ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 13 за 2017 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Клышинский Э.С., Логачева В.К.,
**Белобокова Ю.А.**

# Понимаемость текста на иностранном языке: случай славянских языков

***Рекомендуемая форма библиографической ссылки:*** Клышинский Э.С., Логачева В.К., Белобокова Ю.А. Понимаемость текста на иностранном языке: случай славянских языков // Препринты ИПМ им. М.В.Келдыша. 2017. № 13. 23 с. doi:10.20948/prepr-2017-13
URL: http://library.keldysh.ru/preprint.asp?id=2017-13

**О р д е н а  Л е н и н а**
**ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ**
**имени М.В.Келдыша**
**Р о с с и й с к о й  а к а д е м и и  н а у к**

**Э.С. Клышинский, В.К. Логачева, Ю.А.Белобокова**

# Понимаемость текста на иностранном языке: случай славянских языков

**Москва — 2017**

**Клышинский Э.С., Логачева В.К., Белобокова Ю.А.**

**Понимаемость текста на иностранном языке: случай славянских языков**

В работе рассматривается вопрос понятности текста на иностранном языке для русскоязычного читателя, не знакомого с данным языком. Целью работы было получить численную оценку процента слов, которые читатель может понять в тексте на незнакомом иностранном языке, родственном языку читателя. Была проведена серия экспериментов, где читателю предлагалось восстановить текст с пропусками на русском языке, имея в качестве подсказки его перевод на иностранный язык (украинский, белорусский, болгарский, польский или чешский). В качестве контроля использовался текст на эстонском языке, имеющем множество структурных и лексических различий с русским языком и непонятном не учившим его носителям русского языка. По результатам экспериментов мы выяснили, что понятность иностранного текста определяется процентом слов, созвучных русским и имеющих сходное или то же значение. Улучшения за счет несозвучных слов значительно меньше и при исключении прочих факторов стремится к нулю.

***Ключевые слова:*** понятность текста, славянские языки, L2 языковая сложность.

**Klyshinsky E.S., Logacheva V.K., Belobokova Yu.A.**

**Foreign text intelligibility: case of Slavic language group**

The research concentrates on the intelligibility of languages for a speaker of a related language. Our hypothesis is that people can partially understand text in a language related to their native language, and the closer the relation, the more they can understand. In order to check that we conducted a series of experiments where we asked native speakers of Russian to fill the gaps in a Russian text using its translation into another Slavic language (Ukrainian, Belorussian, Bulgarian, Polish, or Czech). As a baseline we took the Estonian language, which is very different from Russian in many aspects and cannot be understood by a Russian speaker without training.

The results show that a native speaker of Russian performs the gap filling task significantly better when provided with the translation of the text into a Slavic language.

The research is partially supported by RFH grant №15-04-12019.

***Key words:*** text intelligibility, Slavic languages, L2 language complexity

Работа выполнена при поддержке РГНФ, грант № 15-04-12019.

# 1. Введение

В ходе изучения особенностей различных языков исследователи пришли к понятию языковой сложности (см., например, [Kusters, 2003]). На данный момент различают несколько параметров, по которым может оцениваться языковая сложность. Это, например, сложность языковых структур, выражающаяся в их многообразии; стоимость выражения или восприятия некоторой мысли средствами языка; трудность изучения языка для конкретного человека. Эти и другие категории описаны в работах [Miestamo, 2008], [Даль, 2009], [Бердичевский, 2012] и других. При этом все авторы сходятся в том, что следует различать сложность изучения родного языка (сложность L1) и сложность изучения иностранного языка (сложность L2). Очевидно, что сложность L2 существенно зависит от многих факторов: сходства родного и изучаемого языка, наличия в обществе таких явлений, как диглоссия и двуязычие[1], индивидуальных особенностей изучающего.

В целом методы исследования в области языковой сложности можно разделить на две группы. Во-первых, это методы сравнительного языкознания, в которых сравнение нескольких языков проводится по выбранному набору параметров. Сравнение ведется путем анализа ограниченного множества примеров, извлеченных из текстовых корпусов, или набора характеристик, взятых из существующих лингвистических ресурсов (например, World Atlas of Language Structures – WALS, http://wals.info/). Сравнение является скорее качественным, чем количественным, так как в его задачи входит обнаружение новых явлений, а не их численное описание. Второй группой методов являются количественные методы оценки языковой сложности. Самым ярким представителем данной группы является оценка Колмогоровской сложности системы[2], однако применяются и некоторые другие методы.

Обе группы методов обладают рядом недостатков. Так, в ходе сравнения языков авторы обычно либо изучают частные лингвистические вопросы для нескольких (от двух до пяти) языков, либо анализируют большое количество языков, но для ограниченного количества параметров. Например, в работе [Lupyan, 2009] авторы исследовали три характеристики для 2236 языков: численность населения, занимаемая этим населением площадь и количество языков-соседей. Авторами было показано, что чем больше значения признаков, тем ниже сложность языка. Однако как методика оценки сложности, так и причины появления такой зависимости остаются туманны.

---

[1] Под двуязычием здесь мы понимаем попеременное использование двух языков в обществе, как, например, смешивание носителей языков на территории Швейцарии, выделяется также параллельное использование двух языков в различных сферах – диглоссия, например, древнерусского в быту и церковнославянского для богослужений на Руси в IX — XII вв.

[2] Согласно Колмогорову [Колмогоров, 1965], сложность системы определяется длиной программы, порождающей описание данной системы.

Исследование количественных характеристик также зачастую приводит к неоднозначным результатам. Так, использование коэффициента архивации в качестве аппроксимации Колмогоровской сложности текста упирается в тот факт, что параметры подобной аппроксимации неясны. Коэффициент архивации скорее показывает частоту повторов определенного набора букв. Данная характеристика хотя и связана с богатством лексики или сложностью синтаксиса языка, но слишком неявно. Помимо этого, делать выводы о языке в целом после изучения нескольких текстов ограниченного числа авторов было бы слишком смело.

В связи с этим можно утверждать, что, несмотря на отточенность лингвистических методов исследования, изучение языковой сложности нуждается в новых и более точных числовых методах ее оценки.

В рамках сложности L2 рассматриваются такие вопросы, как понятность родственных иностранных языков. Здесь исследования уже проводились, например, для скандинавских [Gooskens, 2007], романских [Ciobanu, 2014] и славянских [Lindsay, 2014] языков. Однако, как и в других исследованиях, в данных работах очевиден недостаток количественных данных. В связи с этим в данной работе мы будем не просто исследовать вопрос сходства языков, но проведем количественную оценку подобного сходства.

Наше исследование посвящено достаточно узкому вопросу: степени понятности иностранных славянских языков (белорусского, украинского, болгарского, польского и чешского) для русскоязычного читателя, не знакомого с ними. Чтобы узнать, насколько понятны иностранные тексты русскоязычному читателю, мы просили информантов заполнить пропуски в русском тексте, пользуясь переводом этого текста на иностранный язык. Этот тест позволил определить процент слов, которые могут быть частично поняты без перевода в тексте на языке, не знакомом информанту, но родственном его родному языку. Чтобы подтвердить, что информанты восстанавливали пропущенные слова, пользуясь иностранной версией текста, а не только контекстом самих слов, мы провели два типа контрольных экспериментов: (1) просили участников восстановить пропуски, не давая им доступа к переводу на родственный язык, (2) просили участников восстановить пропуски, предоставив перевод на эстонский язык. В первом случае мы могли сравнить влияние контекста с влиянием перевода, во втором случае – убедиться, что перевод на язык, далекий от родного языка читателя, не дает ему дополнительной информации о тексте. Проведенные тесты показали, что наличие текста на незнакомом иностранном языке улучшает результаты на 8-31% в зависимости языка. Помимо этого, наблюдается зависимость прироста от процента созвучных слов в иностранном тексте.

# 2. Метод определения понятности иностранного текста

Для проведения экспериментов мы использовали тесты на заполнение пропусков в тексте. В данном виде тестов из текста вычеркиваются некоторые слова или части слов (например, каждое пятое слово, слова на заданную тему или окончания слов). Задачей информанта является восстановить эти слова по контексту. Данная методика часто используется при изучении иностранного языка для проверки уровня знаний студентов или расширения их словарного запаса [Zarei, 2013] или в психологии [Ackerman, 2000]. Кроме того, в работе [Ageeva, 2014] было предложено использовать этот метод для оценки качества машинного перевода. Для того чтобы определить, правильно ли переведен текст, исследователи вычеркивали некоторые слова из текста перевода и просили информантов (носителей языка перевода) восстановить полный текст, пользуясь по возможности исходным текстом. Качество машинного перевода определялось в процентах слов, корректно восстановленных информантами.

## 2.1 Методика проведения теста

В своей работе мы использовали этот метод для определения понятности текста на родственном иностранном языке для русскоязычного читателя, не знакомого с данным языком. Вместо автоматически переведенного текста использовался перевод, сделанный человеком.

В предыдущих исследованиях было показано [Ягунова, 2010], что восстановление слов по тексту возможно за счет его избыточности. При наличии только перевода информант восстанавливает слова, используя избыточность текста, дистрибутивные свойства слов и свой личный опыт. Из контекста могут быть восстановлены часть речи, грамматические характеристики и синтаксические связи пропущенного слова. Например, в предложении «Человек зашел в ____ .» последнее слово не может быть с уверенностью восстановлено без дополнительной информации, однако его контекст (предшествующий предлог «в» и глагол «заходить») дает понять, что пропущенное слово — неодушевленное существительное в винительном падеже. Восстановив часть речи пропущенного слова, информант может угадать само слово, пользуясь лексической сочетаемостью контекста. Например, в предложении «Поезд приехал на ____.» последнее слово может принимать значения из ограниченного списка: например, вокзал, полустанок, платформа. Заметим, что вписываемое слово во многом зависит от личного опыта информанта, его внутренних ассоциаций. Подобный «шум» должен сглаживаться при наборе значимого числа участников эксперимента. При наличии нескольких вариантов для вписываемого слова информант использует грамматические свойства слов. Так, если восстанавливается именная группа, то входящие в нее слова согласуются с пропущенным существительным в роде, что помогает вспомнить нужное слово.

Итак, пропущенные слова могут быть восстановлены из русского текста без дополнительных подсказок. В данной работе мы ставили задачу определить, в какой степени наличие частично понимаемого текста на языке оригинала служит подсказкой информантам, то есть увеличивает процент корректно восстановленных слов. Для этого нам надо отделить влияние контекста от влияния показанного пользователю оригинального текста. Мерой понятности иностранного языка в этом случае будет количество слов, которые не могли быть восстановлены из контекста, но были восстановлены с помощью оригинального текста на иностранном языке. Поэтому для каждого подготовленного текста мы проводили две серии экспериментов: в первой информанту был показан русский текст с пропусками и его оригинал на иностранном языке (или в транскрипции на кириллицу), во второй — только русский текст при помощи которого осуществляется контроль за избыточностью текста. Согласно нашей гипотезе, тесты, содержащие в себе как исходный иностранный текст, так и его перевод на русский язык, должны быть выполнены с меньшим количеством ошибок.

Мы использовали следующую систему оценки ответов информантов:

- Если информант вписывает именно то слово, которое имеется в нашем переводе, или один из его синонимов, ответ считается **корректным**[3].
- Если информант вписывает другое слово той же части речи, ответ считается **частично корректным**, так как информант понял структуру текста, но не смог правильно восстановить пропущенное слово.
- Если информант ошибается даже с частью речи слова или не вписывает ничего, ответ считается полностью **некорректным.**

Эксперименты проводились в два этапа. На первом этапе для выбранных языков было создано по два текста с пропусками. Все тексты были написаны различными авторами. Для всех произведений иностранный язык был языком оригинала, то есть они были написаны на иностранном языке, а затем переведены на русский.

Однако проведение экспериментов на разных текстах обладает существенным недостатком: для проведения экспериментов мы пропускаем различные слова. Однако нам неизвестна разница между простотой восстановления различных слов, то есть разница между текстами на разных языках может быть объяснена как языковыми различиями, так и различиями в выбранных текстах. В связи с этим нам следует зафиксировать как текст, так и список пропущенных слов, оставив его неизменным для всех языков. В этом случае можно будет утверждать, что мы исследуем только различия в понимании выбранных языков.

По указанным причинам на втором этапе мы проводили эксперименты для одного и того же текста, переведенного на выбранные языки. В проведении контрольных экспериментов с эстонским языком здесь уже нет необходимости,

[3] Грамматические параметры при этом могут не совпадать.

так как результаты первого этапа должны быть достаточно показательны. Помимо этого, контрольный тест, содержащий в себе текст только на русском языке, не будет создаваться заново для каждого языка, а используется только один раз.

### 2.2 Подготовка данных и выбор языков для исследования

В качестве исходных языков были выбраны **белорусский, украинский, болгарский, польский** и **чешский.** Этот выбор обусловлен классификацией языков, принятой в сравнительном языкознании. Славянские языки делятся на три подгруппы: восточнославянские, западнославянские и южнославянские. Языки, входящие в одну подгруппу, характеризуются бóльшим количеством общей лексики и общих структурных характеристик, чем языки из разных подгрупп. Желая проверить, насколько эти факторы влияют на понимаемость иностранного языка, мы выбрали для эксперимента два языка, входящих, как и русский, в восточнославянскую подгруппу – белорусский и украинский, и два языка с более далеким родством – чешский и польский языки из западнославянской подгруппы, и язык, занимающий промежуточное положение и относящийся к южнославянской группе, – болгарский.

Для контроля работы методики мы провели еще один тест для эстонского языка – языка, который принадлежит к финно-угорской семье, не имеет родства с русским и далек от него лексически и структурно. Предполагалось, что при восстановлении этого текста не будет значимой разницы между восстановлением текста с использованием оригинала и без него. Если бы для эстонского языка были получены такие же результаты, что и для славянских языков, это означало бы, что предложенная методика неверна.

Подготовка текстов проводилась следующим образом.

1. Выбирался отрывок из художественного произведения, написанного на одном из исследуемых языков. Отрывок должен представлять собой несколько абзацев связного текста (230-260 слов).

2. Каждому предложению или крупному фрагменту предложения (15-20 слов) был поставлен в соответствие его перевод на русский язык.

3. Из русской версии текста вычеркивалось 30-35 слов (примерно каждое восьмое слово текста). Вычеркивались только знаменательные слова: существительные, прилагательные, глаголы, реже наречия и местоимения.

На первом этапе использовались тексты на белорусском, украинском, польском и чешском языках, а также контрольные эксперименты для эстонского языка. Для всех языков, кроме польского, выбирались непрерывные фрагменты из одного и того же произведения, делившиеся между тестами на две части; для польского тесты были взяты из разных произведений, чтобы проверить различия в стиле текста. Удаляемые слова выбирались экспериментаторами, количество удаляемых слов зависело от текста.

Для второго этапа был выбран фрагмент книги Сенкевича «Quo vadis» («Камо грядеши»). Выбор произведения определялся доступностью текста для

всех выбранных языков (белорусский, украинский, польский, чешский и, дополнительно, болгарский). Выбор фрагмента определялся отсутствием специфичной лексики описываемой эпохи, с которой читатели могли быть не знакомы, а также дословным сходством переводов для всех выбранных языков. На втором этапе мы провели две серии экспериментов. В первой серии вычеркивалось каждое пятое значимое слово, чтобы избежать неявного произвола экспериментатора при их выборе. Всего было вычеркнуто 31 слово. Во второй серии слова вычеркивались экспериментатором с учетом их повторений в тексте, баланса понятности вычеркнутых слов и их семантической нагруженности. Всего было вычеркнуто 35 слов.

Помимо части речи мы обращали внимание на фонетическое сходство слов иностранного и русского языков. Предполагалось, что слова, сходно звучащие в обоих языках, будут вписываться информантами более корректно. С другой стороны, «ложные друзья переводчика»[4] должны показывать больший процент ошибок. Таким образом, все слова следует разделить на три категории. В первую категорию входят **созвучные слова**, имеющие одинаковый корень в иностранном и русском языках, например, польские *drgnęła* ([*дргнэнла*])[5] и *domem* ([*домем*]) vs. русские *вздрогнула* и *домом* соответственно. Сюда же были отнесены слова, однокоренные с синонимичным словом, например, польское *myśleć* ([*мышлеч*]) vs. русское *думать, мыслить*. Во вторую категорию были отнесены **несозвучные слова**, не имеющие общих корней или синонимов. И, наконец, в третью категорию были отнесены так называемые «**ложные друзья переводчика**», обладающие сходным корнем с русским словом, но имеющие другое значение, например, белорусское *час* или польское *czas* vs. русское *время*.

В эксперименты с языками различных групп вмешивается еще один фактор – в языках восточнославянской группы принята запись кириллицей, тогда как западнославянские языки используют латиницу. Несмотря на различия алфавитов, слова могут сохранять фонетическое сходство: информант может не узнать слово в его польской или чешской записи, но при этом догадаться о его значении по звучанию. В связи с этим для польского и чешского было подготовлено два вида текстов: русский текст ставился в соответствие оригинальному тексту или его транскрипции на русский язык (кириллической записи, по возможности сохраняющей произношение на языке оригинала). Транскрипция текста предъявлялась для того, чтобы проверить, является ли для информантов более важным фонетический облик слова или его исходное написание. Эксперименты с транскрибированными текстами проводились только на первом этапе.

---

[4] То есть слова иностранного языка, имеющие фонетическое сходство со словами русского языка, не являющимися их переводами, ср. запомнить – zapomnieć (*польск.* забыть).

[5] Здесь и далее запись в квадратных скобках означает практическую транскрипцию иностранного слова на русский язык, иными словами, запись иностранного слова буквами русского алфавита в соответствии с русской орфографией.

## 3. Используемые инструменты и данные

Для автоматизации процесса проверки мы разработали автоматизированную систему тестирования и проверки результатов. Результаты тестов хранятся в базе данных под управлением MySQL. По адресам http://cosyco.ru/under/ и http://cosyco.ru/under2/ была размещена программа, выбирающая тест для информанта случайным образом, соблюдая равномерность распределения ответов по разным тестам.

Тест представляет собой html-страницу с полями ввода на месте пропущенных слов (рис. 1). Пользователю предлагается ввести слова в пропуски и отправить результаты на сервер, где введенная информация сохраняется в базу данных. Из базы данных информация выгружается в csv-файл, который автоматически проверяется с использованием морфологического анализатора «Кросслейтор» [Елкин, 2003] и с использованием списка синонимов, введенных организаторами экспериментов. Каждому слову приписывается одна из трех оценок, введенных выше: корректно, частично корректно, полностью некорректно. Далее информация просматривается организаторами. Это необходимо в связи с тем, что система не исправляет орфографические ошибки, допущенные пользователями, а список синонимов может быть неполным.

| | |
|---|---|
| Рівно о чотирнадцятій він ішов від Історичного музею по тротуару вздовж високої гратниці Александровського саду. | Ровно в четырнадцать он шел от Исторического музея по тротуару вдоль высокой железной [ ] Александровского сада. |
| З одного боку був величний спокій Кремля, з другого - гриміла тисячами машин Москва, попереду - Борис уже бачив її - самотньо стояла на зупинці Тая, в білому платті, тонка, витка, мов дівчинка; здавалося йому, що сміється вона, хоч обличчя її ще й не розрізняв, та коли підійшов ближче, переконався: справді, посміхається. | С одной стороны, было [ ] спокойствие Кремля, с другой - [ ] тысячами машин Москва, [ ] - Борис уже [ ] ее - одиноко стояла на [ ] Тая, в белом платье, тонкая, [ ], словно девочка; казалось ему, что смеется она, хотя лица ее еще не [ ], но, когда подошел ближе, убедился: в самом деле, улыбается. |

*Рис. 1. Пример теста с параллельным текстом*

Тесты создавались вручную, однако в дальнейшем данный этап может быть автоматизирован. В любом случае, последнее слово в создании теста должно оставаться за организатором эксперимента, так как требуется контроль сложности восстановления пропущенных слов, их баланса, проверки созвучности. В последнем вопросе рекомендуется введение системы транскрипции, так как она может дать количественную характеристику фонетического сходства.

В качестве информантов выступали лица, родным языком[6] для которых был русский. Кроме того, они не должны были быть знакомы с другими славянскими языками. В качестве информантов выступали в основном студенты и преподаватели университетов из Москвы и Санкт-Петербурга, школьники 9-10 классов. Помимо них было небольшое количество взрослых

[6] Или хотя бы одним из родных языков.

информантов, не связанных с университетами. Большинство информантов получили лишь один тест, однако несколько добровольцев прошли все составленные тесты в одном из вариантов.

Таблицы 1 и 2 отражают количество информантов, принявших участие в тестировании на момент написания препринта.

*Таблица 1. Число информантов, принявших участие в тестировании на первом этапе (тест1 / тест2)*

| **Язык** | **Оригинал+ перевод** | **Транскрипция+ перевод** | **Только перевод** | **Итого** |
|---|---|---|---|---|
| Белорусский | 16/16 | - | 11/13 | 56 |
| Украинский | 14/12 | - | 13/10 | 49 |
| Польский | 16/12 | 14/17 | 18/11 | 88 |
| Чешский | 15/16 | 16/14 | 14/14 | 89 |
| Эстонский | 17/ - | 17/ - | 18/ - | 52 |
| Итого | | | | 334 |

*Таблица 2. Число информантов, принявших участие в тестировании на втором этапе (серия 1/серия2)*

| **Язык** | **Оригинал + перевод** | **Только перевод** |
|---|---|---|
| Белорусский | 17/17 | |
| Украинский | 17/17 | |
| Болгарский | 18/15 | |
| Польский | 14/17 | |
| Чешский | 15/14 | |
| Русский | | 19/19 |
| Итого | 161 | 38 |

## 4. Результаты тестирования на различных текстах и их статистическая оценка

Вписанные информантами слова были оценены в соответствии с разработанной шкалой (0 – полностью корректны; 1 – частично корректны, совпадение части речи; 2 – полностью некорректны: не совпадает часть речи или нет ответа). Для каждого слова в тесте была рассчитана доля правильных ответов, данных информантами, доля частично корректных ответов и доля полностью некорректных ответов. Каждое слово было помечено как созвучное, несозвучное или «ложный друг переводчика».

Далее были рассчитаны квартили полученного распределения долей правильных ответов внутри каждого теста. Полученные результаты для всех слов каждого теста показаны на рис. 2а в форме диаграммы размаха. На диаграмме показаны только второй и третий квартили, так как минимальное значение почти во всех тестах равнялось 0, а максимальное – 1. Как видно из рисунка, распределение долей корректных ответов отличается от нормального. Также из рисунка видно, что второй и третий квартили пересекаются в разных

вариантах одного теста. При этом среднее значение доли правильных ответов в тестах, где информантам предъявлялся только текст на русском языке, всегда меньше на 10-15%. Исключение составляет эстонский язык, в котором доля правильных ответов выше в тесте без оригинала.

На рисунках 2б и 2в показаны диаграммы размаха, полученные только для созвучных и несозвучных слов. Как видно из рис. 2б, для созвучных слов в тестах с исходным текстом и без него квартили практически не пересекаются. Для несозвучных слов картина сходна с картиной для текста в целом –

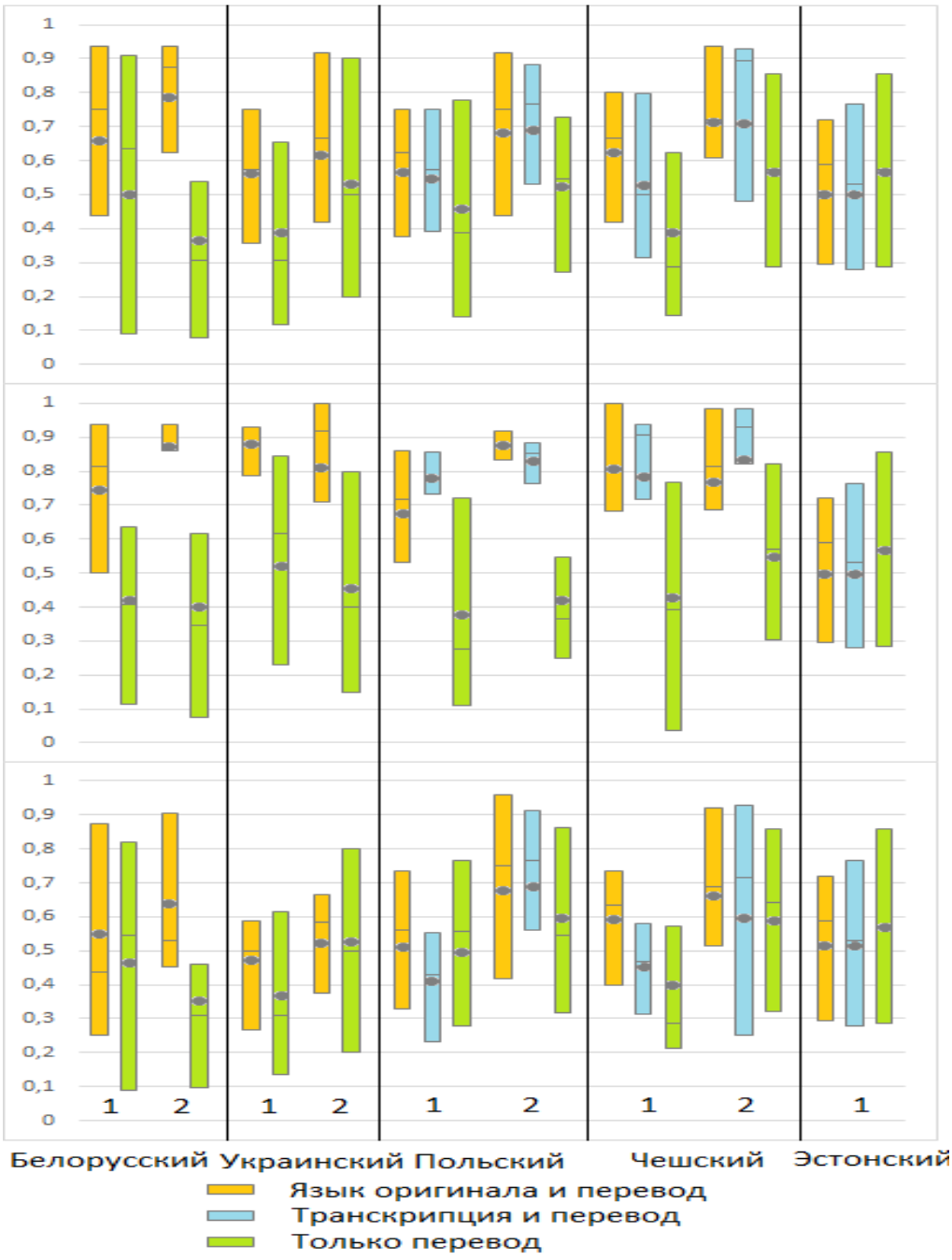



*Рис. 2. Диаграмма размаха правильных ответов по тестам: все слова (а), созвучные (б) и несозвучные (в) слова; цифра указывает на номер эксперимента*

квартили пересекаются, а среднее значение чаще всего уменьшается при отсутствии иностранного текста.

Мы рассчитали доверительные интервалы для каждого из тестов в целом и отдельно для созвучных и несозвучных слов. Интервалы рассчитывались по формуле $M \pm \alpha * se$, где $M$ – среднее значение, $se = \frac{s^2}{\sqrt{n}}$ – стандартная ошибка, $s^2$ – дисперсия, а $n$ – количество слов в тесте. Здесь $\alpha$ выбирается исходя из необходимого уровня достоверности (в нашем случае равнялось 1,96). Результаты приведены на рисунках 3а-в.

Как видно из рис. 3, ситуация повторяется – доверительные интервалы для созвучных слов практически не пересекаются, тогда как для теста в целом и для несозвучных слов наблюдается лишь падение среднего значения при отсутствии параллельного оригинала или его транскрипции. Здесь также

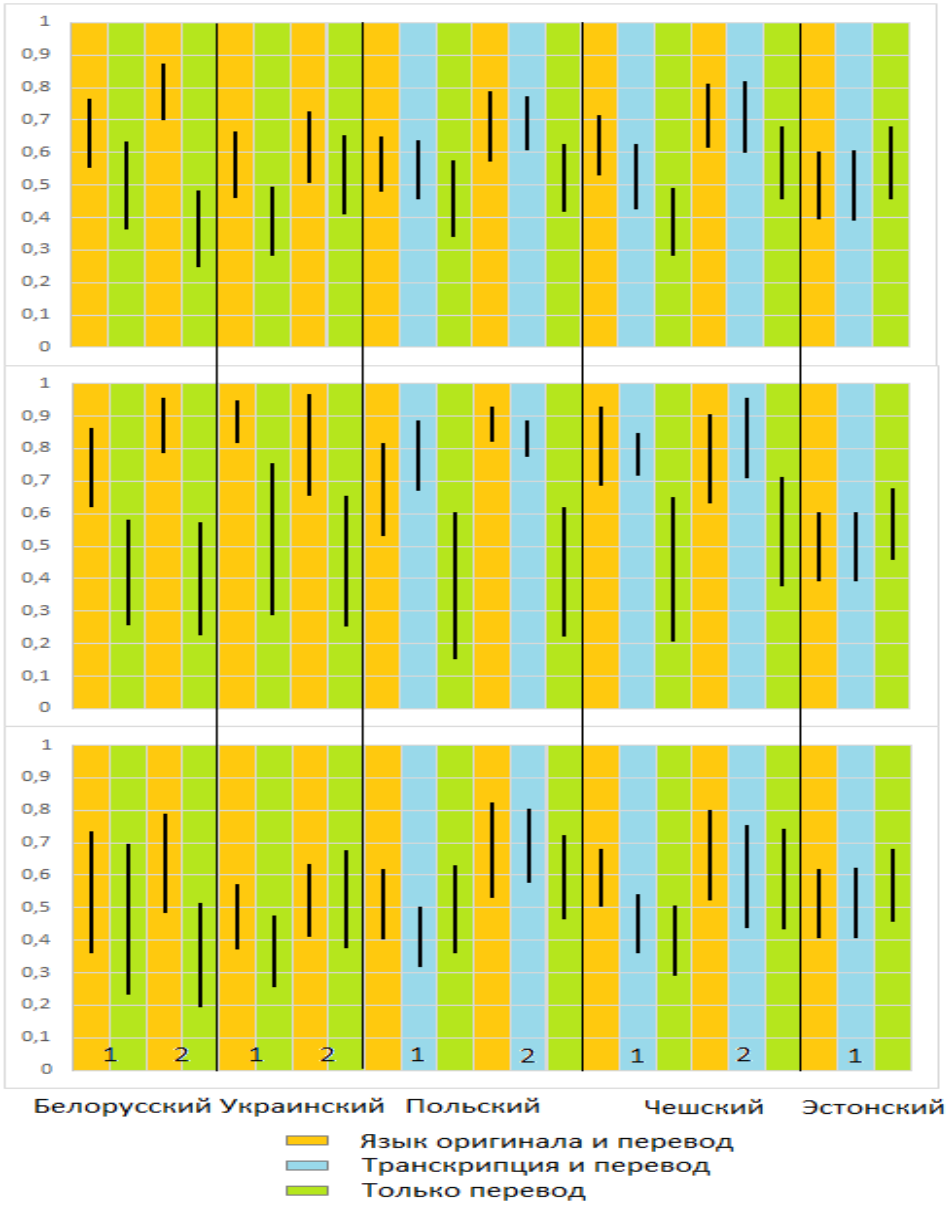
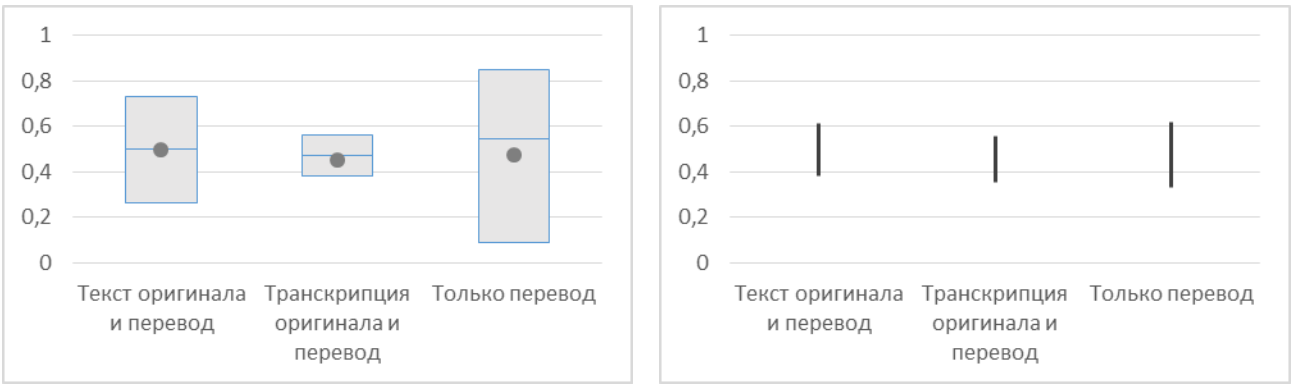



*Рис. 3. Доверительные интервалы для правильных ответов по тестам: все слова (а), созвучные (б) и несозвучные (в) слова*

необходимо заметить, что значение 1,96 выбиралось как значение для 95% доверительного интервала при нормальном распределении значений. Однако, как уже обсуждалось выше, распределение значений отличается от нормального, то есть коэффициент перед стандартной ошибкой должен быть другим. Однако текущее значение позволяет ориентироваться в значимости получаемых результатов.

На рис. 4а и б представлены квартили и доверительные интервалы для «ложных друзей переводчика». В связи с тем, что в тексте их было очень мало, расчет проводился по всем словам всех тестов. Заметим, что, вопреки ожиданиям, картина практически не отличается от несозвучных слов. Это может быть связано либо с усреднением по разным тестам, либо с тем, что информанты чувствовали диссонанс между ложным значением и контекстом и использовали только контекст. В нескольких тестах информанты ухудшали свои показатели при наличии текста оригинала, но чаще всего результаты были сопоставимы.




а) б)

*Рис. 4. Диаграмма размаха и доверительные интервалы для «ложных друзей переводчика» по всем тестам*

Для каждого теста была рассчитана средняя доля ошибок для трех выбранных типов ответов: полностью корректный, частично корректный (правильная часть речи, неправильное слово) и некорректный (см. рис. 5). Как видно из рисунка, при наличии текста на языке оригинала информанты в среднем допускали меньше ошибок. Чаще всего снижалось количество ошибок обоих типов, однако для польского языка это не так.

Для каждого теста было также рассчитано улучшение результатов тестов при добавлении текста на языке оригинала или его транскрипции (см. рис. 6). Для всех языков, кроме польского (и эстонского, в котором улучшения не наблюдается), сократилось количество частично корректных ответов и полных ошибок. В польском языке количество частично корректных ответов увеличилось, при том что общее число ошибок двух видов уменьшилось.

Кроме того, мы рассчитали значения для F-теста между разными вариантами одного теста: параллельные тексты оригинала и перевода с одним переводом, транскрипции и перевода с одним переводом, параллельные тексты между собой. Но так как наши данные не распределены по нормальному

закону, результаты теста могут быть поставлены под сомнение. В связи с этим мы также рассчитали значение корреляции между тестами в тех же комбинациях. Результаты показаны на рис. 7. В полученных результатах наблюдается значительный разброс. Однако можно утверждать, что в целом сходство между двумя параллельными тестами на одном языке выше, чем между параллельным тестом и тестом только на русском языке. Для эстонского языка степень уверенности принимает максимальное значение, различия между разными тестами выражены не так ярко.

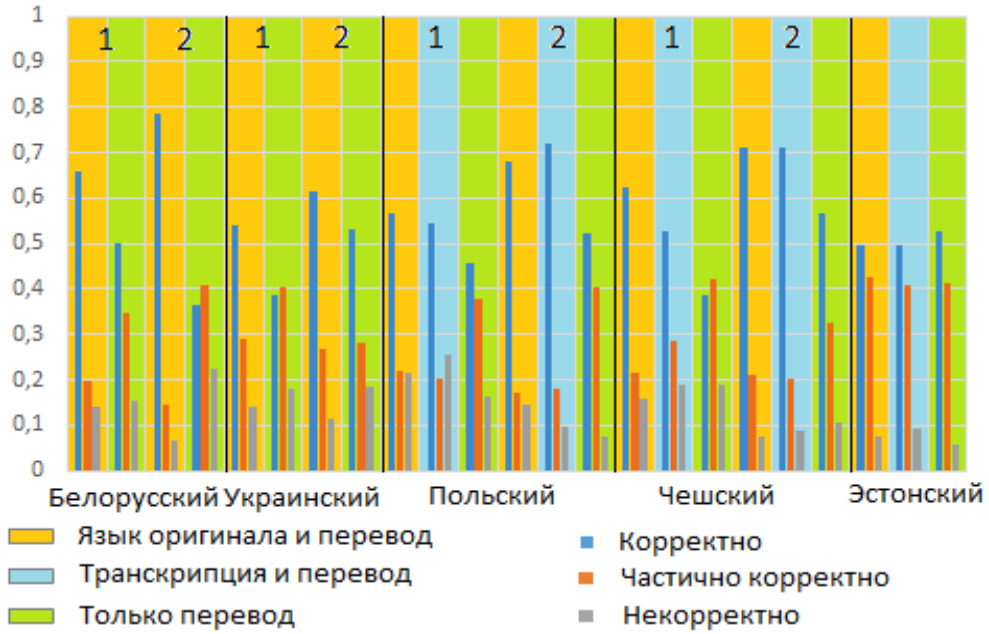



*Рис. 5. Распределение ошибок по видам*

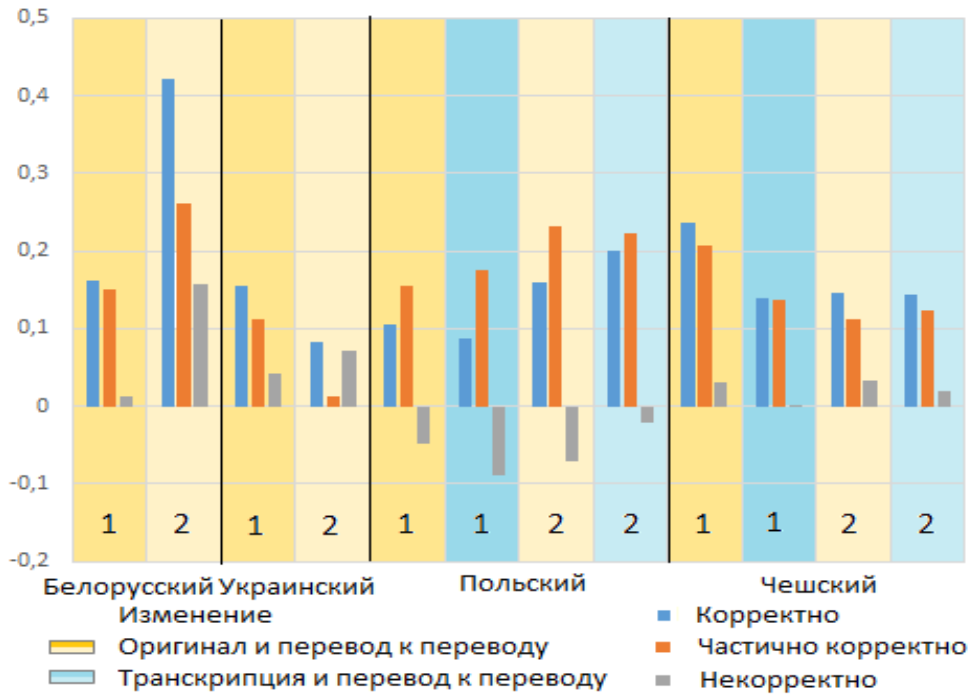



*Рис. 6. Улучшение результатов тестов относительно работы информанта только с русскоязычным текстом*

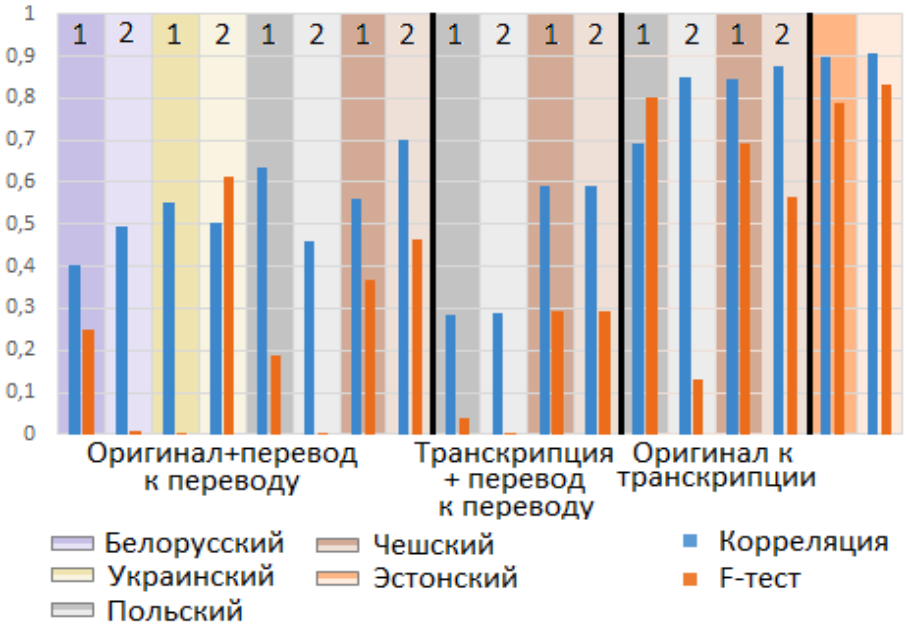



*Рис. 7. Значения корреляции и F-теста для различных комбинаций тестов*

## 5. Результаты тестирования на едином тексте и их статистическая оценка

Мы провели аналогичные тесты для единого текста, переведенного на все рассматриваемые языки. Построенная диаграмма размаха показывает картину, аналогичную полученной на различных текстах. Для созвучных слов пересечение второго и третьего квартилей наблюдается только для польского языка (см. рис. 8б), тогда как для несозвучных слов результаты для всех языков, включая русский, находятся примерно в одном диапазоне.

Как видно из рис. 9, доверительные интервалы для созвучных слов для тестов с параллельными текстами не пересекаются с доверительными интервалами для русского языка. Более того, не пересекаются доверительные интервалы для созвучных (рис. 9б) и несозвучных (рис. 9в) слов. Однако для несозвучных слов картина примерно такая же, как для теста, включающего в себя только русский текст.

На рис. 10 показано, что прирост точности ответов при наличии текста на иностранном языке сохраняется. Как и на первом этапе, тесты для польского и чешского языков показывают всплеск числа ошибок, который может быть объяснен тем, что пользователи ожидают подвоха при анализе текста, уровень понятности которого ниже определенного уровня.

Как и в случае с этапом 1, польский язык показывает относительное высокое сходство с результатами, полученными только для русских текстов (см. рис. 11). Однако уровень значимости F-теста не позволяет сделать однозначные выводы.

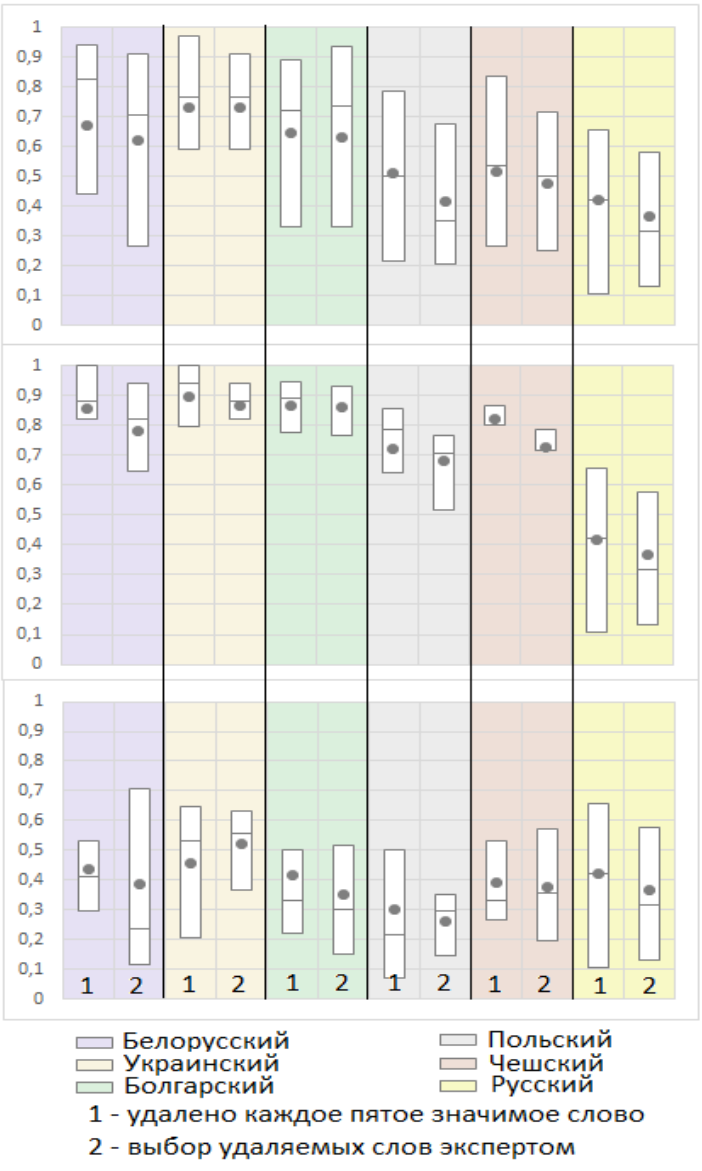



*Рис. 8. Диаграмма размаха правильных ответов по тестам: все слова (а), созвучные (б) и несозвучные (в) слова; цифра указывает на номер эксперимента*

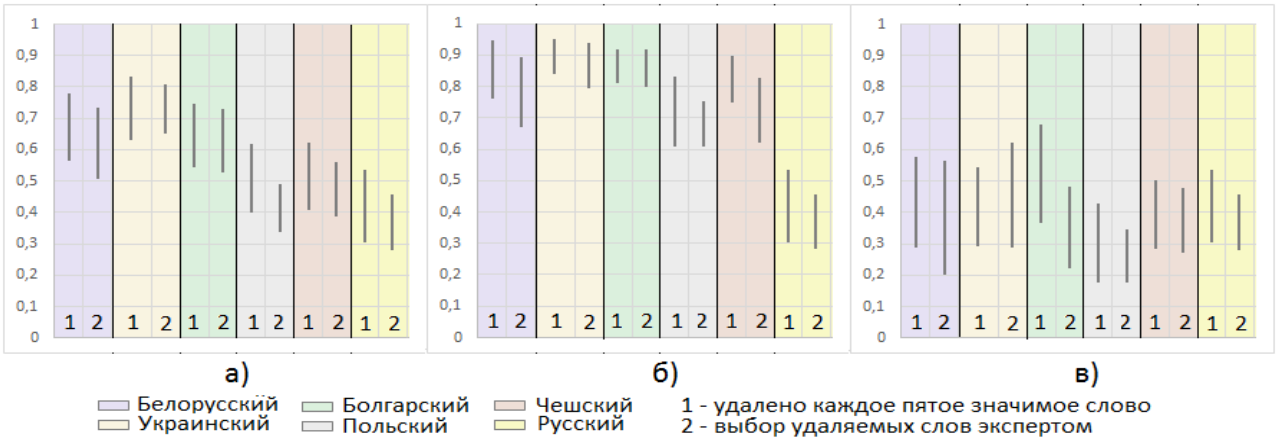



*Рис. 9. Доверительные интервалы для правильных ответов по тестам: все слова (а), созвучные (б) и несозвучные (в) слова*

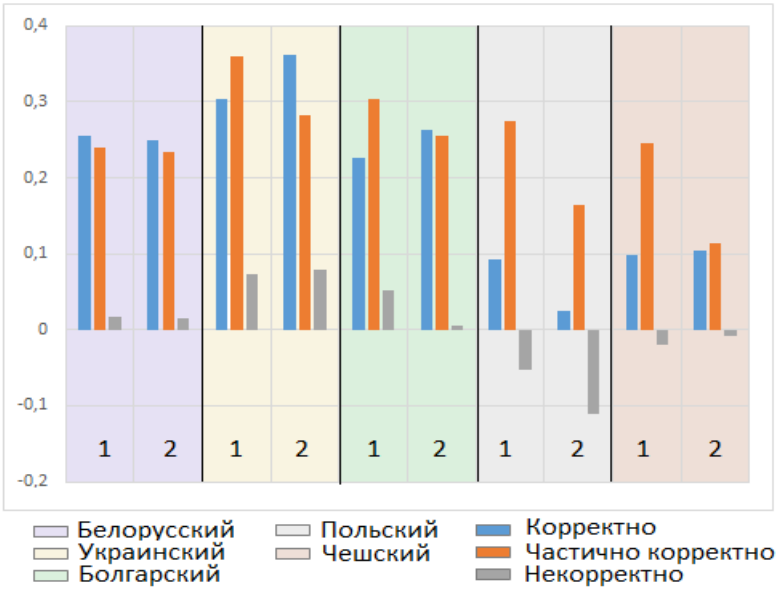



*Рис. 10. Улучшение результатов тестов относительно работы информантом только с русскоязычным текстом (второй этап)*

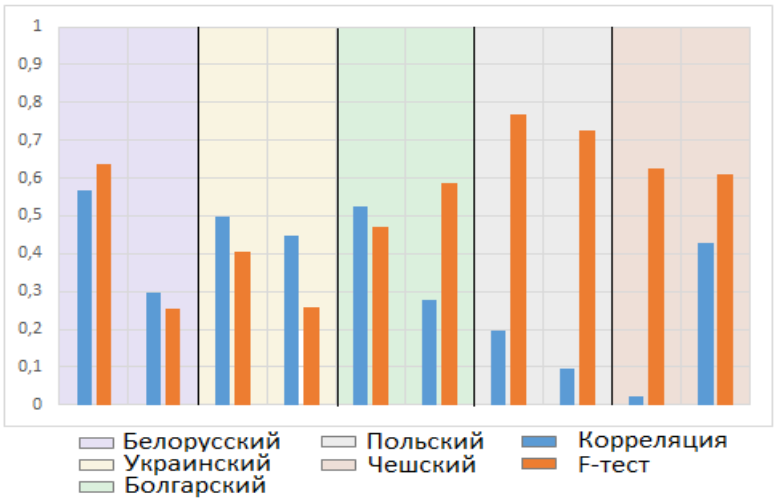



*Рис. 11. Значения корреляции и F-теста для тестов второго этапа*

Из всего вышесказанного можно сделать вывод, что наличие иностранного текста на незнакомом языке той же группы с высокой достоверностью помогает понять этот текст, но только для созвучных слов. Несозвучные слова подобного прироста не дают. Для польского языка уровень сходства результатов тестов высок, но не позволяет делать однозначных выводов. При этом прирост правильных ответов при наличии текста на иностранном языке наблюдается для всех языков.

## 6. Обсуждение результатов

Тесты для эстонского языка показали, что если информанту предъявляется текст на полностью незнакомом языке, из которого информант не может выделить знакомые слова, результаты для разных вариантов проведения тестов

практически не отличаются. При этом для славянских языков такие изменения наблюдаются. В связи с этим мы принимаем гипотезу о корректности самой методики тестирования и дальнейшие рассуждения будем вести лишь для выбранных славянских языков.

Проведенные эксперименты позволили оценить процент слов, которые могут быть поняты в оригинальном тексте без перевода. На всех тестах мы получили устойчивый рост среднего количества правильных ответов при наличии текста на языке оригинала по сравнению с его отсутствием. Рост точности для тестов первого этапа составил от 8% до 42% со средним в 16,9%. Если проанализировать отдельно созвучные и несозвучные слова, то рост составил от 22% до 47% при среднем 35,8% для созвучных и от 0% до 28% при среднем значении 10,5% для несозвучных. Для тестов второго этапа (единый текст для всех языков) рост точности прохождения тестов составил от 4,5% до 36% со средним в 20%. Для созвучных слов рост составил от 30% до 49% со средним в 41%; для несозвучных слов рост составил от -10% до 8,5% со средним в 0%.

Мы выяснили, что для всех тестов первого этапа, кроме двух, доверительные интервалы теста с исходным текстом или его транскрипцией пересекаются с доверительным интервалом теста, в котором предъявлялся только перевод. Для второго этапа все доверительные интервалы не пересекаются. Однако во всех случаях среднее значение точности ответов ниже, если информантам не предъявлялся текст на языке оригинала. Для созвучных слов картина прямо противоположная – пересечение доверительных интервалов наблюдается лишь для двух тестов из восьми. Доверительные интервалы для несозвучных слов всегда перекрываются, хотя бы в минимальной степени.

Если проанализировать пересечения доверительных интервалов для созвучных и несозвучных слов в рамках одного теста с предъявлением оригинала (рис. 2б и 2в, 9б и 9в), то мы увидим, что эти интервалы пересекаются только в двух случаях из восьми (еще в двух случаях пересечение составило всего 0,2% при ширине интервала в 25-30%, то есть пренебрежимо мало). При этом средняя точность ответов для созвучных слов всегда выше (рис. 1б и 1в, 8б и 8в). Для тестов, предполагавших работу только с переводом, доверительные интервалы пересекаются всегда, а средние значения отличаются в обе стороны. Таким образом, можно утверждать, что точность восстановления созвучных слов при наличии перевода статистически выше, чем восстановления несозвучных слов.

На рис. 12 представлена зависимость прироста понятности теста при наличии параллельного иностранного текста от процента созвучных слов в иностранном тексте и понятности текста по результатам экспериментов. Из рис. 12В очевидно, что наличие параллельного текста на иностранном языке той же группы увеличивает понятность текста в целом пропорционально уровню прироста понятности относительно только русского текста. Корреляция здесь

на полученных данных равна 0,8 (0,97 для данных, полученных на едином тексте для всех языков). Рис. 12А показывает, что прирост понятности имеет тенденцию к увеличению с ростом процента созвучных слов. Корреляция для всех данных равна 0,62 (0,75 для данных, полученных на едином тексте для всех языков). Наконец, понятность текста растет с увеличением процента созвучных слов (рис. 12Б; для удобства представления процент созвучных слов взят по соответствующему параллельному тексту, то есть русские тексты скорее показывают ограничение снизу; формально их процент созвучных слов равен нулю). Здесь корреляция равна 0,62 для всех данных и 0,82 для тестов, построенных по единому тексту. Заметим, что русские тексты на рис. 12Б показывают скорее ограничение снизу – заполнение пропусков без каких-либо подсказок.

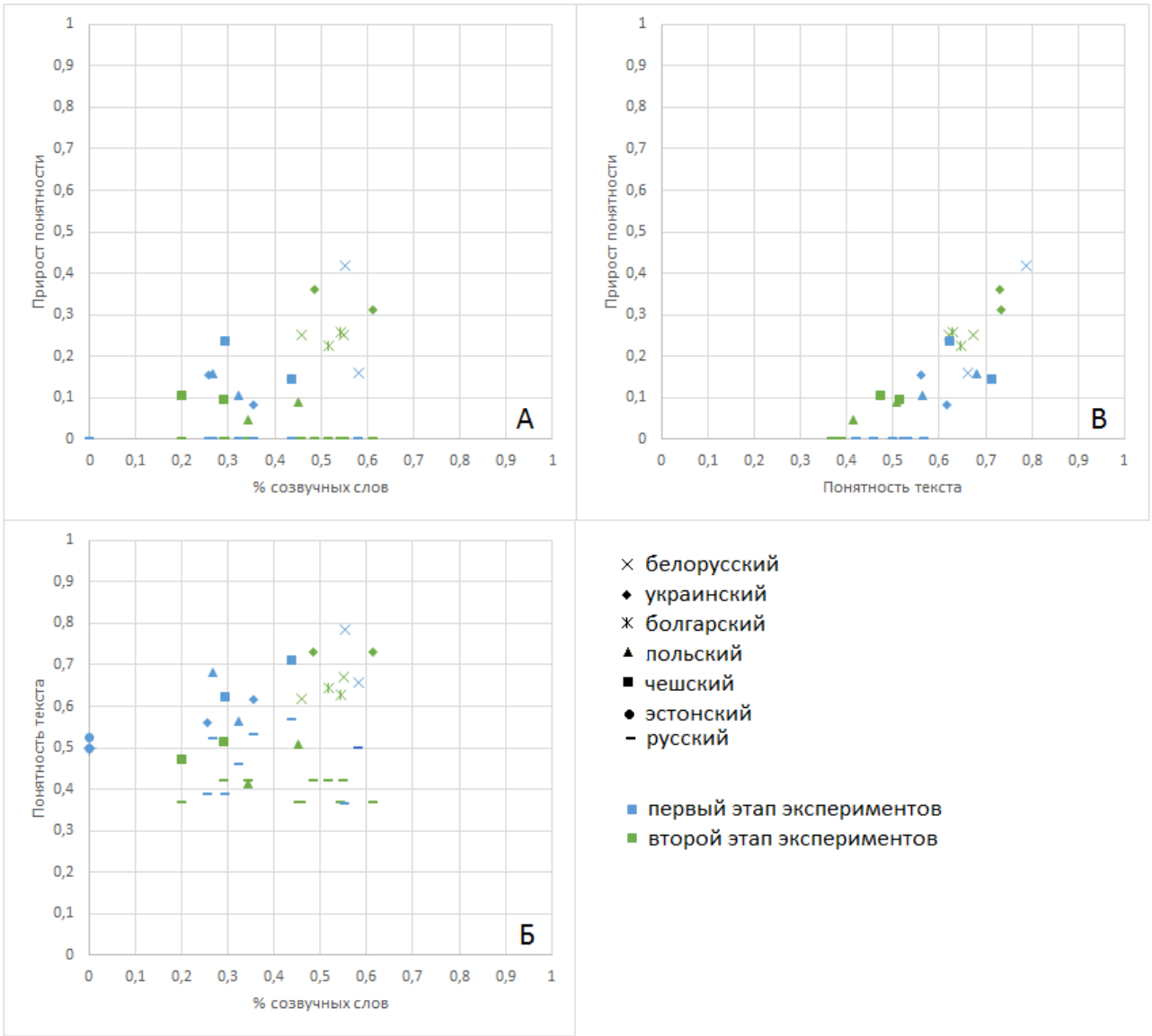



*Рис. 12. Зависимость прироста понятности теста при наличии параллельного иностранного текста от % созвучных слов в иностранном тексте и понятности текста по результатам экспериментов*

Из рис. 12 также видно, что понятность текста (пропущенных слов в тексте) имеет ограничение снизу – примерно треть пропущенных слов может быть восстановлена только исходя из контекста. Помимо этого, она ограничена снизу процентом слов, созвучных словам родного языка перевода и сходных с ними по значению.

Прирост понятности текста также имеет ограничение, на сей раз сверху. С одной стороны, он ограничен сверху процентом созвучных слов. С другой стороны, прирост понятности ограничен общей понятностью параллельного текста, хотя здесь, скорее всего, следует говорить об обратной зависимости – понятности текста от прироста. В любом случае, прирост понятности текста за счет добавления параллельного текста на незнакомом сходном языке не превышает понятности текста за вычетом упомянутой границы понятности снизу (около 35%).

Из этого можно сделать следующие выводы:

1) информанты в среднем лучше восстанавливают отдельные слова при наличии текста на языке, сходном с их родным, или его транскрипции (при том, что они не знакомы с этим языком);

2) информанты лучше восстанавливают созвучные слова из параллельного текста;

3) понятность параллельного текста зависит от процента созвучных слов в текстах на иностранном языке;

4) значимой разницы между восстановлением слов при наличии оригинального текста или его транскрипции не наблюдается, хотя уровень корреляции для текстов с транскрипцией выше;

5) результаты тестов на едином тексте для всех языков показывают лучшую корреляцию, чем тесты на различных языках; в целом такие тесты более адекватно описывают близость языков.

Помимо этого, следует заметить, что проведенные тесты не позволили сделать достоверных выводов о влиянии частоты совместной встречаемости слов на ответы информантов. В большинстве случаев удаленные слова не составляли частотное сочетание с другими словами текста. Однако для связанных между собой слов можно сказать, что при наличии текста оригинала информанты чаще делали ошибки в тех случаях, когда пропущенное слово встречалось чаще (видимо, информанты ожидали в данном случае подвоха). И наоборот, если слова редко встречались вместе, наличие оригинала помогало более корректно восстановить пропущенное слово (видимо, не найдя объяснения в собственном опыте, информант пытался разобраться с текстом оригинала). В любом случае, данный вопрос нуждается в отдельном исследовании и специальной проработке тестов.

## 7. Выводы

Проведенные эксперименты показали, что носители русского языка достоверно лучше восстанавливают текст при наличии оригинала, написанного

на одном из выбранных славянских языков (белорусский, украинский, польский и чешский). При наличии только текста, написанного на русском языке, информанты восстанавливают слова с точностью около 35-55%. В этом случае информанты не могут корректно восстановить часть речи для 5-25% пропущенных слов. В остальных случаях информанты ошибались с выбором слова, но корректно определяли его часть речи. При наличии параллельного текста информанты на 5-40% лучше восстанавливают пропущенные слова (40-80% корректно восстановленных слов). Улучшение достигается за счет наличия в оригинальном тексте слов, которые переводятся созвучными русскими словами. Несозвучные слова восстанавливаются так же, как в случае с отсутствием перевода, то есть за счет контекста.

Из проведенных экспериментов можно сделать вывод о том, что степень понятности текста, написанного на иностранном языке, не известном информанту, но являющимся родственным его родному языку, определяется процентом созвучных слов, обладающих сходным смыслом. Использованный нами подход может быть использован в дальнейшем для определения степени понятности иностранных языков для людей, владеющих языками из той же группы. Полученные результаты позволяют в дальнейшем априори оценивать, например, процент слов, которые могут быть восстановлены информантами из контекста или из текста на иностранном языке, который также является подсказкой.

Также можно сделать вывод о том, что степень сходства языков может быть среди прочего определена через процент созвучных слов в текстах на двух языках. Для оценки подобного сходства могут быть использованы результаты экспериментов, проведенных нами над одним и тем же текстом с удалением каждого пятого значимого слова, так как они ставят всех информантов в одинаковое положение, а результаты тестов являются сравнимыми. В соответствии с результатами экспериментов мы можем расположить исследованные языки в следующем порядке, показывающем убывание степени их сложности как L2 для русскоязычных изучающих: украинский (прирост 31% по сравнению с текстом только на русском языке), белорусский (прирост 25%), болгарский (прирост 23%), чешский (прирост 9%), польский (прирост 8%). В дальнейших исследованиях мы планируем рассчитать процент созвучных слов в текстах на исследованных языках и проверить высказанную гипотезу о языковом сходстве.

## 8. Благодарности

Авторы признательны Заболотней Т.Н. и Аксенову С.А. за подготовку текстов для работы с информантами, а также Файзуллину Р.Л. за слепую проверку полученных результатов.

Данный проект частично поддержан грантами РГНФ № 12-04-00060 и 15-04-12019.

# Список литературы

**[Ackerman, 2000]** Ackerman P.L., Beier M.E., Bowen K.R. Explorations of crystallized intelligence: Completion tests, cloze tests, and knowledge. In Learning and Individual Differences, Volume 12, Issue 1, March 2000, pages 105–121.

**[Ageeva, 2015]** Ageeva E., Tyers F. M., Forcada M. L., and Pérez-Ortiz J. A. Evaluating machine translation for assimilation via a gap-lling task. In EAMT-2015: 18th Annual Conference of the European Association for Machine Translation, pages 137-144, Antalya, Turkey.

**[Ciobanu, 2014]** Ciobanu A. M. and Dinu L. P. (2014). On the Romance Languages Mutual Intelligibility. In LREC-2014: Language Resources and Evaluation Conference, pp. 3313-3318, Doha, Qatar.

**[Gooskens, 2007]** Gooskens C. The Contribution of Linguistic Factors to the Intelligibility of Closely Related Languages. Journal of Multilingual and Multicultural Development, 2007. 28(6): pp. 445-468.

**[Kusters, 2003]** Kusters W. Linguistic complexity: the influence of social change on verbal inflection. Utrecht, 2003.

**[Lindsay, 2014]** Lindsay R. Mutual Intelligibility among the Slavic Languages. http://www.academia.edu/4080349/Mutual_Intelligibility_of_Languages_in_the_Slavic_Family. 2014. [Интернет; доступ 07 августа 2016].

**[Lupyan, 2009]** Lupyan G., Dale R. Language structure is partly determined by social structure // PLoS ONE. 2009. 5(1).

**[Miestamo, 2008]** Miestamo M. Grammatical complexity in a cross-linguistic perspective // Language Complexity: Typology, contact, change. Amsterdam. 2008. pp. 23-41.

**[Zarei, 2013]** Zarei A. A. and Ab, M. A. The Contribution of Word Formation, Code Mixing, Multiple Choice, and Gap Filling Tasks to L2 Vocabulary Comprehension and Production. International Journal of Language Learning and Applied Linguistics World (IJLLALW), 2013. 4(1): pp. 7-55.

**[Бердичевский, 2012]** Бердичевский А. Языковая сложность // Вопросы языкознания №5, 2012. М.: Институт русского языка им. В. В. Виноградова РАН, с. 101-124.

**[Даль, 2009]** Возникновение и сохранение языковой сложности / Эстен Даль; пер. с англ. Д. В. Сичинавы. - Москва : ЛКИ, 2009. - 558 с.

**[Елкин 2003]** Елкин С.В., Клышинский Э.С., Стекляников С.Е. Проблемы создания универсального морфосемантического словаря // Сб. трудов Международных конференций AIS’03 и CAD-2003, т. 1, Дивноморское. 2003

**[Колмогоров, 1965]** Колмогоров А. Н. Три подхода к определению понятия «количество информации». // Проблемы передачи информации, № 1, М.: 1965, с. 3–11.

**[Ягунова, 2010]** Ягунова Е.В. Исследование избыточности русского звучащего текста // Труды института лингвистических исследований, том 4, часть 2. СПб: Наука, 2010, с. 90-114.

# Оглавление

1. Введение ..... 3
2. Метод определения понятности иностранного текста ..... 5
3. Используемые инструменты и данные ..... 9
4. Результаты тестирования на различных текстах и их статистическая оценка ..... 10
5. Результаты тестирования на едином тексте и их статистическая оценка ..... 15
6. Обсуждение результатов ..... 17
7. Выводы ..... 20
8. Благодарности ..... 21

Список литературы ..... 22