to the nature of the Russian language, in our study, all verb stimuli read or produced by the participants were morphologically complex: e.g., *nyr-ja-t’* ‘to dive’ – *nyr-ja-ju* (regular) and *mol-o-t’ ‘to grind’ – *mel-ju* (irregular). The difference in productivity and type frequency is the same as in English. Finally, there was a difference in error rates in our study, indicating that IV’s were more difficult to process. Ideally, all three factors – morphological complexity, regularity and processing difficulty – must be assessed separately and then studied in more detail (for example, to see whether the role of productivity can be dissociated from the role of type frequency etc.). We are infinitely far from this goal now, but arguably, our study lets us make a small step toward it.

Firstly, we observed an increase in functional connectivity between the LIFG and temporal cortex, in particular, the left and right STG, in the RV > IV comparison. Stamatakis et al. (2005)

---

**Table 2 | RV > IV PPI-contrast for a ROI seed in the LIFG (BA 44, LIFGop3).**

<table>
<thead>
<tr>
<th>Brain region</th>
<th>p-value</th>
<th>K</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. STG (BA 22/21)</td>
<td>0.007*</td>
<td>83</td>
<td>-54</td>
<td>-13</td>
<td>-5</td>
</tr>
<tr>
<td>L. putamen</td>
<td>0.028</td>
<td>62</td>
<td>-27</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>R. STG/insula (BA 22/21)</td>
<td>0.0001*</td>
<td>152</td>
<td>54</td>
<td>-11</td>
<td>-1</td>
</tr>
</tbody>
</table>

*Significant clusters after Bonferroni–Holm correction. BA, approximate Brodmann’s area; L/R, left/right hemisphere; k, cluster size in voxels; STG, superior temporal gyrus.

**Table 3 | The effect of task difficulty in the IV > RV PPI-contrast for a ROI seed in the LIFG (BA 44, LIFGop3).**

<table>
<thead>
<tr>
<th>Brain region</th>
<th>p-value</th>
<th>Peak MNI coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>R. ACC (BA 32)</td>
<td>0.013*</td>
<td>29 12 20 31</td>
</tr>
</tbody>
</table>

*Significant clusters after Bonferroni–Holm correction. ACC, anterior cingulate cortex; BA, approximate Brodmann’s area; L/R, left/right hemisphere; k, cluster size in voxels.

\(^{4}\)http://www.nitrc.org/projects/rex/
reported similar findings. They saw a positive influence of LIFG activity on the activity in the left STG and MTG both for regular and irregular real verb trials. Given that the subtractive analysis of the data reported in Tyler et al. (2005) demonstrated that RVs activated the LIFG, bilateral STG and MTG significantly more than irregular ones in this study, the authors conclude that this indicates greater engagement of the fronto-temporal network in RV processing. Since two PPI studies gave similar results in this case, we suggest that this is an effect of regularity.

Stamatakis et al.’s (2005) study and the present study rely on very different languages: morphologically poor English with a clear-cut distinction between regular and IVs vs. morphologically rich Russian, with numerous verb classes and where the notion of regularity is difficult even to define. Experimental tasks were also different: a same/different judgment task for aurally presented stimuli and a production task for visually presented stimuli (which obviously involves a silent reading stage). The fact that our findings partly replicate Stamatakis et al.’s (2005) despite these major differences shows that the observed regularity effect is indeed robust and has cross-linguistic validity. Moreover, it is present both in comprehension and in production, which we consider good news because no major model addressing the problem of regularity defines this notion differently for production and comprehension. An important advantage of our study that strengthens this result is that we used a ROI-whole brain analysis, i.e., did not predefine the set of regions to be analyzed.

Why did Stamatakis et al. (2005) observe coactivation between the LIFG and the left-lateralized temporal brain network, while in our study, both left and right temporal areas were involved? Given the above-mentioned differences between the two studies, explanations can only be very tentative. This could result from task-related differences: for example, in contrast to passive listening, active word production probably involves self-monitoring associated with bilateral STG activation (e.g., Indefrey, 2011). Alternatively, based on the fact that in Tyler et al. (2005) RVs induced an increase of activation in both left and right STG and MTG, one could hypothesize that connectivity changes in the right-lateralized temporal language areas simply did not reach significance in Stamatakis et al. (2005).

In addition to the STG, we also observed an increase in connectivity between the LIFG and the putamen. Since this result did not reach significance after Bonferroni–Holm correction, we will refrain from interpreting it and will only point to some potentially relevant observations in the literature. Numerous studies show that this part of the basal ganglia plays a role in articulation (e.g., Brown et al., 2009, Price, 2010). Some authors also believe that the basal ganglia are part of the system underlying rule-based language processing (e.g., Pinker and Ullman, 2002), but this model is controversial (e.g., Longworth et al., 2005; Macoir et al., 2013). Our observations agree with this model, but could be explained without it. In our stimuli, we matched the length of infinitives, but the present tense forms of some IVs are shorter, so they might require less effort in terms of articulation.

Now let us turn to the results involving the ACC, which did not coincide in the two PPI studies. Stamatakis et al. (2005) found that ACC activity influenced fronto-temporal connectivity, and that the effect was significantly stronger for regulars than for irregulars. They conclude that the ACC plays a monitoring role, “which, in the context of processing real regular inflected words, would reflect greater engagement of an integrated fronto-temporal language system. Morpho-phonological processes, such as the decomposition of regular inflected forms into stems and affixes, may place higher demands on this system, calling on additional resources” (p. 120). Since we did not observe similar results in our study, we hypothesize that this finding is due to the difference in morphological complexity between regular and IVs in English, which is absent in Russian. This hypothesis is very similar to Stamatakis et al.’s (2005) conclusions quoted above, but now we can dissociate morphological complexity from regularity (in the sense of defaultness and/or type frequency).

In our study, we observed covariance between the number of mistakes in the IV trials and functional connectivity changes between the LIFG and the right ACC in the IV vs. RV comparison (see Figure 1C). For participants who had low error rates, LIFG–ACC connectivity was greater during RV trials, while for participants who had high error rates, the opposite was true. We believe that we are dealing with two distinct effects here, and that the former is overshadowed by the latter as processing load increases. We do not have a definitive answer as to why LIFG–ACC connectivity may be greater for RVs. Both regular and irregular forms are morphologically complex in Russian and, if there is any rule-based processing system at all, both engage it (ininfinitival suffixes must be stripped and first person singular endings must be added). However, regular forms might engage this system more than irregular ones: it may also be activated for present tense stem formation. Further research is necessary to test this explanation, but, if it is correct, this would be an argument for the DR approach.

At the same time, LIFG–ACC connectivity increases in IV trials as the processing load they pose grows. This pattern of connectivity changes can be interpreted as a top–down general regulatory effect of the LIFG–ACC interaction, given the fact that the ACC is identified as an important part of the cognitive control network for the detection and resolution of processing conflicts (e.g., Carter and van Veen, 2007; Westerhausen et al., 2010). This effect completely overshadows the one described above when error rates are high. Let us try to formulate more precisely what might be going on. When an irregular form is produced successfully, the stem is simply taken from memory, which is the easiest option for the morphological processing system. But when somebody cannot find the right form and tries to derive it somehow, it is more taxing for the system than dealing with a regular form because the pattern is infrequent and unproductive. In this light, the absence of similar findings in Stamatakis et al.’s (2005) study is expected: different trial types did not differ significantly in terms of processing load in their experiment. This could be due to the fact that Stamatakis et al. (2005) examined comprehension, where one does not have to find or derive any forms. In general, passive comprehension might require more shallow processing than production, and low-status rules or morphological patterns associated with IVs in associative memory might get activated only in the latter case, but not in the former.

Let us briefly comment on the opposite results from Tyler et al. (2005) and Sloufssar et al. (2014). The fronto-temporal language-related areas were activated more for irregulars in the former and
for irregulars in the latter study. We attributed our findings to processing difficulty (more details above in Section “Previous Studies Testing the SR and DR Approaches on Russian”), while Tyler et al. (2005) explained theirs by regularity. In the light of the Section “Discussion” above, we conclude that in both cases, the increased activity levels might reflect greater engagement of the morphological processing system (this does not contradict the conclusions made in these studies and only clarifies the picture). In English, regular forms rely on it more than irregular ones because the former are morphologically complex, while the latter are simpler and do not require any morphological processing at all. In Russian, all forms are morphologically complex, but when people cannot retrieve an irregular form or try to construct a form from a nonce verb, especially from an INV, the morphological processing system has to work harder.

Now, what do our conclusions mean for the DR and SR approaches? In the SR approach, only the frequency of a morphological pattern really matters. In this respect, regular stimuli had the same properties in both PPI studies, yet the results diverged. The canonical version of the DR approach postulates one default rule and argues that all other forms are stored in memory. Again, prima facie this does not predict any differences between regular stimuli in the two studies. One could go on to argue that Russian irregular stimuli must undergo morphological decomposition (at least to get rid of the infinitival affix), and some combination of morphological analysis and memory retrieval processes makes them more difficult than regular stimuli on a certain scale, while English irregular stimuli are the easiest on this scale because no morphological analysis is required at all. Potentially, hybrid models with several rules of different status such as the ones in Yang (2002) are better suited to account for the data. As we mentioned in Section “Previous Studies Testing the SR and DR Approaches on Russian,” such models were proposed for Russian based on the results from behavioral experiments. In any case, it is clear that simplistic views must be discarded.

Further studies are needed to give more definitive answers to the questions above. In particular, the Russian verb system with its numerous classes has much more to offer than what we have used so far. In the present study, we compared verbs from the least frequent unproductive classes to verbs from the most frequent productive AJ class. However, Russian has other highly frequent productive classes. This might allow us to explore the nature of the effects we have observed so far in more detail: what (if anything) would be associated with the morphological default, with productivity, with type frequency, with the complexity of the morphological pattern (e.g., whether it involves stem and suffix alternations etc.)? This might eventually let us figure out what precisely stands behind the regularity effect. Then it will be clear whether it can be accounted for in terms of the DR or SR approach.

Now let us turn to the results for noun stimuli. The fact that the RN > IN comparison gave no significant results is not surprising, given that the main effect of (ir)regularity also did not reach significance for nouns in the factorial analysis in Slioussar et al. (2014). Most probably, this is because the irregular feature we had to select for our noun stimuli was rather minor – the Russian noun system is not very complex in this respect.

Finally, let us look at our data in the context of recent research arguing that fronto-temporal language brain regions are spatially and functionally distinct from the domain-general fronto-parietal multiple demand (MD) system (e.g., Duncan, 2010; Fedorenko et al., 2013; Fedorenko, 2014). In our study, an increase in connectivity between the LIFGop3 region located in one of the language-specific areas and the bilateral STG was driven by linguistic properties of the stimuli (regularity in the sense of defaultness, type frequency and/or productivity). At the same time, we observed how connectivity between this very same region and the right ACC, which is argued to be part of the domain-general cognitive control network, depends on the processing difficulty. As the discussion above shows, the source of this processing difficulty might also be language-specific, namely, it might be a morphological processing difficulty. However, it has an effect on response selection demands, so the cognitive control network must be invoked. In total, in contrast to recent functional connectivity studies arguing for the independence of language-related and domain-general cognitive control systems (Blank et al., 2014), our data demonstrate how these systems can be functionally integrated.

To summarize, the present PPI study allowed us to tease apart processing difficulty and regularity effects in the domain of inflectional morphology: We not only observed the processing difficulty effect we identified earlier in Slioussar et al. (2014), but were also able to find a novel effect of regularity, and gained a better understanding of these two effects by comparing our study to the only other published PPI study of inflectional morphology (Stamatakis et al., 2005). In Slioussar et al. (2014) some regularity-related differences in functional activity could be attenuated by the processing load effect, but the PPI analysis was sensitive enough to reveal such differences in functional connectivity. The present study makes us reevaluate some findings from Stamatakis et al. (2005), Slioussar et al. (2014) and several other previous studies.

ACKNOWLEDGMENTS

The study was partially supported by the grant “Integrative Physiology” from the Department of Physiological Studies of the Russian Academy of Sciences (to Maxim Kireev, Alexander D. Korotkov, and Syryatlaya V. Medvedev) and by the grant #0.38.518.2013 from St. Petersburg State University (to Natalia Slioussar and Tatiana V. Chernigovskaya). We are very grateful to the reviewers for their most valuable comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fnhum.2015.00036/abstract

REFERENCES


**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 17 September 2014; accepted: 14 January 2015; published online: 18 February 2015.


This article was submitted to the journal Frontiers in Human Neuroscience. Copyright © 2015 Kireev, Slioussar, Korotkov, Chernigovskaya and Medvedev. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
Appendix C.

Paper “Forms and features: the role of syncretism in number agreement attraction”


**Abstract**

Many experiments have studied attraction errors in number agreement (e.g. ‘The key to the cabinets were rusty’). It has been noted that singular heads with plural dependents (attractors) trigger larger attraction effects than plural heads with singular attractors, and that in languages with morphological case, morphologically ambiguous attractors trigger larger effects (accusative plural forms coinciding with nominative plural were compared to unambiguous case forms). In Russian, the nominative plural forms of some nouns coincide not only with their accusative plural forms but also with the genitive singular. In one production and two comprehension experiments, such genitive singular forms were found to trigger larger attraction effects than morphologically unambiguous genitive plural forms. Accusative plural forms coinciding with the nominative plural were shown to be the most effective attractors. These results have implications for different models of attraction and for other discussions in morphology concerning ambiguity processing, different approaches to syncretism and the problem of lexical insertion.

**The link on the publisher's website**

This is not an open access paper, so according to the publisher’s policy it cannot be reproduced. It can be accessed on the publisher’s website: https://www.sciencedirect.com/science/article/pii/S0749596X18300305
Appendix D.

Paper 4. “Gender agreement attraction in Russian: production and comprehension evidence”

Gender Agreement Attraction in Russian: Production and Comprehension Evidence

Natalia Slioussar1,2* and Anton Malko3

1 School of Linguistics, Higher School of Economics, Moscow, Russia, 2 Faculty of Liberal Arts and Sciences, Saint-Petersburg State University, Saint-Petersburg, Russia, 3 Department of Linguistics, University of Maryland, College Park, MD, USA

Agreement attraction errors (such as the number error in the example “The key to the cabinets are rusty”) have been the object of many studies in the last 20 years. So far, almost all production experiments and all comprehension experiments looked at binary features (primarily at number in Germanic, Romance, and some other languages, in several cases at gender in Romance languages). Among other things, it was noted that both in production and in comprehension, attraction effects are much stronger for some feature combinations than for the others: they can be observed in the sentences with singular heads and plural dependent nouns (e.g., “The key to the cabinets...”), but not in the sentences with plural heads and singular dependent nouns (e.g., “The keys to the cabinet...”). Almost all proposed explanations of this asymmetry appeal to feature markedness, but existing findings do not allow teasing different approaches to markedness apart. We report the results of four experiments (one on production and three on comprehension) studying subject-verb gender agreement in Russian, a language with three genders. Firstly, we found attraction effects both in production and in comprehension, but, unlike in the case of number agreement, they were not parallel (in production, feminine gender triggered strongest effects, while neuter triggered weakest effects, while in comprehension, masculine triggered weakest effects). Secondly, in the comprehension experiments attraction was observed for all dependent noun genders, but only for a subset of head noun genders. This goes against the traditional assumption that the features of the dependent noun are crucial for attraction, showing the features of the head are more important. We demonstrate that this approach can be extended to previous findings on attraction and that there exists other evidence for it. In total, these findings let us reconsider the question which properties of features are crucial for agreement attraction in production and in comprehension.

Keywords: agreement, gender, attraction, production, comprehension, Russian
1. INTRODUCTION

1.1. The Phenomenon of Agreement Attraction

Grammatical agreement is one of the most basic linguistic operations. It is well-known, however, that it is not always accurate. In the last 20 years many studies have looked at so-called agreement attraction errors, exemplified in (1). In (1a) the verb agrees not with the head of the subject NP key\(^1\), but with another, embedded NP cabinets (we will further call such NPs “attractors”). In (1b) the verb in a relative clause agrees with the subject of the matrix clause.

\[(1)\]
\[\begin{align*}
  & a. \text{The key to the cabinets were rusty} \quad (\text{Bock and Miller, 1991}) \\
  & b. \text{The musicians who the reviewer praise so highly will probably win a Grammy} \quad (\text{Wagers et al., 2009}).
\end{align*}\]

Agreement attraction errors are observed in spontaneous speech and in well-edited texts. They have also been studied experimentally, mostly in English, but also in French, Spanish, Italian, Dutch, German, and some other languages (Bock and Miller, 1991; Vigliocco et al., 1995, 1996; Pearlmutter et al., 1999; Anton-Mendez et al., 2002; Hartsuiker et al., 2003, to name just a few). The first accounts suggested that the verb simply agrees with the linearly closest noun (Jespersen, 1924; Quirk et al., 1972; Francis, 1986, a.o.). However, later studies demonstrated that agreement attraction is a structural phenomenon. For example, Vigliocco and Nicol (1998) showed that people make attraction errors producing questions, e.g., “Are the helicopter for the flights safe?” Various factors that influence attraction have also been identified. However, the overwhelming majority of studies focused on number agreement in the languages where number has only two values: singular and plural. It is not clear to what extent these results can be generalized to other cases.

In this paper, we analyze subject-predicate gender agreement. Gender attraction has been investigated only in a few studies, and mostly in Romance languages, which have two genders. We report one production and three comprehension experiments on Russian, a language with three genders. To the best of our knowledge, this is the first comprehension study looking at agreement attraction in a non-binary category. Below we present several findings from the research on number agreement, which will be most important for our study, and different accounts of attraction. Next, we review the few existing studies on gender attraction, providing rationale for the present work.

1.1.1. Plural Markedness Effect

In all studied languages, attraction effects were found to be asymmetric. They can be observed when the head is singular, and the attractor is plural [as in (1) above], but are much weaker or virtually non-existent in the opposite configuration. In the majority of agreement attraction studies, this asymmetry is explained in terms of feature markedness. Plural is assumed to be the marked value of number feature\(^2\), and the asymmetry is attributed to the fact that attractors with a marked feature are more disruptive. Hence it is known under the name of “plural markedness effect.”

However, the concept of markedness is not widely agreed upon. Different authors adopt different theoretical approaches and different tests to determine marked and unmarked feature values [including frequency, presence of a non-zero affix, default use of a form (e.g., in impersonal sentences), various semantic tests etc.; see Haspelmath, 2006]. It is impossible to evaluate them looking only at singular and plural. To figure out which of these properties may be relevant for the asymmetry between feature values (and whether it makes sense to attribute it to markedness in a particular theoretical framework), it is crucial to look at other features systems. As we will show below, Russian gender is interesting in this respect because the results of different markedness tests do not converge, letting us tease several approaches apart.

1.1.2. Parallel Results in Production and Comprehension

Experimental studies demonstrated that attraction exists not only in production, but also in comprehension. In production it manifests itself as agreement errors. In comprehension attraction errors have been observed to trigger more grammaticality judgment mistakes and to provoke less pronounced effects in reading time and EEG studies than other agreement errors. In other words, people perceive ungrammatical sentences as if they were grammatical or had a minor violation. This is often called a “grammaticality illusion.”

The results from production and comprehension are largely parallel (in particular, significant attraction effects are observed only with plural attractors). This is often used to conclude that the mechanism of attraction is the same in both modalities. We will come back to this problem discussing our findings because we did not observe parallelism that we expected based on the previous studies.

1.1.3. Debate on Ungrammaticality Illusions

We just mentioned that in comprehension, attraction causes grammaticality illusions, making ungrammatical sentences more acceptable. Can it also lead to ungrammaticality illusions, and make grammatical sentences less acceptable? For example, if people tend to miss agreement errors in sentences like (2a), do they sometimes see non-existent errors in sentences like (2b)? As we show below, different approaches to attraction make opposing predictions about ungrammaticality illusions, so this is an important question.

\[(2)\]
\[\begin{align*}
  & a. \text{The key to the cabinets were rusty.} \\
  & b. \text{The key to the cabinets was rusty.}
\end{align*}\]

Several studies (e.g., Nicol et al., 1997; Pearlmutter et al., 1999) suggested that ungrammaticality illusions do arise. However, Wagers et al. (2009) demonstrated that at least on-line findings may be artifactual (they might be due to the fact that processing

---

\(^1\)Here and further, the following standard symbols are used: N, noun; NP, noun phrase; P, preposition; PP, prepositional phrase; V, verb; M, masculine gender; F, feminine; N, neuter.

\(^2\)Notably, in semantics there is an ongoing debate whether singular or plural is the default (e.g., Sauerland et al., 2005; Farkas and de Swart, 2010).
plural nouns carries an additional cost compared to singular ones, not to any aspects of subject-verb agreement processing). This hypothesis can be tested by analyzing some cases where this problem does not apply, and we do so in the present study looking at gender agreement 3.

1.1.4. The Role of Morphophonology

Hartsuiker et al. (2003) showed that when the form of the attractor is morphologically ambiguous and coincides with nominative, the rate of attraction errors increases. They compared German sentences like (3a,3b). People made more errors in (3a), where the attractor (die Demonstrationen) is ambiguous between accusative and nominative, compared to (3b), where the attractor (den Demonstrations) is unambiguously dative. We do not explore the role of morphophonology in the present study, but take this factor into account. Several studies also demonstrated that heads with regular inflections are more resistant to attraction, but no similar counting errors, but also measuring RTs during elicitation tasks (e.g., Staub, 2009, 2010; Brehm and Bock, 2013). They demonstrated that participants slow down when the subject contains a singular head and a plural attractor both when they eventually answer correctly and when they do not [to be precise, Staub observed this for the subjects containing a PP attractor, but not for the subjects contained within relative clauses, as in (1b)].

The more plural the subject NP, the higher the possibility of choosing a plural verb. In such accounts there is no way to avoid ungrammaticality illusions: if the agreement controller can be mis-construed or ambiguous, there is no way to restrict such mis-construals to only ungrammatical sentences. They happen even before we encounter the verb, i.e., even before it is clear whether the sentence is or is not grammatical.

Now let us turn to the retrieval account (Solomon and Pearlmutter, 2004; Lewis and Vaislishth, 2005; Badecker and Kuminiak, 2007; Badecker and Lewis, 2007; Wagers et al., 2009; Dillon et al., 2013). Research on memory suggests that the amount of material a person can hold in a ready-to-process state is extremely limited (McElree, 2006; Cowan, 2001). Thus, it can be hypothesized that when we reach an agreeing predicate, the subject needs to be reactivated. This reactivation can be done via so-called cue-based retrieval (Lewis and Vaislishth, 2005; McElree, 2006): we query the memory with a set of cues (e.g., “number: plural; “case: nominative” etc.) and select an element that matches the maximum number of cues.

This process is not error-free, and the retrieval account argues that attraction arises at this stage. For example, in a sentence like “The key to the cabinets is rusty” the form of the verb suggests that we need to look for an NP with the features “subject” and “plural.” However, no NP perfectly satisfies these conditions: key is the subject, but is not plural, and cabinets is plural, but is not the subject. It is hypothesized that in such conditions we may mistakenly select the wrong NP. The retrieval approach predicts the absence of ungrammaticality illusions: if a sentence is grammatical, the true subject is a perfect match and will always be selected. Thus, unlike in the representational account, there is nothing wrong or ambiguous in the syntactic structure, errors are access failures. Such cases with several elements competing for retrieval are an instance of “retrieval interference.” Other examples are discussed in Van Dyke and Johns (2012).

1.2. Models of Agreement Attraction

There exist two major approaches to agreement attraction. Here they will be referred to as the “representational account” and the “retrieval account.” Models that belong to the representational account share one crucial assumption: agreement attraction takes place because the mental representation of the number feature on the subject NP is faulty or ambiguous (Nicol et al., 1997; Vigliocco and Nicol, 1998; Franck et al., 2002; Eberhard et al., 2005; Staub, 2009, 2010; Brehm and Bock, 2013). In some models, it is assumed that syntactic features can “percolate” or otherwise move to neighboring nodes: for example, sometimes number features from the embedded NP percolate to the subject NP (which normally has the same number marking as its head).

Another model known as Marking and Morphing (Eberhard et al., 2005) postulates that the number value of the subject NP is a continuum, i.e., it can be more or less plural. For example, if a subject NP contains a singular head and a plural dependent NP it is more plural than a subject NP with a singular modifier. A subject NP that is formally singular, but refers to a collective entity is more plural than the ones referring to singular entities.

3In production, looking for symmetric effects in ungrammatical and grammatical sentences is less straightforward. However, several authors suggested not only counting errors, but also measuring RTs during elicitation tasks (e.g., Staub, 2009, 2010; Brehm and Bock, 2013). They demonstrated that participants slow down when the subject contains a singular head and a plural attractor both when they eventually answer correctly and when they do not [to be precise, Staub observed this for the subjects containing a PP attractor, but not for the subjects contained within relative clauses, as in (1b)].

The more plural the subject NP, the higher the possibility of choosing a plural verb. In such accounts there is no way to avoid ungrammaticality illusions: if the agreement controller can be mis-construed or ambiguous, there is no way to restrict such mis-construals to only ungrammatical sentences. They happen even before we encounter the verb, i.e., even before it is clear whether the sentence is or is not grammatical.

Now let us turn to the retrieval account (Solomon and Pearlmutter, 2004; Lewis and Vaislishth, 2005; Badecker and Kuminiak, 2007; Badecker and Lewis, 2007; Wagers et al., 2009; Dillon et al., 2013). Research on memory suggests that the amount of material a person can hold in a ready-to-process state is extremely limited (McElree, 2006; Cowan, 2001). Thus, it can be hypothesized that when we reach an agreeing predicate, the subject needs to be reactivated. This reactivation can be done via so-called cue-based retrieval (Lewis and Vaislishth, 2005; McElree, 2006): we query the memory with a set of cues (e.g., “number: plural; “case: nominative” etc.) and select an element that matches the maximum number of cues.

This process is not error-free, and the retrieval account argues that attraction arises at this stage. For example, in a sentence like “The key to the cabinets is rusty” the form of the verb suggests that we need to look for an NP with the features “subject” and “plural.” However, no NP perfectly satisfies these conditions: key is the subject, but is not plural, and cabinets is plural, but is not the subject. It is hypothesized that in such conditions we may mistakenly select the wrong NP. The retrieval approach predicts the absence of ungrammaticality illusions: if a sentence is grammatical, the true subject is a perfect match and will always be selected. Thus, unlike in the representational account, there is nothing wrong or ambiguous in the syntactic structure, errors are access failures. Such cases with several elements competing for retrieval are an instance of “retrieval interference.” Other examples are discussed in Van Dyke and Johns (2012).

1.3. Studies of Gender Agreement Attraction

Relatively few studies of gender agreement attraction have been conducted so far. Their results do not always converge, but one thing seems to be certain: attraction effects are present. They have been observed in several experiments on different languages.

1.3.1. Previous Studies on Languages with Two Genders

As far as we know, the first attempt to induce gender agreement attraction was made in the production study on Italian by Vigliocco et al. (1995). Virtually no evidence of attraction was found: out of 1920 responses only four (0.2%) contained a gender error. However, in a later study Vigliocco and Franck (1999) observed gender agreement attraction in Italian.

Vigliocco and Franck carried out four production experiments: two on Italian and two on French. Both languages have two genders: masculine and feminine. In all experiments, participants saw a masculine and a feminine adjective at the same time (one above the other) and then a noun phrase, and had to combine them saying the resulting sentence aloud. The gender of the head and the attractor were manipulated. When the genders
mismatched, people were found to make more agreement errors. In Italian, there was no significant difference between FM and MF conditions\(^4\). In French, more errors were made in FM conditions (the difference was significant in Experiment 2 and marginally significant in Experiment 4). Whether the head gender was purely grammatical (on inanimate nouns) or conceptual (on animate nouns) also played a role. Participants made fewer errors in the latter case. Thus, semantic factors do enter the picture in case of gender agreement attraction, but, as far as we can judge, only to suppress it (on the contrary, conceptual numerosity can increase the number agreement attraction rate).

The observed pattern of attraction errors was different from number agreement studies. Firstly, a significant number of errors was made in all mismatch conditions, while in case of number agreement, the error rate in the conditions with plural heads and singular attractors was very low, often the same as the error rate without attraction. Secondly, both in French and in Italian, masculine is used as the grammatical default (for example, it appears in impersonal constructions and in the cases where the predicate must agree with several masculine and feminine nouns) and is more frequent. So the pattern observed in French (more errors in FM conditions) is the reverse of the number agreement attraction pattern found across languages.

The authors concluded that feature markedness does not matter for gender agreement and outlined an explanation based on inflectional differences between Italian and French. However, this explanation was undermined by Anton-Mendez et al. (2002) who conducted a production study on Spanish. Spanish is similar to Italian in terms of adjectival inflections, but the results were the same as in French. In addition to that, Vigliocco and Zilli (1999) and Franck et al. (2008) demonstrated in a number of experiments on Italian, Spanish, and French that the morphophonological properties of the head influence the error rate in gender agreement attraction. As in the studies of number agreement attraction, there were fewer errors when heads had regular inflections, but no similar effects were found for attractors.

We could find only two studies examining gender agreement attraction in comprehension: Acuña-Fariña et al. (2014) and Martin et al. (2014). Both looked at Spanish, eye-tracking was used in the first and ERPs in the second. Attraction effects were detected, but no differences between M and F genders were reported.

### 1.3.2. Previous Studies on Languages with Three Genders

Badecker and Kuminiak (2007) (henceforth, B&K) report results of three production experiments on Slovak. Slovak has three genders: masculine, feminine, and neuter. M is the most frequent, N is the least frequent, but is used in impersonal constructions. In all experiments, participants were given subject NPs (often called “preambles”) and asked to generate complete sentences. In Experiment 1, B&K compared the number of errors in two groups of conditions: MM, MF, FF, FM and MM, MN, NN, NM. As in the previous studies, there were significantly more errors in mismatch conditions than in match conditions. But the pattern was different: there were more errors in the MF condition compared to the FM and in the NM compared to the MN.

Experiment 2 confirmed the results of Experiment 1 (it contained MM, MF, FF, and FM conditions and was designed to test the role of morphophonological factors). In Experiment 3, NN, NM, and NF conditions were compared. NM and NF preambles provoked more errors than NN preambles; but the number of errors in NM and NF conditions was comparable. Explaining this pattern, B&K adopt an optimality-theoretic approach and argue that there is no single markedness hierarchy in the Slovak gender system (such as N < M < F), but markedness is defined in pairs (N < M, N < F, M < F). Among other things, the results of this study show that frequency does not play a role for feature asymmetries.

Another production experiment was conducted on Russian (Lorimor et al., 2008). The authors manipulated both the number and the gender of heads and attractors (only M and F genders were used). In all trials, participants saw and heard the predicate and then saw the preamble. Their task was to construct a sentence using these two parts and to say it aloud. Out of 1155 answers where gender agreement was necessary (in Russian, as well as in Slovak, verbs agree in gender only in past tense singular forms), only seven (0.6%) contained an agreement error. Based on this, the authors concluded that gender agreement attraction does not exist in Russian.

To summarize, in all gender agreement attraction studies, if any effects are observed, error rates in all mismatch conditions are higher than in match conditions (unlike in number attraction studies, where significant effects are found only in one mismatch condition: with singular heads and plural attractors). Otherwise, the results of gender agreement studies are different: larger effects are found in the FM condition (compared to the MF condition) in Spanish and French, and in the MF and NM conditions (compared to the FM and MN conditions) in Slovak. The results from Slovak are closer to the pattern observed for number, if we assume that feminine and masculine genders and plural number are marked.

Out of several approaches to attraction outlined above, the existence of gender agreement attraction is hardly compatible with the Marking and Morphing model, primarily because in the absolute majority of cases, gender features are semantically empty. Moreover, even if we take nouns with conceptual gender, as malčík “boy\(_M\)” or sestra “sister\(_F\)” in Russian, it makes little sense to assume that, for example, having an M dependent NP could make an F noun “more masculine.” Notably, we do not want to say that the existence of attraction with semantically empty features implies that conceptual numerosity cannot play any role for number agreement attraction - various experimental findings clearly indicate that it does (e.g., Bock and Cutting, 1992; Eberhard, 1999; Haskell and MacDonald, 2005; Mirkovic and MacDonald, 2013). We would only like to stress that attraction is possible without any semantic effects of this sort and therefore should result from some process that does not depend on them.
(e.g., from the formal properties of features). Semantic effects can be added to the picture, but this is optional.

1.4. The Present Study

Apparently, gender agreement attraction errors are more difficult to induce than number errors. For example, Vigliocco et al. (1995) did not observe them in Italian, although they were found in subsequent experiments. So we decided to run another production experiment on Russian replicating B&K's first experiment on Slovak (which, in terms of its gender system, is very close to Russian). Our goal was to see whether any attraction errors would be induced, and, if yes, whether the pattern would be similar to B&K's study or to what has been observed for French, Spanish, or Italian. We also planned comprehension experiments because no existing studies had looked at comprehension in a language with three genders. We were particularly interested to find out whether production and comprehension results would be parallel and whether ungrammaticality illusions would be found. Before we move on to the experiments, let us present a brief overview of the Russian gender system.

1.4.1. Russian Gender System

Russian nouns are inflected for number and case, and the ones that have the same endings in the majority of forms are grouped into declension classes. Russian has three declension classes for nouns (and a separate class for substantivized adjectives). The first class includes almost all M nouns (they have zero endings in nominative singular, like mal’čik “boy”) and all N nouns (they have -o or -e endings, like okno “window”). These M and N nouns use the same set of endings in all cases except for genitive plural and nominative and accusative in singular and plural (in plural, all declension classes have the same endings, like mužˇcina “man.”) The third class includes F nouns with zero endings in nominative singular, like devoˇcka “girl.”

In addition to that, there are some irregular and uninflected nouns.

Thus, in most cases, it is impossible to determine the gender of the noun unambiguously looking at the noun itself, and, at least prima facie, we cannot speak of something like morphological markedness in the noun system. Let us add that M nouns are the most frequent and N nouns are the least frequent. M nouns constitute about a half of the lexicon, F nouns - about 30–35%, N nouns are the rest (Yanovich and Fedorova, 2006; Slioussar and Samoilova, 2014).

Gender agreement can be observed only in singular, on adjectives, participles and past tense verb forms. Russian adjectives and participles have so-called full forms (used attributively and predicatively) and short forms (used only in predicates and inflected for number and gender, but not for case). M form is the citation form (i.e., the form would appear in dictionaries, grammatical descriptions etc.).

Verb forms and short forms of adjectives and participles have zero endings in M gender (e.g., byl “wasM” - byla “wasF” - bylo “wasN”), otherwise all forms have non-zero endings (e.g., krasivý “beautifulM,NOM,SG” - krasivá “beautifulF,NOM,SG” - krasivoe “beautifulF,NOM,PL”). Thus, we cannot say that M forms are morphologically unmarked, even if we limit ourselves to predicates. In impersonal sentences, where unmarked forms are expected, N predicates are used, as (4) shows.

(4) Svetalo.

dawnPST.N,SG

It dawned.

As for gender conflict resolution, another classical test for markedness, it is of limited use in Russian because there is no gender agreement in plural. Gender conflict resolution can be observed only in constructions like “X and Y each did something.” We conducted an informal questionnaire, asking about 30 native speakers.

As we discuss below, acceptability of such sentences differs depending on animacy of the nouns and the genders that are combined, and there is substantial individual variation among speakers. However, one crucial generalization can be made: examples with the feminine or neuter forms of každyj “each” are never found even marginally acceptable, only some examples with the masculine forms are.

Firstly, let us consider sentences with M and F nouns, like in (5). Not all speakers of Russian find these examples acceptable, but for those who do, this construction sounds better with human animates (5a) than with non-human animates (5b). Nobody accepts this construction with inanimate nouns, as in 6a), although they can be used in such sentences if both nouns are of the same gender, as in (6b).

(5) a. Mužˇcina i Ženˇšˇcina každyj

manM,NOM,SG and womanF,NOM,SG eachM,NOM,SG

sješi po jabloku

atePST.PL PREP,DISTR appleDAT,SING

b. Jož i svin’ja

hekhekogM,NOM,SG and swineF,NOM,SG

dadog PL

kadogj sjeli po jabloku.

eachM,NOM,SG atePST.PL PREP,DISTR appleDAT,SING

(6) a. Divan i krovat’ každyj

sofM,NOM,SG and bedF,NOM,SG eachM,NOM,SG

stolj celoj sostojanije.

costPST.PL whole,ACC,SG fortune,ACC,SG

b. Kušetka i krovat’ každaja

couchF,NOM,SG and bedF,NOM,SG eachF,NOM,SG

stolj celoj sostojanije.

costPST.PL whole,ACC,SG fortune,ACC,SG

Now let us look at M and N nouns. More than half of the speakers we asked rejected this construction even with animate human nouns (7a) as ungrammatical, but those who accepted it used masculine form. All our informants rejected examples with non-human animates like (7b) or

5Since acceptability ratings for some sentences vary from speaker to speaker, we do not mark any of the examples below with asterisks or question marks used to indicate ungrammaticality or marginal acceptability.
found them only marginally acceptable. This might be at least partly due to independent factors (the relevant neuter words, like mlékopitajúšce “mammal,” životnoe “animal,” nasekomoe “insect,” tend to be abstract), but is still telling.

(7) a. Voin i diťa každyj
    warriorM.NOM.SG and childN.NOM.SG eachM.NOM.SG
    sjeli po jabloku.
    atePST.PL PREP.DISTR appleDAT.SG

b. Gryzun i nasekomoe každyj
    rodentM.NOM.SG and insectN.NOM.SG eachM.NOM.SG
    vypili po kaple.
    drankPST.PL PREP.DISTR dropDAT.SG

Finally, such constructions with F and N nouns, as in (8), were rejected by most of our informants. The few people who accepted them again preferred the masculine form.

(8) Ženščina i diťa každyj
    womanN.NOM.SG and childN.NOM.SG eachM.NOM.SG
    sjeli po jabloku.
    atePST.PL PREP.DISTR appleDAT.SG

Let us add that M nouns are used to refer to groups of people of mixed or uncertain gender, or to an arbitrary member of such groups. This generalization is discussed by Yanovich (2012) who shows that it does not hold for animals. For example, the word sobaka “dog” is feminine. There are specific words to denote male and female dogs, but they are much more often used as swearwords, like the English bitch. To sum up, N appears to be the grammatical default as the gender used in impersonal constructions, while all cases where M is used as the standard gender are limited to the nouns denoting humans and sometimes other animals. In all our experiments, we used only inanimate nouns as heads and attractors (we wanted to avoid additional factors before the general picture becomes clear)\(^6\).

2. EXPERIMENT 1

Experiment 1 was designed to check whether the findings of Badecker and Kuminiak (2007) would be replicated in Russian, which is very close to Slovak in the relevant part of the grammar. In particular, both languages have three genders, M is the most frequent, N is the least frequent, but is used in impersonal sentences. There are no articles. Gender agreement can be observed on adjectives and participles (in singular) and on verbs (in past tense singular). The system of declensions is very similar as well.

2.1. Participants

Thirty native speakers of Russian (8 male, 22 female) participated in Experiment 1. Ages ranged from 18 to 50 (mean age 28.7, SD 9.4). No participant took part in more than one experiment. All experiments reported in this paper were carried out in accordance with the Declaration of Helsinki and the existing Russian and international regulations concerning ethics in research. All participants provided informed consent. They were tested at the Laboratory for Cognitive Studies of Saint-Petersburg State University.

2.2. Materials

In this experiment, participants first saw a predicate, then on the next slide a subject at which point they were asked to produce a complete sentence. In half of the cases, predicates did not agree with the subject in gender, and participants were asked to modify them. Like in B&K’s study, subject noun phrases were always built according to the following schema: NP1–preposition–NP2, e.g., okno vo dvor “windowN,SG to yardM,SG.” NP1 was always in nominative singular, NP2 was in accusative singular. We selected inanimate nouns that have the same form in accusative and nominative, since this was shown to inflate the error rate (Badecker and Kuminiak, 2007). As in many other agreement attraction studies, we had both adjunct and argument PPs.

The predicates always consisted of two words: the copula byt’ “to be” in the past tense (where gender agreement can be observed) and an adjective or participle. We opted for such predicates because they are short and do not contain any objects or other nouns that could cause additional disturbance of subject-predicate agreement (initially, we wanted to use simple verbs, but could not come up with such predicates for all experimental stimuli). Adjectives and participles were always in instrumental singular form\(^7\).

The genders of NP1 and NP2 were manipulated. As Table 1 shows, these two factors were not fully crossed. Like in B&K’s Experiment 1, we used only seven out of nine possible combinations of genders. Additionally, we manipulated the agreement marking on the predicate\(^8\). Sample stimuli in conditions 1–4 in Table 1 represent one set: two variants of the subject NP (one head and two different dependent nouns, or attractors) and two variants of the predicate (matched or mismatched in gender with the subject). We constructed 48 sets, 12 for each of the four combinations of conditions. This approach to the construction of materials (one head noun and several attractors of different genders, plus a grammatical and an ungrammatical version of the predicate) holds for all experiments in this article. All materials are listed in Appendices in Supplementary Material.

In addition to that, we constructed 100 fillers, also consisting of a predicate and a subject. Subject NPs had singular or plural

\(^6\)Vigliocco and Franck (1999) demonstrated that the gender agreement error rate was lower when the gender of the head noun was conceptual, rather than purely grammatical, but we would not expect markedness patterns to be reversed in such cases.

\(^7\)As we explained in the introduction, participles, adjectives, and nouns in predicates can appear either in nominative or in instrumental, and adjectives and participles also have short forms used only in predicates and inflected for gender and number, but not for case. Often only one variant is grammatical, but sometimes two or even three are, or one is fine, while the others are marginally acceptable. Meaning nuances associated with them can be very subtle. It will suffice to say that we chose instrumental forms because, unlike nominative and short forms, they suited all our stimuli. But, if the participants occasionally responded with nominative or short forms, we did not count this as a mistake.

\(^8\)We opted for this design primarily to facilitate the comparison with B&K’s study. In addition to that, when we were pretesting the experiment, we found that the experimental session was relatively short, but very intense because we prompted the participants to respond very fast. We concluded that making it 1.5 times longer to fully cross the two factors could make it too taxing.
heads and adjectival or prepositional modifiers (the NPs inside these PPs were not in accusative). Predicates were similar to the ones in target stimuli and did not agree with subjects in gender in one third of the cases.

Each participant saw only one target stimulus from each set. Consequently, we had four experimental lists with 148 items (48 stimuli and 100 fillers). The number of conditions was balanced for every list. Thus, every participant saw three target items per condition: for example, three FF stimuli (having an F head and an F attractor) with a matched F predicate, three FF stimuli with a mismatched M predicate etc. All lists began with ten fillers, and then fillers and experimental items were presented in a pseudo-random order, with the constraint that no more than two experimental items occur consecutively.

2.3. Procedure

In a pilot experiment, we used the same procedure as in B&K’s study: participants listened to preambles and were asked to generate complete sentences. But after running six subjects, we did not get any attraction errors. This can be explained by the fact that such errors are in general relatively infrequent. In B&K’s study, they occurred in 3% cases on average. Since the number of errors varies from subject to subject, the probability to elicit no errors from several people in a row is considerably high. However, we decided to switch to a different method in the main experiment in hope to elicit more errors.

The experiment was run on a Macintosh computer using PsyScope software (Cohen et al., 1993). In every trial, participants experiment in hope to elicit more errors. However, we decided to switch to a different method in the main

<table>
<thead>
<tr>
<th>Condition</th>
<th>NP1 gender</th>
<th>NP2 gender</th>
<th>Predicate gender</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 / 2</td>
<td>M</td>
<td>M</td>
<td>M / F</td>
<td>had prosročeným / had prosročenjá + recept na porošak</td>
</tr>
<tr>
<td>3 / 4</td>
<td>M</td>
<td>F</td>
<td>M / F</td>
<td>had prosročnený / had prosročnenjá + recept na maz’</td>
</tr>
<tr>
<td>5 / 6</td>
<td>F</td>
<td>F</td>
<td>F / M</td>
<td>had otkrytym / had otkrytym + okno v dvar</td>
</tr>
<tr>
<td>7 / 8</td>
<td>F</td>
<td>M</td>
<td>F / M</td>
<td>had otkrytym / had otkrytym + okno v dvar</td>
</tr>
<tr>
<td>9 / 10</td>
<td>M</td>
<td>M</td>
<td>M / N</td>
<td>byl prosročeným / had prosročenjá + recept na porošak</td>
</tr>
<tr>
<td>11 / 12</td>
<td>M</td>
<td>N</td>
<td>M / N</td>
<td>had prosročeným / had prosročenjá + recept na maz’</td>
</tr>
<tr>
<td>15 / 16</td>
<td>N</td>
<td>M</td>
<td>N / M</td>
<td>had prosročeným / had prosročenjá + recept na maz’</td>
</tr>
</tbody>
</table>

TABLE 1 | Gender combinations used in Experiment 1.

To encourage participants to respond faster, a time counter appeared on the screen after both the predicate and the subject were presented. As soon as the participant responded, the experimenter pressed a key, and the next trial started. All participants’ responses were tape-recorded. An experimental session lasted around 7.5 min.

2.4. Results

The participants’ responses were transcribed, and each of them was assigned into one of the following categories:

- i. Correct response: the sentence is grammatical, the subject and the predicate provided as stimuli are repeated faithfully.
- ii. Agreement error: the sentence is correct except for a gender agreement error.
- iii. Repetition error: the sentence is grammatical, but the subject or the predicate is repeated incorrectly (for example, the word *krem* “cream” was used instead of the word *maz’* “ointments”).
- iv. A combination of a repetition error and an agreement error.
- v. Incomplete response: the participant utters only a part of the sentence or says nothing at all.
- vi. A combination of an incomplete response and an agreement error: the sentence is incomplete, but a verb, a participle or an adjective was uttered and did not agree with the subject (cf. 9a – 9b).

(9) a. Recept na maz’ byla ... recipeM,NOM,SG for ointmentF,ACC,SG wasF,SG ...

b. Recept na maz’ recipeM,NOM,SG for ointmentF,ACC,SG prosročenjá... expiredF,SG ...

Errors in subject-verb gender agreement were the only grammar errors participants made, all other errors involved incorrectly repeating or omitting lexical material (we did not expect any
other grammar errors, for example, in number or case, but they could have occurred accidentally). To exclude mishearings during transcription, both authors of this paper and two other native speakers of Russian listened to all responses to target stimuli. The number of errors in each category is given in Table 2. In case of self-corrections, only the first variant was counted, both when participants changed an answer with an error to a correct one and when they did the opposite (this happened in three cases).

At the following stage of analysis, we collapsed all agreement errors together. The distribution of errors by experimental conditions is given in Table 3. In total, there were 77 agreement errors (5.4% from all responses). Only 13 out of them were not due to attraction (they are discussed in more detail below). The difference between the number of agreement errors with and without attraction is statistically significant according to the chi-square test $\chi^2(1, N = 77) = 18.97, p < 0.01$, so our results show that gender agreement in Russian is subject to attraction.

As Table 3 shows, agreement errors were more frequent in predicate mismatch conditions, but were not limited to them. Out of 13 errors without attraction, in eight cases, a mismatched predicate was not changed, but there were also five cases where participants produced a neuter predicate with an MF subject, a masculine predicate with an NN subject etc., although they were provided with other forms, matched or mismatched with the subject. Out of 64 attraction errors, 11 errors occurred in predicate match conditions, i.e., participants changed the correct gender of the predicate they were provided with to an incorrect one due to attraction.

Conditions with matched and mismatched predicates are collapsed in Table 4 showing that the number of agreement attraction errors differs depending on the combination of genders of the head and attractor nouns. To test whether these differences are statistically significant, we modeled the data with a mixed-effects logistic regression in the statistical software program R (R Core Team, 2014) using the glmer function from the lme4 package (Bates et al., 2015).

Firstly, we compared MF and FM conditions. The logistic regression evaluated the likelihood of an agreement attraction error (coded as 1) vs. a correct response (coded as 0). The combination of genders was treated as a fixed effect. For the predictors we used contrast coding: MF was coded as 0.5, FM was coded as $−0.5$. Random intercepts by participant and by item were also included in the model. The results of the analysis are reported in Table 5. The coefficient for the intercept was significant, reflecting that most responses were correct. There was also a significant main effect of Gender Combination indicating that F attractors trigger significantly more errors than M attractors.

Secondly, we compared MN and NM conditions in the same way. MN was coded as 0.5, NM was coded as $−0.5$. The coefficient for the intercept was again significant because most responses were correct. But the main effect of gender combination did not reach significance. We also compared MF and MN conditions and FM and NM conditions, as well as the number of non-agreement (“other”) errors in different conditions, but did not find any significant differences.

---

TABLE 2 | The distribution of responses in Experiment 1.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Correct responses</th>
<th>Agreement errors</th>
<th>Repetition errors</th>
<th>Repetition and agreement errors</th>
<th>Incomplete responses</th>
<th>Incomplete responses with agreement errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM + M</td>
<td>69</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MN + M</td>
<td>69</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FM + M</td>
<td>69</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FF + M</td>
<td>66</td>
<td>3</td>
<td>1</td>
<td>20</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>MN + N</td>
<td>69</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MM + N</td>
<td>69</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FM + N</td>
<td>66</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FF + N</td>
<td>50</td>
<td>10</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Note:** Agreement errors in parentheses indicate partial agreement.

---

TABLE 3 | The distribution of responses by condition in Experiment 1.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Correct response</th>
<th>Agr. error (attraction)</th>
<th>Agr. error (no attraction)</th>
<th>Other errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM + M</td>
<td>69</td>
<td>0</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>MN + M</td>
<td>66</td>
<td>0</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>FM + M</td>
<td>69</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>FF + M</td>
<td>53</td>
<td>19</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>MN + N</td>
<td>66</td>
<td>1</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>MM + N</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>FM + N</td>
<td>58</td>
<td>11</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>FF + N</td>
<td>66</td>
<td>6</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>MN + M</td>
<td>63</td>
<td>13</td>
<td>0</td>
<td>14</td>
</tr>
</tbody>
</table>

---

TABLE 4 | The Number of gender agreement attraction errors by condition in Experiment 1.

<table>
<thead>
<tr>
<th>Head/attractor gender</th>
<th>Correct responses</th>
<th>Attraction errors</th>
<th>Other errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>119</td>
<td>22</td>
<td>39</td>
</tr>
<tr>
<td>FM</td>
<td>116</td>
<td>11</td>
<td>53</td>
</tr>
<tr>
<td>MN</td>
<td>136</td>
<td>12</td>
<td>42</td>
</tr>
<tr>
<td>NM</td>
<td>125</td>
<td>19</td>
<td>36</td>
</tr>
</tbody>
</table>

---

46
2.5. Discussion
The results of Experiment 1 are similar to the results of B&K’s first experiment, which can be explained by the fact that the two languages have similar gender systems, as we demonstrated in the introduction. In both studies, F attractors triggered more errors than M attractors. N attractors triggered fewer errors than M attractors, but this difference was statistically significant only in B&K’s study. As we mentioned in the introduction, other authors studying gender attraction in French and Spanish (which have two genders and where M is grammatical default), observed a different pattern: there were more errors with M attractors than with F attractors. We postpone further discussion until the general discussion section.

3. EXPERIMENT 2A
Experiment 2a was designed to find out whether gender agreement attraction can also be detected in comprehension. For the sake of comparison with Experiment 1, we used the same combinations of head and attractor noun genders.

3.1. Participants
Forty-eight native Russian speakers (19 female and 29 male) took part in the experiment. Ages ranged from 19 to 26 (mean age 20.9, SD 1.9).

3.2. Materials
The materials consisted of target and filler sentences. All target sentences were 9–10 words long and followed the schema: NP1–preposition–NP2–copula (byt’) - adjective/participle - four-five words modifying the predicate. We had the same 16 conditions as in Experiment 1 (see Table 1 above). Almost all subject NPs and predicates were based on the materials from Experiment 1 and followed the same constraints. In half of the conditions, the predicate did not agree with the subject. Given existing findings on number agreement attraction, we expected parallel results in production and comprehension. In particular, we expected to find grammaticality illusions in conditions MFF, FMM, MNN, and NMM (this would mean that they would be read significantly faster than the other four ungrammatical conditions: MFF, FFM, MMN, NNM).

As in Experiment 1, conditions were grouped in sets, each set containing four conditions with the same head nouns. An example of a stimuli set is given in (10). For each condition set we constructed 12 sentences, 48 target sentences in total.

Additionally, we constructed 120 fillers, which had roughly the same structure as experimental sentences. Subject NPs in fillers consisted of a single noun modified by an adjective, or of a complex NP, where the embedded noun was not in accusative. All fillers were grammatical. Thus, we had 24 ungrammatical and 144 grammatical sentences, making the grammatical-to-ungrammatical ratio 6:1. Experimental sentences and fillers were distributed in four counterbalanced experimental lists. Every list started with ten fillers; then stimuli and fillers were presented in pseudo-random order with the constraint that a maximum of two stimuli could occur consecutively.

3.3. Procedure
The sentences were presented on a PC using Presentation software (http://www.neurobs.com). We used the word-by-word self-paced reading methodology (Just et al., 1982). Each trial began with a sentence in which all words were masked with dashes while spaces and punctuation marks remained intact. Participants were pressing the space bar to reveal a word and re-mask the previous one. One third of the sentences was ungrammatical ratio 6:1. Experimental sentences and fillers were distributed in four counterbalanced experimental lists. Every list started with ten fillers; then stimuli and fillers were presented in pseudo-random order with the constraint that a maximum of two stimuli could occur consecutively.

As in Experiment 1, conditions were grouped in sets, each set containing four conditions with the same head nouns. An example of a stimuli set is given in (10). For each condition set we constructed 12 sentences, 48 target sentences in total.

10 The translation for all sentences is identical, so we only give it for the first one.

TABLE 5 | Results of the analysis for Experiment 1.

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Predictor</th>
<th>Coefficient</th>
<th>Std. error</th>
<th>Wald Z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF vs. FM</td>
<td>(Intercept)</td>
<td>−3.04</td>
<td>0.43</td>
<td>−7.01</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td></td>
<td>GenComb</td>
<td>−0.95</td>
<td>0.48</td>
<td>−1.96</td>
<td>0.05</td>
</tr>
<tr>
<td>MN vs. NM</td>
<td>(Intercept)</td>
<td>−2.68</td>
<td>0.30</td>
<td>−8.82</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td></td>
<td>GenComb</td>
<td>0.62</td>
<td>0.39</td>
<td>1.59</td>
<td>0.11</td>
</tr>
</tbody>
</table>
3.4. Results
We analyzed participants’ question-answering accuracy and reading times. Two participants answered more than 20% questions incorrectly, so their data were discarded. Otherwise no participant made more than two mistakes when answering questions to target sentences (i.e., 10% at most). Reading times that exceeded a threshold of 2.5 standard deviations, by region and condition, were excluded (Ratcliff, 1993). For two participants, this led to the exclusion of more than 15% responses, so we did not include their data in further analysis.

After four participants were excluded, we had 44 participants (11 in each experimental list). In total, 2.3% of the data were excluded as outliers (never more than 3.6% per region and condition). Average RTs per region in different conditions are presented in Figure 1.

The data for each set of conditions (e.g., MMM - MFM - MMF - MFF) were entered in a 2 × 2 Repeated Measures ANOVA with grammaticality and gender match between the attractor and the head nouns as factors. We used IBM SPSS software (www.ibm.com/software/analytics/spss/). Analyses by items and by participants were performed. Data from all regions were tested, but there were significant results only in regions 4–6 in the conditions with M heads and in regions 5–6 in the conditions with F and N heads. Region 4 is the copula, region 5 is an adjective or participle, regions 6–10 contain several words modifying the predicate. The results of the tests for the relevant regions are given in Table 6.

3.4.1. Feminine Head, Masculine Attractor
The main effect of Grammaticality is significant in analysis by subjects and by items in regions 5–6, reflecting the fact that ungrammatical sentences were read slower than grammatical ones. The main effect of Gender Match is not significant in any region. The interaction of Grammaticality and Gender Match is significant in analysis by subjects and by items in region 5 and only in analysis by subjects in region 6. Ungrammatical sentences were read faster if the head and the attractor were mismatched in gender (i.e., in the FMM condition compared to the FFM condition). This is the classical attraction pattern.

3.4.2. Neuter Head, Masculine Attractor
The main effect of Grammaticality is significant in regions 5–6, reflecting longer RTs in ungrammatical conditions. The main effect of Gender Match is significant only in analysis by subjects in regions 5–6. The interaction of Grammaticality and Gender Match is significant in regions 5–6, which is again a reflection of the classical attraction pattern: NMM condition was read faster than NNM and, in fact, almost as fast as grammatical conditions.
3.4.3. Masculine Head, Feminine Attractor
The main effect of Grammaticality is significant in analysis by subjects in region 4 and in analysis by subjects and by items in regions 5–6. This reflects the fact that RTs were longer in ungrammatical conditions. The main effect of Gender Match is significant in analysis by subjects and by items in region 4, and only in analysis by subjects in regions 5–6. This corresponds to longer RTs in conditions where the genders on the nouns were mismatched. The interaction of Grammaticality and Gender Match did not reach significance in any regions, which points to the absence of agreement attraction.

3.4.4. Masculine Head, Neuter Attractor
The main effect of Grammaticality is significant in analysis by subject and by items in regions 5–6: the ungrammatical conditions are read slower than grammatical. The main effect of Gender Match is significant only in analysis by subjects in regions 4–6. The interaction of Grammaticality and Gender Match is not significant in any region, so these conditions also show no agreement attraction.

3.5. Discussion
As can be seen from the analyses, the results fall into two groups. In the conditions with F or N heads and M attractors there is clear evidence for gender agreement attraction. RTs exhibit the classical attraction profile with grammaticality illusions: ungrammatical sentences where the attractor and the predicate have the same gender (FMM and NMM) are read faster than other ungrammatical sentences (FFM and NNM). Discussing comprehension studies of number agreement attraction in the introduction, we outlined different approaches to this phenomenon, but will opt for one of them ourselves only in
4. EXPERIMENT 2B

In this experiment we follow up on potential frequency effects in the conditions with M heads from Experiment 2a.

4.1. Participants
Thirty-five native Russian speakers (17 female, 18 male) took part in the experiment. Ages ranged from 21 to 47 (mean age 31.3, SD 6.2).

4.2. Materials
We constructed 32 sets of stimuli according to the same schema as in Experiment 2a and observing the same constraints. Head nouns were always masculine. In 16 sets, the attractors were masculine and neuter; in the other 16 sets, the attractors were masculine and feminine. Most of the head nouns were re-used from the Experiment 2a, but we replaced attractors so that their frequencies were closely matched inside the two groups of conditions. We used The Frequency Dictionary of Modern Russian Language (Lyashevskaya and Sharoff, 2009). Average frequencies of head and attractor nouns in Experiments 2a and 2b are shown in Table 7. As in Experiment 2a, half of the predicates did not agree with the subject in gender. Additionally, we used 80 fillers from Experiment 2a. Experimental sentences were distributed into four experimental lists, with factors counterbalanced. As a result, we had 112 sentences per list (16 ungrammatical and 96 grammatical), making the grammatical-to-grammatical ratio 6:1.

4.3. Procedure
The procedure was the same as in Experiment 2a. An experimental session lasted around 9 min.

4.4. Results
Like in Experiment 2a, we analyzed participants’ question-answering accuracy and reading times. At the first stages of analysis, the data from three participants were discarded: one of them had <75% accuracy in comprehension questions; the other two read too slowly compared with the others, so more than 15% of their RTs would have to be excluded as outliers (exceeding the threshold of 2.5 standard deviations). As a result, we had 32 participants, eight for each experimental list.

After three participants were excluded, on average, 1.5% RTs were excluded as outliers (never more than 3.1% per region and condition). Average RTs per region in different conditions are presented in Figure 2.

2 x 2 Repeated Measures ANOVAs with grammaticality and gender match as factors were used to analyze RTs, as in Experiment 2a. Significant results were found only in regions 5 (adjective/participle) and 6–7 (spillover regions). They are presented in Table 8.

4.4.1. Masculine Head, Feminine Attractor
The main effect of Grammaticality was significant in analysis by subjects and by items in regions 5–6, and only in analysis by subjects in region 7. This reflects the fact that ungrammatical sentences were read slower than grammatical ones. The main effect of Gender Match was significant only in analysis by subjects in regions 5–7. The interaction between Grammaticality and Gender Match was not significant in any region.

4.4.2. Masculine Head, Neuter Attractor
The results were almost the same as in the other set of conditions. The main effect of Grammaticality was significant in regions 5–7. The main effect of Gender Match was significant only in analysis by subjects in regions 5–7. The interaction between the factors never reached significance.
### 4.5. Discussion

The results of this experiment show that the basic finding from Experiment 2a holds: there is no evidence for agreement attraction in the sentences with M heads. The plots of the data also suggest that the unbalanced frequencies in Experiment 2a had some influence on reading times. In Experiment 2b, where this confounding factor was excluded, two ungrammatical and two grammatical conditions pattern more closely together within each condition set. Still, the conditions where the genders of heads and attractors are mismatched have longer RTs.

Notably, this difference in RTs is not an instance of ungrammaticality illusion, since it is observed in both grammatical and ungrammatical conditions. In case of illusions, a different pattern would be expected: gender mismatch between the head and the attractor should increase RTs in grammatical conditions and decrease RTs in ungrammatical ones. Rather, it can be suggested that gender mismatch carries some processing cost in the sentences with M heads. In any case, our data do not allow for strong claims: the main effect Gender Match is significant in by subjects analysis in regions 5–7, but never reaches significance in by items analysis.

Since the outcome of comprehension experiments was not parallel to the results of Experiment 1 and earlier experiments on Slovak (Badecker and Kuminiak, 2007), we decided to look...
at the remaining combinations of head and attractor genders in Experiment 3 before suggesting an explanation.

5. EXPERIMENT 3

In this experiment, we studied sentences with N heads and N, F, and M attractors and sentences with F heads and F, N, and M attractors in comprehension. NF and FN combinations have not been examined before, and we added M attractors to be able to compare sentences with all possible attractors.

5.1. Participants

Thirty-nine native Russian speakers (22 female, 17 male) took part in the experiment. Ages ranged from 19 to 40 (mean age 25.4, SD 6.4).

5.2. Materials

We constructed 36 sets of stimuli according to the same schema as in Experiments 2a and 2b and observing the same constraints. Half of the sets had F head nouns and the other half had N head nouns. In all sets, we used M, N, and F attractors. Their frequency was closely matched inside the three groups of conditions, as Table 9 shows. Half of the predicates were grammatical, and half were not. As a result, every target sentence appeared in six conditions: NNN, NNF, NMN, NMM, NFN, NFN for the sentences with N heads and FFF, FFN, FNF, FMM, FMF, FNN for the sentences with F heads. Thus, out of all possible combinations of head, attractor and predicate genders, we did not use NNM and FFM. We decided to do so to keep the number of grammatical and ungrammatical conditions equal and sacrificed two conditions without any potential for agreement attraction that we have already looked at in Experiment 2a. Additionally, we used 100 fillers from Experiment 2a. Experimental sentences were distributed into six experimental lists, with factors counterbalanced. As a result, we had 136 sentences per list (18 ungrammatical and 118 grammatical), making the grammatical-to-ungrammatical ratio 6.6:1.

5.3. Procedure

The procedure was the same as in Experiments 2a and 2b. An experimental session lasted around 11 min.

5.4. Results

We analyzed participants’ question-answering accuracy and reading times. The data from three participants were discarded because they had <75% accuracy in comprehension questions. As a result, we had 36 participants, six for each experimental list. None of them made more than two mistakes when answering questions to target sentences (i.e., 12.5% at most).

As in the previous experiments, reading times that exceeded a threshold of 2.5 standard deviations, by region and condition, were excluded. In total, 1.8% of the data were excluded (never more than 3.7% per region and condition). Average RTs per region in different conditions are presented in Figure 3 (notice that coloring conventions are different from the previous plots).

5.4.1. Feminine Head, Neuter Attractor

The main effect of Grammaticality was significant in analysis by subjects and by items in regions 5–6. This reflects the fact that ungrammatical sentences were read slower than grammatical ones. The main effect of Gender Match was significant only in analysis by subjects in regions 5–6. The interaction of Grammaticality and Gender Match was significant in analysis by subjects and by items in region 6 and only in analysis by subjects in region 5. Ungrammatical sentences were read faster if the head and the attractor were mismatched in gender (i.e., in the FNN condition compared to the FFN condition). This is the classical attraction pattern, also known as a grammaticality illusion. At the same time, there are no differences between grammatical conditions, i.e., no evidence of ungrammaticality illusions was found.

5.4.2. Neuter Head, Feminine Attractor

The results were the same as in the other set of conditions. Thus, the answer to our first experimental question was positive, so we proceeded to compare the size of the attraction effect for attractors of different genders. We compared two groups of conditions: FNF, FNN, FMF, FMM and NFF, NFN, NMN, NMM. We used 2 x 2 Repeated Measures ANOVAs with grammaticality and gender match as factors. Only the main effect of Grammaticality in region 6 was statistically significant [for conditions with F heads, \( F_1(1, 35) = 19.31, p < 0.01, M_{S\text{effect}} = 86064.00; F_2(1, 17) = 10.17, p = 0.01, M_{S\text{effect}} = 24457.35\); for conditions with N heads, \( F_1(1, 35) = 55.80, p < 0.01, M_{S\text{effect}} = 126973.44; F_2(1, 17) = 7.32, p = 0.02, M_{S\text{effect}} = 52915.47\)]. The main effect of Attractor Gender or the interaction between the factors were not significant in any region.
5.5. Discussion

Let us summarize the results of Experiments 2a, 2b, and 3. Gender agreement attraction was observed with F heads and M or N attractors and with N heads and M or F attractors, but not with M heads and F or N attractors. This leads us to the conclusion that attraction depends primarily on the features of the head rather than on the features of the attractor. If the features of the attractor played an additional role, ungrammatical sentences with M attractors would be read faster than ungrammatical sentences with other attractors. However, when we compared sentences with F heads and N or M attractors and sentences with N heads and F or M attractors, the Attractor Gender or the interaction between this factor and Grammaticality never reached significance, and average RTs even showed the opposite pattern: they were longer in the ungrammatical sentences with M attractors. This goes against the assumptions entertained in the absolute majority of previous agreement attraction studies, so a detailed analysis of this result will be presented in the General Discussion Section.

6. GENERAL DISCUSSION

In this paper we reported four experiments on gender agreement attraction in Russian. We observed attraction effects both in production and in comprehension. Badecker and Kuminiak (2007) is the only previous production study where gender agreement attraction was examined in a language with three genders (Lorimor et al., 2008 elicited very few gender errors in their experiments on Russian). In this paper, we replicated one of Badecker and Kuminiak’s experiments and conducted the first comprehension experiments analyzing attraction with non-binary features.

Two outcomes of our experiments can be identified as the most important. Firstly, our results suggest that gender attraction works differently in production and comprehension. This does not agree with previous studies of number agreement attraction, in which production and comprehension results were largely parallel: only the combination of a singular head and a plural attractor triggered attraction. Secondly, our reading experiments...
suggest that the features of the head, rather than the features of the attractor are crucial to determine the pattern of agreement attraction, while the absolute majority of previous agreement attraction studies rely on the opposite assumption.

6.1. Overview of Experimental Findings
In our comprehension experiments, attraction was observed in some combinations of head and attractor genders, but not in the others, while in the production experiment, all combinations exhibited attraction, only to a different extent. We will first consider production results, and then comprehension findings. The outcome of the production study was similar to the results of the first experiment conducted by Badecker and Kuminiak (2007): there were more errors with MF subjects than with FM subjects and with NM subjects than with MN subjects. Both differences were statistically significant in the Slovak study, while in our experiment, only the first one was.

Badecker and Kuminiak ran an additional experiment comparing NF and NM preambles and found that the error rates in these conditions were roughly the same. They claim that this pattern can be explained only in an optimality-theoretic framework where markedness effects are by definition relational. We believe that this is not the case. Given the impressive body of literature on number and gender features, we do not think that we can select a particular approach based on experimental data without a detailed consideration of other arguments. So we chose two models that have been applied to Russian to demonstrate that they are also compatible with the pattern described by Badecker and Kuminiak and may be better suited to explain other findings we reported.

In Kramer (2015), F is encoded as [+FEM], M is [−FEM] and N corresponds to no gender features. When zero and non-zero feature values are compared, the latter are marked, and it can be argued that for this comparison, it is not important whether non-zero values are plus or minus. Therefore, the same error rates are observed with NF and NM preambles. When non-zero values are compared, plus values are more marked. In Nevins (2011), F is [+FEM], [−MASC], M is [−FEM], [+MASC] and N is [−FEM], [−MASC]. N is less marked than M and F because it contains only minus values, while M and F both contain one plus value. But when we compare F and M directly, it can be argued that feature hierarchy becomes important. [FEM] is standardly assumed to be lower than [MASC], so F is more marked than M.

Now let us focus on another property of production findings from Slovak and Russian that is not discussed by Badecker and Kuminiak (2007), but seems crucial to us. In case of gender agreement, attraction errors are produced with all preambles in which the genders of the head and the attractor are mismatched, while in case of number agreement, errors are virtually absent with plural heads and singular attractors. One way to capture this would be to assume that all genders are marked by some feature combinations, as Nevins (2011) suggests, while singular corresponds to no number features.

Another important problem is the difference between experimental findings from Slovak and Russian on the one hand and Romance languages on the other. In Russian and Slovak, more errors are produced with MF preambles than with FM preambles, while in Spanish and French the situation is the opposite. Badecker and Kuminiak (2007) do not comment on this discrepancy, and we cannot offer any explanation for it so far. We can only note that the pattern observed in Slovak and Russian is similar to what we see with number: more errors when the head is less marked than the attractor.

Now let us turn to comprehension experiments. Attraction was observed in NMM, NFF, FMM, and FNN conditions, but not in MFF and MNN conditions. As we already noted, this indicates that features of the heads rather than features of the attractors play a crucial role for attraction. Before discussing this finding in the next section, we want to make two important observations. Firstly, the M gender exhibits a different pattern from the F and N genders. This can hardly be attributed to feature markedness: N is the grammatical default in Russian, and the psycholinguistic relevance of this fact is confirmed by the production data discussed above. We will explore alternative explanations below. Secondly, no ungrammaticality illusions (differences between grammatical conditions depending on whether the head and the attractor have matched or mismatched gender features) were detected in our experiments, which lends further support to the retrieval approach to agreement attraction.

6.2. The Role of Head and Attractor Features in Attraction
In the literature on agreement attraction, the presence or absence of the effect is traditionally associated with the features of the attractor. There are at least two reasons for this. Firstly, experimental findings suggest that some properties of attractors do influence attraction effects [e.g., as we discussed in the introduction, Hartsuiker et al. (2003) showed that the incidence of agreement errors was much higher when attractors were formally similar to nominative plural forms]. The second reason is tradition: the first proposed account of agreement attraction relied on feature percolation, which means focusing exclusively on the attractor whose features can erroneously spread upwards.

The assumption that the features of the attractor are crucial has been maintained in the more recent retrieval account. However, it is important to realize that in this account the properties of the head can influence the agreement process as well. For example, to explain the plural markedness effect, it is traditionally assumed that singular nouns are not marked for number, and “the system is biased to return explicitly number marked constituents” (Wagers et al., 2009, p. 233), therefore plural attractors can easily be retrieved, while singular ones almost never are. But another interpretation is possible: the plural feature makes the heads easier to retrieve and thus more stable, less prone to attraction errors. This is why attraction in the plural-singular configurations is virtually non-existent. On the other hand, the retrieval of singular heads is prone to error, hence the abundance of errors in singular-plural configurations.

While we look at binary features or at the cases where attraction is observed in all feature combinations (as in production experiments on Slovak and Russian), we can only

---

11 Let us note that under this scenario an attractor can also be retrieved in a singular-singular configuration, but this will not provoke any agreement errors.
use indirect evidence to estimate the contribution of head and attractor features to the agreement process. Our reading experiments allow for the first direct comparison and show that at least in comprehension, the features of heads, not attractors play the crucial role. We observed attraction with attractors of all three genders, but only with N and F heads. The gender of the attractor did not even influence the size of the effect. These results suggest that the gender of the attractor has very little or no influence on its chances to be retrieved (it should only match the gender of the incorrect verb form).

Notably, Julie Franck expressed similar ideas in a recent talk (Franck, 2015). The first part of the talk was dedicated to summarizing existing data on agreement attraction. Franck adopted the retrieval approach for production and comprehension and identified the following groups of factors that can lead to attraction: semantic factors (primarily related to the conceptual numerosity of the subject NP), stability of the head's features, accessibility of the attractor (defined by its structural position) and similarity between the head and the attractor. Discussing stability of the head's features Franck examined asymmetries between feature values, morphophonological and semantic influences.

Franck's reexamination of attraction phenomena was driven by the findings on morphophonology (other data she considered could be accounted for in the old models). As we noted in the introduction, studies on several languages demonstrated that number and gender agreement attraction errors are less frequent when heads have regular inflections, but this plays no role for attractors (e.g., Bock and Eberhard, 1993; Vigliocco et al., 1995; Vigliocco and Zilli, 1999; Franck et al., 2008). For attractors, only morphological ambiguity making them more similar to a subject is important (e.g., Hartsuiker et al., 2003; Badecker and Kuminiak, 2007). This led Franck to conclude that the features of the head are crucial, and she reanalyzed existing data according to this idea. She argued that features that have a semantic correlate are more resistant to attraction (for example, Vigliocco and Franck, 1999 observed lower error rates when heads had conceptual rather than purely grammatical gender) and that the same is true for marked feature values. The latter conclusion was based on number agreement attraction findings and on the results of Badecker and Kuminiak's and our production experiments.

Thus, the findings summarized by Franck and the outcome of our reading experiments point into the same direction, but we still have to explain the difference between our comprehension and production results. Of course, to make definitive conclusions, it would be great to have data from several languages (for example, comprehension data from Slovak), but let us suggest several hypotheses based on existing findings. Our reading experiments strongly indicate that M heads are resistant to attraction, while N and F heads are not. The data from production experiments on Russian and Slovak are open to several interpretations because attraction was observed in all head-attractor combinations with mismatched genders. Therefore, we assume that M heads in general are the most stable ones and the least prone to attraction, and production data need an independent explanation. This assumption is supported by independent evidence: several production experiments on number agreement attraction in Russian reported by Nicol and Wilson (1999) and Yanovich and Fedorova (2006) demonstrated that the incidence of number errors depends on the gender of the head noun. Errors arise most often with N heads and least often with M ones.

If our assumption is on the right track, M heads and plural heads exhibit similar properties in comprehension. But why should they do so, given that M features are neither the most marked nor the least marked in Russian? Let us come back to the idea expressed in the previous subsection: number is privatively marked (i.e., singular nouns have no number features), while gender is not (all nouns have some gender features with plus and minus values). We hypothesize that with privative features, the non-zero value is the most stable, while with non-privative features, where all values are non-zero, other considerations come into the picture. We are reluctant to appeal to frequency, but maybe it plays a role that M gender vastly outnumber F and N in Russian. In any case, our data indicate that there is no straightforward relation between feature markedness and stability. The next subsection considers some differences between comprehension and production and how these differences could explain our results.

### 6.3. Differences between Production and Comprehension

Based on parallel results from number agreement attraction experiments most authors assume that the same mechanisms underlie attraction in production and comprehension. The opposite view has been recently advocated by Tanner et al. (2014). They claim that the mechanisms responsible for attraction in comprehension are a subset of those involved in production. In particular, they argue that attraction in comprehension is due to retrieval interference, while attraction in production is best described by the representational account, namely, by the Marking and Morphing model (Eberhard et al., 2005), although retrieval interference is also present.

As we noted above, the Marking and Morphing model is incompatible with gender agreement attraction. We believe that the core mechanism underlying number and gender agreement attraction in production is the same, so we opt for the retrieval approach. Evidently, in case of number, semantic factors influence agreement, and it is expected that their influence is much more readily detected in production than in comprehension: in production, we start with the conceptual structure, while in comprehension, it is our goal. Vigliocco and Franck (1999) demonstrated that gender agreement attraction errors are less frequent when head nouns have conceptual, rather than purely grammatical gender. So semantic factors also play a role here, but, given the relevant distinctions between number

---

12 Let us add that Badecker and Kuminiak (2007) demonstrated that ambiguity is important not only for attractors, but also for heads: if the form is ambiguous between nominative and accusative, the chances of the head to be retrieved are lower.

13 In case of number, we can find many words that are formally singular, but denote plural entities, for example, nouns like crowd or heads of the phrases like the label on the bottles that have a distributive interpretation. Gender is usually semantically
and gender, this role is different: they mainly reduce the size of the effect. It would be very interesting to assess their influence on gender agreement attraction in comprehension: we expect that it should be much smaller, as in case of number agreement. Thus, the differences between production and comprehension noted by Tanner et al. (2014) may also be relevant for gender agreement, but the picture revealed by our experiments cannot be explained by them.

In the previous subsection we argued that agreement attraction patterns in comprehension are due to the fact that heads with plural features and M features are resistant to attraction, i.e., that during the retrieval process, they tend to be identified correctly, while the retrieval of heads with other features can be disturbed by attractors. Findings summarized by Franck (2015) show that the stability of head's features should also be relevant for agreement attraction in production. This is further confirmed by the results from Nicol and Wilson (1999) to Yanovich and Fedorova (2006) indicating that heads with M features are indeed more stable when we look at number agreement production in Russian. Based on these data, we would expect to see no errors in MF and MN conditions in production experiments on gender agreement, but this is not what we found.

To address this problem, we should specify in more detail how retrieval may work in comprehension and production. Wagers et al. (2009) who analyze comprehension show that the retrieval account has two versions that may be difficult to tease apart based on the current experimental data. On the one hand, cue-based retrieval may be initiated every time we deal with an agreeing verb. On the other hand, we may predict the features of the upcoming verb relying on the subject NP and initiate retrieval only when our predictions are not met. Both versions give roughly the same results if we assume that when the true subject matches all the cues, it is successfully retrieved in the absolute majority of cases. Then in both scenarios, problems are expected only when we encounter an incorrect verb form and the sentence contains an attractor a non-subject NP that matches the incorrectly specified feature of the verb.

We believe that two similar scenarios can also be distinguished for production: we can decide which features we need on an agreeing predicate while processing the subject or once we get to the predicate. Accordingly, retrieval might be initiated every time we deal with an agreeing predicate or only when a wrong verb form that does not match our predictions is spuriously generated. The models proposed by Solomon and Pearlmutter (2004) or by Badecker and Kumiñak (2007) instantiate the first scenario. For example, Solomon and Pearlmutter argue that attraction in production arises because two nouns, the head of the subject NP and the attractor, are simultaneously active in the syntactic structure, and a wrong agreement controller may be selected. However, we argue for the second scenario below.

To summarize, in comprehension, we construct the set of retrieval cues based on the verb form that is provided to us. As we demonstrated above, different versions of the account share this basic observation. If the first scenario is adopted for production (the features of the upcoming verb are predicted, and retrieval is initiated only when we spuriously generate a wrong verb form), the picture should be quite similar: the set of cues will be based on this form.

However, we do not believe that this scenario is the most plausible. In particular, it implies that we generate the subject NP with all its feature specifications before we turn to the verb. In reality, the process should be much more complicated. On the one hand, we cannot determine the case of an NP before we select the predicate (for example, experiencers may receive nominative, accusative, or dative case in Russian, depending on the verb, so it is impossible to plan a nominative NP having only some abstract V in mind). On the other hand, we cannot select some features of the verb form without looking at the subject.

This leads us to adopt the second scenario, in which the relevant features are retrieved at some point during the derivation, rather than predicted and then rechecked. Then we do expect certain differences between production and comprehension. Namely, under the second production scenario it is not the case that we look for an NP with a particular number or gender feature. Rather, we look for the values of number and gender features inside the subject NP. These features should belong to the head of this NP, but sometimes we spuriously pay attention to the features of other nouns. We hypothesize that feature markedness plays a role in this process, and this is what causes different outcomes in our production and comprehension experiments.

To explain how markedness effects may arise, let us summarize different factors that have been shown to play a role for retrieval. More stable head nouns have more chances to be retrieved than less stable ones. Structurally accessible attractors looking like subjects have more chances to be retrieved than the attractors without these characteristics. This is true both for production and for comprehension. And, independently of these factors, marked features have more chances to be retrieved. In comprehension, when we encounter a particular verb form and construct a set of retrieval cues based on it, different number or gender features do not compete with each other: we always look for a particular value. In production, we need to find the value of the gender feature of the subject NP, there is no value that is provided in advance, thus different values may enter the competition. Thus, production involves competition and comprehension does not, therefore we can observe feature markedness effects in production, but not in comprehension. This is why production and comprehension results for gender agreement are different. We do not observe any differences in case of number agreement because plural is at the same time a more stable feature and a marked one. This is a very tentative hypothesis, so further experiments are necessary to test it or to suggest an alternative explanation for the observed asymmetry between production and comprehension findings.

\[^{14}\text{In our production experiment, participants were provided with predicates in a particular form. Still, we also expect competition here because participants had to produce a correct form if the provided form was wrong, and to do so, they had to retrieve the subject NP and determine its gender.}\]
AUTHOR CONTRIBUTIONS
All authors listed have made substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS
The study was partially supported by the grant #16-18-02071 from the Russian Science Foundation. We are grateful to many colleagues for their valuable comments and would especially like to thank Colin Phillips. We are also very grateful to the reviewers.

SUPPLEMENTARY MATERIAL
The Supplementary material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.01651/full#supplementary-material


**Conflicts of Interest Statement**: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Slioussar and Malko. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
Appendix E.

Paper “Gender, declension and stem-final consonants: an experimental study of gender agreement in Russian”

GENDER, DECLENSION AND STEM-FINAL CONSONANTS: AN EXPERIMENTAL STUDY OF GENDER AGREEMENT IN RUSSIAN

Slioussar N. A. (slioussar@gmail.com)
Higher School of Economics, Moscow, and Saint-Petersburg State University, Saint-Petersburg, Russia

Every adult native speaker of Russian knows that kon’ is masculine and lan’ is feminine, although 3rd declension nouns present some difficulties in the first and second language acquisition. However, will the fact that these nouns are less frequent than masculine nouns ending in a consonant or feminine nouns ending in -a/ja play a role for online subject-predicate agreement processing? Or will subject-predicate agreement processing be more problematic with subjects of a certain gender? Finally, some final consonants are more characteristic for feminine gender, while the others for masculine gender. Are speakers sensitive to this? We present two experiments addressing these questions. We found that all three factors play a role, but for different tasks (online agreement processing or determining the gender of a novel word) and at different processing stages.

Keywords: grammatical gender, declension, experimental, Russian

1. Introduction

The gender of many Russian nouns cannot be determined from their inflectional affixes. But nouns are not distributed evenly among genders and declensions. Firstly, masculine nouns are in general more frequent than feminine and neuter (48% vs. 35% and 17% forms in the grammatically disambiguated subcorpus of the Russian
Gender, Declension and Stem-final Consonants

National Corpus (RNC, http://www.ruscorpora.ru)). Secondly, a nominative singular form ending in -a/ja or with a zero inflection may be masculine or feminine, but the former is much more likely to be feminine, while the latter is much more likely to be masculine (more details about are given below). Thirdly, if the final consonant of a noun with a zero inflection is taken into account, we will see that, for example, most nouns ending in -s’ are feminine, while most nouns ending in -r’ are masculine.

In this paper, we present two experimental studies exploring whether native speakers of Russian are sensitive to the distributional properties outlined above. The first experiment is dedicated to the online processing of gender agreement. In the second pilot experiment, we study how participants determine the gender of real and nonce nouns.

Now let us discuss the relevant properties of nouns in more detail. Table 1 shows the distribution of nouns among genders and declensions in the grammatically disambiguated subcorpus of the RNC. Masculine nouns ending in -a/ja and feminine nouns with a zero inflection are termed non-prototypical due to their low relative frequency. These nouns are known to be problematic for the L1 and L2 acquisition (e.g. [Janssen, 2016]; [Rodina & Westergaard, 2012]; [Schwartz et al., 2015]; [Tseitlin 2000]). But [Rusakova 2013] who studied naturally occurring errors in spoken Russian found that adult native speakers do not make more gender agreement errors with such nouns¹. Still, what does not show up in the number of errors may influence online sentence processing. Experiment 1 is dedicated to this question.

In this paper, we focus on consonant-final nouns. The distributional picture gets more complex if the nature of the final consonant is taken into account. The gender of the words ending in /ž/, /š/, /č/, /šč/ can be determined orthographically, while words ending in other non-palatalized consonant can be only masculine. No such clues are available for the words ending in other palatalized consonants, but, as Table 2 shows, their distribution between the two genders varies greatly (to represent both the relative frequencies of different forms and the number of existing lemmas we relied not only on the RNC, but also on the Grammatical Dictionary of the Russian Language [GDRL, Zaliznjak, 1987]).

The majority of nouns ending in labials are feminine. Most nouns ending in -r’ and -l’ are masculine, many of them have agentive suffixes like -tel’ or -ar’. Nevertheless, feminine nouns are not just singular cases in this group. Forms of nouns ending in other lingual consonants are either predominantly feminine or are evenly distributed between the two genders (but the number of feminine lemmas is still larger). The majority of nouns ending in -t’, which are especially numerous, have the suffix -ost’.

In our pilot experiment 2, we explored whether adult native speakers are sensitive to these differences.

¹ Cases where gender variation is observed are not taken into account (a corpus-based study of such cases was conducted by [Savchuk 2011]).
Table 1. The distribution of nouns among genders and declensions in the grammatically disambiguated subcorpus of the RNC

<table>
<thead>
<tr>
<th>Declension and gender</th>
<th>Percentage of nouns in the RNC</th>
<th>Ending in Nom.Sg and prototypicality</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st decl. feminine</td>
<td>29% nouns</td>
<td>end in -a/-ja, 'prototypical F'</td>
<td>zhena 'wife'</td>
</tr>
<tr>
<td>1st decl. masculine</td>
<td>1% nouns</td>
<td>end in -a/-ja, 'non-prototypical M'</td>
<td>djadja 'uncle'</td>
</tr>
<tr>
<td>2nd decl. masculine</td>
<td>46% nouns</td>
<td>end in a consonant, 'prototypical M'</td>
<td>syn 'son', gel 'gel'</td>
</tr>
<tr>
<td>2nd decl. neuter</td>
<td>18% nouns</td>
<td>end in -o/e, 'prototypical N'</td>
<td>pole 'field'</td>
</tr>
<tr>
<td>3rd decl. feminine</td>
<td>5% nouns</td>
<td>end in a consonant, 'non-prototypical F'</td>
<td>mel 'shallow'</td>
</tr>
<tr>
<td>irregular and indeclinable</td>
<td>1% nouns</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Nouns in the grammatically disambiguated subcorpus of the RNC and in the GDRL

<table>
<thead>
<tr>
<th>Final consonant</th>
<th>RNC (Nom.Sg forms)</th>
<th>GDRL (lemmas)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>/b/</td>
<td>34 (24%)</td>
<td>110</td>
</tr>
<tr>
<td>/p/</td>
<td>3 (2%)</td>
<td>169</td>
</tr>
<tr>
<td>/v/</td>
<td>13 (1%)</td>
<td>1,448</td>
</tr>
<tr>
<td>/f/</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>/m/</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>/d/</td>
<td>748 (51%)</td>
<td>707</td>
</tr>
<tr>
<td>/t/</td>
<td>713 (5%)</td>
<td>13,184</td>
</tr>
<tr>
<td>/z/</td>
<td>319 (49%)</td>
<td>327</td>
</tr>
<tr>
<td>/s/</td>
<td>80 (14%)</td>
<td>491</td>
</tr>
<tr>
<td>/n/</td>
<td>2,354 (45%)</td>
<td>2,842</td>
</tr>
<tr>
<td>/r/</td>
<td>2,160 (76%)</td>
<td>677</td>
</tr>
<tr>
<td>/l/</td>
<td>6,648 (71%)</td>
<td>2,653</td>
</tr>
</tbody>
</table>

2. Previous experimental studies

Two groups of experimental studies are relevant for the present paper: analyzing gender agreement and nouns with more or less morphologically regular inflections.

---

2 The counts are taken from [Slioussar and Samoilova 2015]. Substantivized adjectives were not taken into account.

3 These three forms are rup' (a reduced form of the noun rubl’ ‘ruble’).
There are relatively few experimental studies of gender agreement in Russian. In three of them ([Akhutina et al. 1999, 2001]; [Romanova & Gor 2017]) adjectives were presented before nouns audially or visually. In congruent conditions, adjectives agreed with the following nouns, in incongruent ones they did not, and some experiments also included a baseline condition where bare adjective stems without inflections or adverbs were presented. Several methods were employed, including lexical decision (answering whether the presented stimulus is a real word or a nonce word), grammaticality judgment (answering whether the presented fragment is grammatical) and cued-shadowing in which participants must repeat the second presented word (the target noun).

However, the question was always the same: would participants answer significantly faster and more accurately in congruent conditions compared to incongruent ones, and would there be any differences associated with the gender of the nouns? In experiments with a baseline condition, it was also possible to check whether the difference between congruent and incongruent conditions was primarily due to facilitation in the former, or to inhibition in the latter, or both effects were equally prominent. In brief, [Akhutina et al. 2001] observed significant facilitation and inhibition effects for feminine nouns, while for masculine nouns, only inhibition was significant, and for neuter ones, only facilitation was significant. Results from other studies were similar.

The explanations offered in these studies go along the same lines. Masculine gender as the most frequent is assumed to be unmarked, or default, while neuter is considered the most marked. Thus, masculine is expected by default, and strengthening this expectation by a masculine adjective does not produce a big difference (hence no significant facilitation effects). Neuter is the least expected option, so priming a neuter noun with a neuter adjective has the largest effect compared to the baseline condition (hence facilitation effects for neuter nouns are larger than for feminine nouns). Inhibition effects are explained by rechecking, which is especially costly for masculine nouns presented after non-masculine adjectives.

None of these three studies looked at 3rd declension feminine nouns, while the experiments by [Taraban and Kempe 1999] specifically focused on them. Taraban and Kempe selected masculine and feminine nouns ending in a palatalized consonant (opaque condition) and in non-palatalized consonants or in -a/ja, which are unambiguously masculine or feminine (transparent condition). They examined the role of such transparency for subject–predicate agreement using word-by-word self-paced reading and forced choice tasks. Participants were asked to read sentence beginnings like (1a) or (2a) and then to select one of the two verb forms in the remaining fragment like (1b) or (2b). In some conditions, sentence fragments contained adjectives. Participants were adult native speakers and L2 learners. For native speakers, transparency and the presence of a gender-marked adjective did not play any role.

(1) a. Daže even (obyčnaja) ordinaryF muka/sol' flourF/salt a now teper' now...
b. isčez/isčezla disappearedM/F iz from magazinov stores

This study also involved aphasic patients, while [Romanova and Gor 2017] compared native speakers to second language learners, but we will not discuss these groups here.
Slioussar N. A.

(2) a. Nakanune (otěkšij) palec/lokot’ sil’no…
the-day-before swollen finger strongly

b. bolel/bolela ot udara.
hurt from injury

[Slioussar and Malko 2016] studied gender agreement attraction. To give an example, an attraction error is present in the English sentence “The key to the cabinets are rusty”, where the verb agrees not with the head of the subject phrase, but with another noun, termed attractor. In production, such errors are more frequent than agreement errors without attraction. In comprehension, they are missed more often and produce smaller delays in reading times and less pronounced ERP responses.

Number agreement attraction is widely discussed in the literature, while gender agreement has been analyzed only in a few studies so far. Among other things, it was noted that both in production and in comprehension, attraction effects can be observed in the sentences with singular heads and plural dependent nouns (e.g., “The key to the cabinets...”), but not in the sentences with plural heads and singular dependent nouns (e.g., “The keys to the cabinet...”). Almost all proposed explanations appeal to feature markedness, although approaches to markedness may be very different, from representational to frequency-based. Looking for similar asymmetries in gender agreement attraction, several studies of Romance languages obtained controversial results (e.g. [Acuña-Fariña et al., 2014]; [Anton-Mendez et al., 2002]; [Martin et al., 2014]; [Vigliocco & Franck, 1999]). [Badecker and Kuminiak 2007] found that neuter behaves as unmarked in a series of production experiments on Slovak, in which neuter is the least frequent gender, but is used in impersonal sentences, like in Russian.

[Slioussar and Malko 2016] conducted one production and three comprehension experiments. The results of the former were similar to the Slovak study, while in the latter, masculine behaved differently from feminine and neuter. Namely, attraction was observed for all dependent noun genders, but only for neuter and feminine heads. In other words, masculine heads were significantly more resistant to attraction: readers detected agreement errors irrespective of possible attractors’ interference.

This result can be reconciled with the observations made in [Akhutina et al. 1999, 2001]; [Romanova & Gor 2017]. However, given that different patterns were observed for production and comprehension, we cannot explain them by a particular single property of gender features anymore. This reminds us that the notion of markedness usually invoked to explain all asymmetries between features is problematic because some studies rely on representational markedness (primarily counting the number of positive feature values in formal morphological models), the others consider the most frequent value to be the default etc. From the representational point of view, neuter is the unmarked gender in most accounts, while if we rely on frequency, masculine is. Maybe, these approaches should be seen as complementary, because different properties of features appear to be relevant in different experimental tasks.

---

5 It is traditionally assumed that the features of the dependent noun are crucial for attraction, but both this study and some other findings suggest that the features of the head might be more important. We will not discuss this problem here.
Now let us turn to the problem of inflections that are more or less typical for a particular gender. Many experimental studies used different tasks (for example, determining a noun's gender or selecting an article) to show that nouns with more typical inflections are associated with faster and more accurate answers. This was demonstrated for Italian, French, Hebrew, Bulgarian etc. (e.g. [Andonova et al., 2004]; [Bates et al., 1995]; [Gollan & Frost, 2001]; [Spalek et al., 2008]).

However, fewer studies investigated the role of this factor in sentence processing. To give an example, Caffarra et al. (2015) looked at Italian nouns with more and less typical endings presented in the same sentences. They were preceded by articles, which carried gender information. Nouns from the two groups elicited different ERP responses. [Franck et al. 2008] and [Vigliocco and Zilli 1999] demonstrated for Italian, Spanish, and French that heads with regular inflections are more resistant to agreement attraction. The same is true for number agreement attraction (e.g. [Bock & Eberhard, 1993]; [Vigliocco et al., 1995]).

As for Russian, only [Taraban and Kempe’s 1999] study discussed above addressed this problem. The authors found no differences between prototypical and non-prototypical subject nouns. However, the task they used required selecting a correctly agreeing verb form, which is a less immediate measure than simple reading times. For this reason, we decided to come back to this factor in the present study.

3. Experiment 1

This experiment was conducted together with Pavel Shilin, an MA student at Saint-Petersburg State University. The goal of the experiment was to find out whether subject-predicate gender agreement is processed differently depending on the declension and gender the subject noun belongs to. 33 native speakers of Russian aged 18–55 (12 women) took part in the experiment.

It is impossible to test all potentially interesting combinations of gender and declension in one experiment, so we selected the following three groups of inanimate nouns as subjects: masculine nouns of the 2nd declension (prototypical masculine, 2DM), feminine nouns of the 1st declension (prototypical feminine, 1DF), and feminine nouns of the 3rd declension (non-prototypical feminine, 3DF). The materials included 36 sets of target sentences in six conditions exemplified in (3a-f). All sentences had the same structure: a subject noun, byl/byla “was”, an adjective or participle, and a three-word PP.

(3) a. 2DM G(rammatical): Xalat byl potrepannym ot mnogoletnej noski. b. 2DM U(ngrammatical): Xalat byla potrepannoj... c. 1DF G(rammatical): Kurtka byla potrepannoj... d. 1DF U(ngrammatical): Kurtka byl potrepannym...
Three fourths of the sentences contained gender agreement errors on the predicate because taking previous studies of agreement into account (primarily agreement attraction experiments), the effects could be expected to be different in grammatical and ungrammatical sentences. Subject nouns in the three declension groups were balanced for frequency and length using the StimulStat lexical database (http://stimul.cognitivestudies.ru, [Alexeeva et al., 2018]). Frequency information in this database is taken from the Frequency Dictionary of Modern Russian Language [Lyashevskaya & Sharoff, 2009].

Target sentences were distributed into six experimental lists so that each participant saw only one sentence from each set. The lists also contained 80 grammatically correct filler sentences. The sentences were presented on a PC using Presentation software (http://www.neurobs.com). We used the word-by-word self-paced reading methodology. Each trial began with a sentence in which all words were masked with dashes while spaces and punctuation marks remained intact. Participants were pressing the space bar to reveal a word and re-mask the previous one. One third of the sentences were followed by forced choice comprehension questions to ensure that the participants were reading properly.

We analyzed participants’ question-answering accuracy and reading times. On average, participants answered 12% questions to target sentences incorrectly, no participants made more than 3 errors. Reading times that exceeded a threshold of 2.5 standard deviations, by region and condition, were excluded [Ratcliff, 1993]. In total, 2.0% of the data were excluded as outliers. Average reading times per region in different conditions are presented in Figure 1.

![Figure 1. Average reading times per region (in ms) in different experimental conditions](image)

---

6 We selected predicates that consisted of a copula and an adjective or participle because such predicates were used in the previous experimental studies of subject—predicate gender agreement in Russian.
For each region, we made pairwise comparisons between the three conditions using a $2 \times 2$ Repeated Measures ANOVA with grammaticality and declension as factors. Analyses by participants ($F_1$) and by items ($F_2$) were performed. In region 1 (the subject noun), there were no significant results, which means that nouns in different conditions were properly balanced and that nouns of a particular gender or declension are not intrinsically more difficult to process.

Region 2 contains the verb byl / byla ‘wasM/F’—this is where agreement errors appear in ungrammatical sentences. Figure 1 suggests that participants’ reaction to these errors was different depending on the declension of the subject noun. In the conditions 1DF and 2DM (with prototypical feminine and masculine subjects), reading times in ungrammatical sentences are longer than in grammatical ones, while no such difference can be seen in the 3DF conditions (with non-prototypical feminine subjects), which indicates that the error tends to remain undetected in the latter case.

Statistical analyses support this intuition. In the comparison between 1DF and 3DF conditions, grammaticality and the interaction between declension and grammaticality are significant ($F_1(1,32) = 8.13, p < 0.01, F_2(1,35) = 4.20, p = 0.05$; $F_1(1,32) = 7.41, p = 0.01, F_2(1,35) = 4.05, p = 0.05$), while the main effect of declension does not reach significance. This means that the influence of grammaticality depends on the declension of the subject. In the comparison between 2DM and 3DF conditions, grammaticality reaches significance, while the interaction between declension and grammaticality is marginally significant ($F_1(1,32) = 8.01, p < 0.01, F_2(1,35) = 4.09, p = 0.05; F_1(1,32) = 3.98, p = 0.05, F_2(1,35) = 3.17, p = 0.08$). When 1DF and 2DM are compared, only the grammaticality factor is significant ($F_1(1,32) = 18.66, p < 0.01, F_2(1,35) = 10.21, p < 0.01$).

In region 3 that contains an adjective or participle, differences between grammatical and ungrammatical sentences become visible in all conditions. Accordingly, the grammaticality factor is significant in all pairwise comparisons ($F_1(1,32) = 15.90, p < 0.01, F_2(1,35) = 21.24, p < 0.01$ for 1DF vs. 2DM; $F_1(1,32) = 11.98, p < 0.01, F_2(1,35) = 6.20, p = 0.02$ for 1DF vs. 3DF; $F_1(1,32) = 9.73, p < 0.01, F_2(1,35) = 7.83, p < 0.01$ for 2DM vs. 3DF). No other factors or interactions reach significance.

Regions 4–6 contain a three-word PP. In region 4, a tendency that can be already detected in region 3 becomes statistically significant: the error-related delay in reading times is more pronounced in the 2DM conditions (with masculine subjects) than in the 1DF and 3DF conditions (with feminine subjects). In the comparison between 1DF and 2DM conditions, grammaticality and the interaction between declension and grammaticality are significant ($F_1(1,32) = 36.95, p < 0.01, F_2(1,35) = 15.91, p < 0.01$; $F_1(1,32) = 9.77, p < 0.01, F_2(1,35) = 6.45, p = 0.02$), while declension is not significant. The same is true for the comparison between 3DF and 2DM ($F_1(1,32) = 50.11, p < 0.01, F_2(1,35) = 13.17, p < 0.01; F_1(1,32) = 11.38, p < 0.01, F_2(1,35) = 5.51, p = 0.03$). When 1DF and 3DF are compared, only the grammaticality factor is significant ($F_1(1,32) = 12.34, p < 0.01, F_2(1,35) = 3.65, p = 0.07$).

In region 5, only the grammaticality factor is significant in all pairwise comparisons ($F_1(1,32) = 18.51, p < 0.01, F_2(1,35) = 17.67, p < 0.01$ for 1DF vs. 2DM; $F_1(1,32) = 14.78, p < 0.01, F_2(1,35) = 6.10, p = 0.02$ for 1DF vs. 3DF; $F_1(1,32) = 18.07, p < 0.01, F_2(1,35) = 10.07, p < 0.01$ for 2DM vs. 3DF). In region 6, there are no significant differences.
Finally, let us note that when we planned the experiment, we did not consider assessing the role of the final consonant of 3DF nouns. But 10 out of 36 nouns we selected ended in -l' or -r', which is more characteristic for masculine, while other nouns had final consonants characteristic for feminine. So the role of this factor could be estimated, and there were no hints of any relevant differences.

4. Pilot experiment 2

This pilot experiment was included in a study we conducted together with Varvara Magomedova (SUNY, Stony Brook) and Natalia Chuprasova, an MA student at Saint-Petersburg State University. The main goal of the study was to find out how Russian speakers determine the gender of real and nonce nouns with diminutive and augmentative suffixes. However, to make the materials more diverse, other nouns had to be included, and we selected 12 real and 12 nonce nouns ending in palatalized consonants (as well as some indeclinable nouns etc.).

Participants were 30 native Russian speakers (17 women), aged 19–30. They received a list of seven adjectives and then were presented with nouns one by one. They were asked to pick a matching adjective and pronounce the resulting phrase. Adjectives had meanings like 'big', 'small', 'cool', 'bad' etc., to make participants think that the experiment was about semantic connotations of different nouns.

Analyzing the gender of the adjectives selected by the participants, for 12 real nouns ending in palatalized consonants (6 masculine and 6 feminine) we found only 7 errors out of 360 responses. There were three errors with žen'sen' ‘ginsengM’, two errors with stupen’ ‘stepF’, and one error with kisel’ ‘starch drinkM’ and with prorub’ ‘ice-holeF’. The low number of errors agrees with the previous findings by Rusakova (2013): adult native speakers of Russian do not experience particular difficulties determining the gender of such nouns.

As for 12 nonce nouns, we had two examples with each of the following endings: -b', -d', -s', -n', -l' and -r'. The number of answers with masculine adjectives in these groups was 22 (out of 60), 26, 38, 30, 44 and 51, respectively. We can see that the nature of the final consonant played a role. To estimate it statistically, we used mixed effects logistic regression with random slopes and random intercepts by participants and by items. Only nonce nouns ending in -l' and -r' were significantly different from the other groups ($\beta = 2.48$, SE = 1.15, $z = -2.15$, $p = 0.03$; $\beta = 4.00$, SE = 1.54, $z = -2.60$, $p < 0.01$). Unlike all other palatalized consonants, these final consonants are more characteristic for masculine nouns.

5. Conclusions

In the introduction, we outlined three distributional properties of Russian nouns. Firstly, masculine gender is more frequent than feminine and neuter. Secondly, some combinations of genders and declensions are more frequent than the others (we called them prototypical). Thirdly, for nouns ending in palatalized consonants, some consonants are more characteristic for masculine nouns and the others for feminine nouns.
Previous studies indicate that these factors do not increase the number of naturally occurring gender agreement errors for adult native speakers of Russian [Rusakova 2013]. But speakers might still be sensitive to them, and in the present paper we demonstrated this in two experiments.

Experiment 1 revealed the role of the two first factors. It was not designed to assess the role of the third factor, but, as far as we could estimate from the data, this factor does not play a role in online agreement processing. However, the influence of this factor can be detected when speakers try to guess the gender of a novel noun—like in the pilot experiment 2. In further studies, we plan to use more stimuli, testing speakers’ sensitivity to different suffixes etc.

Now let us discuss the results of experiment 1 in more detail. It demonstrated that both gender and declension of the noun influence online processing of the subject–predicate gender agreement in Russian. But, firstly, this influence can be detected only in the sentences with agreement errors, i.e. no gender or declension is intrinsically more difficult to process (at least, in the sentence context7). Secondly, declension plays a role at a very early stage and its effect is very short-lived, while the role of gender becomes visible later and its effect is more pronounced.

The fact that a masculine verb form is less readily detected after a 3rd declension subject noun can be explained by the fact that its ending is more typical for masculine nouns that for feminine ones. However, alternative explanations are also possible, for example, all agreement errors (in masculine or in neuter) may be harder to detect after 3rd declension subject nouns, i.e. their gender can be in general harder to retrieve. To exclude this and some other possibilities, other experiments should be conducted. Another line of further research should look at non-prototypical masculine nouns like papa ‘dad’. The picture may be different not only because of their different gender, but also because all these nouns denote humans, so the gender feature is not semantically empty in this case, which may aid its processing and retrieval.

As for the role of gender as such, we saw that agreement errors with masculine subjects cause a larger delay in reading times compared to errors with feminine subjects, i.e. were costlier for processing. This is in line with the previous findings on gender agreement in comprehension reported in the literature [Akhtina et al., 1999, 2001]; [Romanova & Gor, 2017]; [Sloussar & Malko, 2016]. However, to have a full picture, neuter subjects and predicates should be introduced in further experiments.

So far, several explanations are possible. It is well known that while reading, we generate expectations about the upcoming predicate based on the features of the subject and rechecking is prompted if these expectations are violated (which is associated with increased reading times). Perhaps, the masculine form of the predicate, being the most frequent, causes less disruption if used incorrectly—similarly, using a frequent word instead of an infrequent one provokes less surprise than the opposite mistake. Maybe, these expectations are more robust for masculine subjects, so violating them is more disruptive. Maybe, if an agreement error is detected and rechecking is initiated, masculine subjects are retrieved more readily and reliably—this is what

7 It is well known that many differences that can be detected in the processing of isolated forms disappear when these forms are embedded in an appropriate context.
Slioussar N. A.

[Slioussar and Malko 2016] suggested based on their agreement attraction results where all combinations of genders on subjects, attractors and predicates were examined. All these explanations are compatible with the observed difference between ungrammatical sentences with masculine and feminine subjects. Further experiments are necessary to tease them apart and to gain a better understanding of the patterns observed in previous studies.

The study was supported in part by the grant #16-18-02071 from the Russian Science Foundation.

References

Appendix F.

Paper “Processing of a free word order language: The role of syntax and discourse context”


**Abstract**

In languages with flexible constituent order (so-called *free word order languages*), available orders are used to encode given/new distinctions; they therefore differ not only syntactically, but also in their context requirements. In Experiment 1, using a self-paced reading task, we compared Russian S V IO DO (canonical), DO S V IO and DO IO V S constructions in appropriate vs. inappropriate contexts (those that violated their context requirements). The context factor was significant, while the syntax factor was not. The less pronounced context effect evidenced in previous studies (e.g., Kaiser and Trueswell in *Cognition* 94:113–147, 2004) might be due to the use of shorter target sentences and less extensive contexts. We also demonstrated that the slow-down starts at the first contextually inappropriate constituent, which shows that the information about context requirements is taken into account immediately, but that it develops faster on preverbal subjects and postverbal indirect objects (occupying their canonical positions) than on preverbal indirect objects (occupying a noncanonical position, or *scrambled*). In Experiment 2, these findings were replicated for IO S V DO and IO DO V S orders. S V IO DO orders with a continuation were used to show that there is no additional effect of inappropriate context at the end of the sentence.

**The link on the publisher’s website**

This is not an open access paper, so according to the publisher’s policy it cannot be reproduced. It can be accessed on the publisher’s website:

Appendix G.

Paper “StimulStat: a lexical database for Russian”


Abstract

In this article, we present *StimulStat* – a lexical database for the Russian language in the form of a web application. The database contains more than 52,000 of the most frequent Russian lemmas and more than 1.7 million word forms derived from them. These lemmas and forms are characterized according to more than 70 properties that were demonstrated to be relevant for psycholinguistic research, including frequency, length, phonological and grammatical properties, orthographic and phonological neighborhood frequency and size, grammatical ambiguity, homonymy and polysemy. Some properties were retrieved from various dictionaries and are presented collectively in a searchable form for the first time, the others were computed specifically for the database. The database can be accessed freely at [http://stimul.cognitivestudies.ru](http://stimul.cognitivestudies.ru).

The link on the publisher’s website

This is not an open access paper, so according to the publisher’s policy it cannot be reproduced. It can be accessed on the publisher’s website:

[https://link.springer.com/article/10.3758/s13428-017-0994-3](https://link.springer.com/article/10.3758/s13428-017-0994-3)