УДК 81'322.2:81'373.423+519.257

Статистический метод количественной оценки понятности иностранных славянских языков для русскоязычных носителей

Э.С. Клышинский
Институт прикладной математики им. М. В. Келдыша РАН
Москва, Россия

Аннотация

В данной статье разбирается вопрос понимаемости иностранного текста на славянском языке для неподготовленного информанта. Целью статьи было выяснить, какую долю слов иностранного текста информанты смогут понять при условии, что они не знакомы с этим языком. Для определения понимаемости текста мы использовали параллельный текст с пропущенными словами. В русской версии текста пропускалась часть слов, задачей информанта было восстановить эти слова используя в качестве подсказки параллельный текст на одном из славянских языков: украинском, белорусском, польском, чешском, словацком, сербском, словенском и болгарском. Часть информантов использовалась в качестве контрольной группы, и параллельный текст им не предъявлялся.

Мы высказали гипотезу о том, что понятность текста на иностранном языке может быть определена как увеличение доли корректно восстанавливаемых слов группы, которой предъявляется параллельный текст на иностранном языке, над долей слов, корректно восстановленных контрольной группой. Результаты экспериментов подтвердили нашу гипотезу.

Также мы разделили все пары «пропущенное слово — перевод» на четыре группы: полные и частичные когнаты, генетические когнаты, не когнаты и ложные друзья. Корреляция средней понятности текста по всем информантам для данного языка с долей полных и частичных когнатов составила 0.7, тогда как для остальных групп была отрицательной. За счёт этого можно утверждать, что понятность иностранного текста по большей части определяется долей полных когнатов, но при этом зависит от некоторых других параметров.

Все результаты экспериментов и программное обеспечение для их анализа размещены по адресу https://github.com/klyshinsky/mutual_intelligibility_Russian .

Ключевые слова

понятность текста, славянские языки, тест с пропусками, корреляция, квантитативная лингвистика

Quantitative Estimation of Intelligibility of Foreign Slavic Languages: Case of Russian Native Speakers

Eduard S. Klyshinsky[1], Olesya V. Karpik[2]
[1] National Research University Higher School of Economics
[2] Keldysh Institute of Applied Mathematics RAS
Moscow, Russia

Abstract

In this article, we investigate the issue of intelligibility of a foreign Slavic text for a Russian-speaking person which don't know this language. The aim of this article is to find out what is the percentage of intelligible words in foreign text for such a person. As a main measuring tool, we used parallel cloze tests with omitted words in the Russian part. The task was to restore omitted words using the foreign part of a test (written in Ukrainian, Belorussian, Polish, Czech, Slovak, Serbian, Slovene, and Bulgarian languages) as a clue. As a baseline, we used a control group which solved a test without the foreign part.

Our hypothesis was that the foreign text intelligibility could be defined as a difference between the mean percentage of correctly restored words for a group used a parallel text and the same percentage for a control group. The results of our experiments proved our hypothesis.

All the pairs "omitted word – its translation" was divided into four groups: full and partial cognates, genetic cognates, non-cognates and false friends. The correlation between the mean intelligibility of a

text in a given foreign language and the percentage of full and partial cognates was as high as 0.7; the same correlation for the other word groups was negative but not so deep. Therefore, we can state that the foreign text intelligibility is defined by the percentage of full and partial cognates but that is not the only parameter.

The gathered data, containing the used tests, users' answers and their background, and the software for its analysis is placed at https://github.com/klyshinsky/mutual_intelligibility_Russian .

Keywords

text intelligibility, Slavic languages, cloze test, correlation, quantitative linguistics

1. Введение

В данной работе исследуется понятность славянских языков для носителей русского языка. Под понятностью иностранного языка мы понимаем ситуацию, когда носители языка А могут без предварительной подготовки понимать тексты и речь на языке В. Как правило, близкородственные языки интуитивно понятны даже непрофессионалам: их родство легко заметить самим носителям. Так, например, любой говорящий по-русски, услышав речь на любом из славянских языков, без труда определяет его принадлежность к славянской группе. Феномен «понятности» родственных языков, по мнению лингвиста М. Бейкера (M. Baker) служит маркером, позволяющим отличать «своих» от «чужих» [1]. Разнообразие языков разделяет человечество на группы. Принадлежность к одной языковой группе способствует возникновению «внутригрупповой солидарности», которая, в свою очередь, побуждает проявлять альтруизм по отношению к людям, с которыми мы имеем общие гены. При этом заметим, что ни одно объяснение разницы между языками с точки зрения культурных, климатических или социологических параметров не описывает текущую языковую картину.

Понятность родственного языка зависит от многих факторов: знакомства индивидуума с другими культурами и обсуждаемой темой, его общей эрудированности, степени владения родным языком и словарного запаса. Традиционно считается, что русский язык, как представитель восточнославянской подгруппы, наиболее близок другим представителям этой же подгруппы – белорусскому и украинскому. Взаимопонятность русского и белорусского языков почти полная, то есть, представители этих языков могут общаться на своем родном языке и понимать друг друга. Однако, как на белорусском, так и на украинском языке легко составить такое предложение, которое будет совершенно непонятно неподготовленным носителям русского. При этом некоторый опыт восприятия украинской речи, хорошая лингвистическая интуиция, знакомство с русскими архаизмами значительно повышают понятность украинского языка, даже без его знания. При прочих равных условиях, письменная речь понимается значительно лучше устной, так как первая усваивается со скоростью, удобной для читателя, тогда как ритм второй задает произносящий.

В серии работ Шарлотты Гускенс [2-5] было статистически доказано, что взаимная понятность языков не является симметричной, то есть текст на языке А может быть менее понятен носителям языка В, чем тот же текст на языке В носителям языка А. В работах [2, 6] показано, что использование условной энтропии вместо меры Левенштейна позволяет более корректно моделировать подобную несимметричность и предсказывать взаимную понятность текстов с ее помощью. Несмотря на построение регрессионых прямых, в среднем описывающих взаимную понятность текстов, а также классификаторов, позволяющих на основе методов машинного обучения различать разные типы когнатов, вопрос о причинах несимметричности взаимной понятности остается открытым.

Целью данной работы была разработка статистически достоверного формального метода, позволяющего оценить понятность иностранного текста для неподготовленного информанта. Для этого мы провели серию экспериментов, показавших на примере носителей русского языка, что разные виды когнатов (близких слов двух языков) понимаются с разной точностью, а общая понятность текста хорошо коррелирует с долей полных и частичных когнатов. Также мы показали, что вместо взаимной понятности языков следует говорить о взаимной понятности текстов, так как их лексика может существенно отличаться по степени понятности для носителя языка.

2. Обзор методов оценки понятности иностранных языков

Задача определения взаимной понятности языков была поставлена достаточно давно, однако в течение длительного времени изучалась скорее ее описательная часть, чем количественные методы оценки. Например, в работах [7-9] изучался вопрос различия омофонов и когнатов. Само применение математических методов для оценки взаимной понятности языков также было не всегда удачным. Так, например, в работах Роберта Линдсей (Robert Lindsay) проделан большой труд по сопоставлению длинного списка языков. Однако основой для количественных оценок служит опрос читателя, при котором у него спрашивается процент понятного ему текста. Очевидно, что подобный метод не может служить прочной основой для проведения количественных исследований. Еще одним примером не до конца удачного исследования служит работа [10], в которой автор на основе теории раскрашенных графов предлагает формализм различения диалектов и отдельных языков. Несмотря на прекрасный с математической точки зрения труд, количественные эксперименты не были проделаны, что позволяет усомниться в корректности математических рассуждений и их практической применимости. Важным является и вопрос исследования уровня сходства не только отдельных языков, но и диалектов (так сказать, проведение «водораздела» между языком и диалектом) [11].

Наиболее полными на данный момент представляются результаты проекта MICReLa (Mutual intelligibility of closely related languages), возглавляемого Шарлоттой Гускенс (Charlotte Gooskens). Основой для данного проекта послужила диссертация Вильберат Яна Хеерига (Wilbert Jan Heeringa) 2004 года [12], в которой было обосновано применение расстояния Левенштейна для определения когнатов в параллельных текстах. В работе [2] данный подход был развит. Как известно, взаимная понятность языков не является симметричной: носители языка А могут понимать язык В лучше, чем носители языка В язык А [4]. Вместо расстояния Левенштейна, которое является симметричным, было предложено использовать условную энтропию. Однако по большей части участники проекта сосредоточились на вычислении количественных значений взаимной понятности языков в ходе экспериментов с информантами. В разные годы материалом для исследования становились скандинавские [3], славянские [4] и германские [5] языки.

Авторы выделили два вида понятности иностранного языка: фонетическую понятность и понятность текста. Сами эксперименты также строились по нескольким методикам: информант должен был выбрать одно из четырех изображений, соответствующих предъявленному стимулу; вставить в текст пропущенные слова, выбирая их из предъявленного списка; пересказать прочитанный текст и т. д. Все эксперименты проводились в двух форматах: с письменной и звучащей речью. Авторы сумели получить убедительные результаты, доказывающие все основные теоретические положения. Однако к методике их исследования можно предъявить целый ряд претензий. При выборе одной из двух картинок есть четкое постоянное контрольное значение: вероятность случайного угадывания равна 0,5. Однако данное значение довольно велико, а превышение над ним не всегда статистически значимо. В тестах, заключающихся во вставке слова из списка, информант ограничен самим этим списком, то есть проверяется скорее его умение вставлять фиксированные слова на правильные позиции, чем собственно его понимание текста и иностранного языка. Таким образом, можно утверждать, что методика проведения подобных экспериментов должна быть улучшена.

Имеются и некоторые претензии к статистическим результатам работ. Так, в работе [3] для каждого языка было взято около 30 информантов (что само по себе может быть достаточным), которые были разделены по городам (то есть разделены на более мелкие группы, которые не являются достаточными для твердых выводов).

Следует заметить, что при наличии небольшого материала более корректно рассуждать о понятности конкретного текста, а не о понятности языка в целом. Проводя эксперименты с информантами на ограниченном числе текстов мы не можем надежно сделать вывод о взаимной понятности языков, но лишь выявить некоторые закономерности. Этот недостаток

исправлен в работе [6], где авторы проводили эксперименты не с информантами, а на материале параллельных корпусов текстов, написанных на языках западной славянской группы (около 7500 слов для чешского, словацкого и польского языков). Но с формальной точки зрения, автор доказали тот факт, что взаимная энтропия текстов на западнославянских языках коррелирует с нашими представлениями о взаимной понятности языков. Анализ других работ, использующих ту же методику расчетов для других групп языков, позволяет обобщить результаты до большинства европейских языков.

Еще одним недостатком является тот факт, что в большинстве работ не исследуются параметры, от которых зависит понимание отдельного текста. Каждая из работ приводит цифры, подтверждающие несимметричность взаимной понятности языков. Только в работе [5] проведен анализ влияния лексических и синтаксических параметров на понимаемость текста. Авторами был сделан вывод, что основную роль играет лексическое сходство слов в разных языках. Но для получения результатов использовался лишь один текст. Полученные на нем результаты, безусловно, нельзя обобщать на языки в целом.

Наконец, подобные исследования не проводились на материале русского языка. В своей предыдущей работе [13] мы уже поднимали данный вопрос, но объем исследованного материала позволяет говорить лишь о предварительном исследовании.

В данной работе мы исправляем некоторые из указанных недостатков. Исследование проведено на шести фрагментах текстов двух авторов в переводе на восемь славянских языков всех групп. Для достижения статистической значимости результатов, для каждого перевода текста было опрошено от 35 до 80 человек.

3. Метод оценки понятности текста и постановка эксперимента

Из описанных в предыдущем разделе вариантов проведения экспериментов мы выбрали тесты на заполнение пропусков в тексте: из текста вычеркиваются некоторые слова или части слов (например, каждое пятое слово, слова на заданную тему или окончания слов). Данная методика часто используется при изучении иностранного языка для проверки уровня знаний студентов или расширения их словарного запаса [14] или в психологии [15]. Кроме того, в работе [16] было предложено использовать этот метод для оценки качества машинного перевода. В нашем случае, информанту предъявлялся текст на русском языке с пропущенными словами и параллельный текст на иностранном славянском языке. Задачей информанта являлось восстановить слова по контексту. Данный метод позволяет собрать объективную количественную информацию по понятности фрагментов текста, не ограничивая фантазию информанта. Также данный метод прост и в программной реализации, и в прохождении информантами, что позволило выложить тест в сеть Интернет.

Как это было показано в [17], восстановление слов текста может производиться за счет его избыточности. При отсутствии других подсказок, информант может использовать дистрибутивные свойства и грамматические характеристики слов контекста. Также на подбор слов влияет личный опыт информанта. При этом верно и обратное: информант может проигнорировать наличие подсказки и попытаться восстановить слово по контексту. Для того, чтобы определить влияние наличия перевода, мы использовали контрольную группу, которой предъявлялся текст только на русском языке.

При принятии решения о том, какое слово должно быть вычеркнуто из текста мы руководствовались следующими соображениями. Задача должна иметь видимое простое решение, поэтому предпочтение отдавалось словам, которые скорее всего будут знакомы информанту. При этом доля незнакомых слов также должны быть заметной для проверки наших гипотез. Пропускались только слова со значимыми частями речи. Часть слов выбирались так, чтобы встретиться более чем в одном тесте. В дальнейшем, это позволит проверить влияние контекста на угадываемость слова. Задача осложнялась тем, что для разных языков степень понятности слова и его перевода существенно варьируется. Для каждого из параллельных языков пропускались одни и те же русские слова. Следуя за работой [4], мы расширили список слов до 186 различных лемм.

Вписанные информантом слова оценивались по следующей шкале.

- Если вписано слово, которое имеется в исходном тексте, один из его синонимов, или близкое понятие, ответ считается корректным. Форма слова при этом может меняться, если при этом не нарушается повествование. Например, пользователь свободно может изменить падеж или число слова, но изменение времени или совершенности глагола не должно менять смысл фразы. В этом случае считается, что пользователь полностью и корректно понял контекст слова и его значение.
- Ответ считается частично корректным, если информант вписывает однокоренное слово, но путает часть речи, либо вписывает слово с той же частью речи, не нарушающее логики повествования. В этом случае считается, что пользователь понял само слово, но не понял синтаксическую структуру текста, либо понял его синтаксическую структуру, но не смог корректно перевести слово.
- Если информант вписывает неправильное слово с ошибочной частью речи, оставляет поле пустым, либо вписывает слово, идущее в разрез повествованию, ответ считается полностью некорректным.

Например, для фразы *...рассуждал он по дороге...* информанты должны были восстановить последнее слово. Ответы *дороге, пути* засчитывались как полностью корректные; *ходу, наитию, ночам* — как частично корректные; ответы *своему, обычному, римский, себе* — как полностью некорректные.

Каждое корректно вписанное слово оценивалось как одно очко, частично корректное слово — 0,5 очка и некорректное слово — ноль очков. Проверка проводилась одним экспертом.

Таким образом, понятностью текста для информанта будет среднее значение оценок за все слова теста. Понятность пропущенного слова теста может быть определена как среднее значение оценок пользователей, полученных за данное слово. Понятность слова заданного иностранного языка может быть определена как разница между понятностью слова при наличии параллельного текста на данном языке и понятностью того же слова при отсутствии параллельного текста. Наконец, понятность текста на иностранном языке может быть определена как разница между средней понятностью текста при наличии перевода и средней понятностью только русского текста.

Для того, чтобы гарантировать понятность контекста и не предъявлять к пользователям требований по знаниям предметной области, мы выбрали фрагменты художественных произведений. В качестве параллельных текстов были взяты художественные переводы, изданные в последней части XX века. В качестве материала для тестов использовались фрагменты из произведений М.А. Булгакова “Мастер и Маргарита” и Г. Сенкевича “Камо грядеши”. Первое произведение исходно написано на русском языке, использованы его художественные переводы на сербский, словенский, болгарский, польский, чешский, словацкий, украинский и белорусский. Второе произведение исходно написано на польском языке, использовались художественные переводы на сербский, словенский, болгарский, чешский, словацкий, украинский, белорусский и русский. Для обоих произведений было выбрано по три фрагмента:

- “Мастер и Маргарита”, Пилат покидает место казни Иешуа — 42 пропущенных слова из 390, 16 предложений (тест 1);
- “Мастер и Маргарита”, диагностика Бездомного в сумасшедшем доме — 39 пропущенных слов из 396, 23 предложения (тест 2);
- “Камо грядеши”, поездка Виниция в горящий Рим — 35 пропущенных слов из 278, 20 предложений (тест 3);
- “Камо грядеши”, размышления Виниция о сбежавшей Лигии — 44 пропущенных слова из 408, 20 предложений (тест 4);
- “Камо грядеши”, Виниций думает о похищенной Лигии — 37 пропущенных слов из 380, 23 предложения (тест 5);
- “Мастер и Маргарита”, беседа с Воландом на Патриарших Прудах после описания сцены казни — 35 пропущенных слов из 337, 16 предложений (тест 6).

Заметим, что для словацкого языка было проведено только пять тестов из шести, так как нам не удалось найти в открытых источниках один из параллельных фрагментов.

Так как у пользователей могли возникать проблемы с чтением текста, записанного в расширенной латинице, все тексты на соответствующих иностранных языках были транслитерированы в кириллицу с применением простых правил преобразования, единых для всех языков. Для того чтобы сохранить единообразие, тексты на языках, использующих кириллицу, были транслитерированы в стандартную латиницу, также с использованием простых правил преобразования.

Для участия в тестах привлекались студенты из Москвы, Санкт-Петербурга и Владивостока, школьники из Москвы, а также взрослые участники из разных городов. Часть информантов была привлечена при помощи сервиса Яндекс Толока. Все информанты проходили тест на сайте, наблюдение за ними не велось.

Перед началом теста пользователь отмечал свою возрастную группу (школьник, бакалавриант, магистрант, аспирант, закончил обучение), владение иностранными славянскими языками и профиль обучения (специальности, связанные с изучением языков, против других специальностей). Пользователь имел возможность не вводить информацию о себе. Если пользователь отмечал, что владеет каким-либо иностранным славянским языком, данный язык не предъявлялся ему в качестве параллельного. Из остальных языков тест выбирался случайным образом. Таким образом, информант (если он не скрыл эту информацию) проходил тест только на незнакомом ему языке или контрольный тест. Задачей пользователя было за ограниченное время (20-25 минут) вписать пропущенные слова.

Также был разработан интерфейс разметчика, расставляющего оценки вписанным словам, и администратора, контролирующего число пройденных тестов и их предварительные результаты. Работа разметчика была существенно ускорена за счет применения следующих правил автоматической разметки. Отсеивались все анкеты, в которых информант расходился не больше, чем в двух словах с текстом автора. В этом случае считалось, что информант помнит текст произведения наизусть или очень близко к тексту, что ставит под сомнение чистоту эксперимента. Также эти информанты часто находили произведение в сети и копировали слова из текста. Помимо этого отсеивались информанты менее чем с семью вставленными словами. В этом случае считалось, что пользователь не понял большую часть предоставленного ему текста. Эти фильтры очень помогли отсеивать информантов, пришедших с Толоки и желающих быстро получить вознаграждение. При разметке использовались следующие правила. Во-первых, если информант не вписывал свой вариант ответа, он автоматически помечался как полностью некорректный. Во-вторых, если слово, вписанное информантом, находилось в базе данных с уже проставленной отметкой, эта отметка дублировалась для новой анкеты. Как выяснилось в ходе проверки анкет, ответы информантов распределены в соответствии с законом Ципфа [18, 19]. В результате, для последних проверяемых анкет из 35-40 слов приходилось размечать лишь 3-10, для которых информант проявил нестандартное мышление, попался в языковую ловушку или допустил грамматическую ошибку.

Наша гипотеза состояла в том, что при наличии параллельного текста на иностранном славянском языке, которого не знает пользователь, ответы должны быть более корректными, при этом рост точности ответов должен коррелировать с понятностью этого иностранного языка. Например, в контрольном тесте пользователи чаще должны вставлять некорректные или частично корректные ответы, тогда как при наличии когната в параллельном тексте ответы должны смещаться к большему числу корректных или частично корректных.

4. Результаты экспериментов

Итак, для каждого участника мы рассчитали среднюю корректность его ответов (понятность текста) исходя из системы оценок отдельных слов [0; 0.5; 1]. В таблице 1 приведены значения для общего числа участников, проходивших тест с параллельным текстом на данном языке, среднее значение понятности текста для участников, нижняя и верхняя граница 95% доверительного интервала для этого среднего. Доверительный интервал

рассчитывался по формуле $ci = 1.96 * s/\sqrt{n}$, где s — стандартное отклонение, а n — количество объектов в выборке. На рисунке 1 точками показаны сами значения понятности текста для отдельных информантов. Доверительные интервалы показаны линиями, цвет линий и точек зависит от языка параллельного текста. Темная полоса на рисунке 1 означает доверительный интервал для контрольного теста.

Для тестов с параллельным текстом была посчитана статистическая мощность разницы их средних значений с контрольным тестом. Статистическая мощность позволяет оценить, достаточно ли имеющееся количество информантов для того чтобы считать разницу статистически значимой. Мощность i-го теста рассчитывалась по следующей формуле: $B_i(\Theta) = 1 - \Phi\left(1.64 - \frac{v_i - v_c}{s_i/\sqrt{n_i}}\right)$, где v_i, v_c — средние значения для i-го и контрольного тестов, s_i — стандартное отклонение i-го теста, n_i- количество информантов в i-м тесте, а Φ— нормальное распределение. Значение статистической мощности, меньшие чем 0.05, означают, что количество информантов достаточно для того, чтобы утверждать статистическую значимость отличий на уровне 95% (задается константой 1.64). Результаты отдельных тестов и языков приведены в таблице 2 и на рисунках 2 и 3. Несимметричность доверительного интервала относительно среднего значения связана с округлениями.

Как видно из полученных данных, за доверительный интервал контрольного теста вышли все языки, кроме чешского и словацкого. Среднее значение для словацкого языка находится чуть выше верхней границы доверительного интервала контрольных тестов. Среднее значение для чешского языка находится внутри доверительного интервала контрольных тестов.

Из таблицы 2 видно, что тесты 5 и 6 являются самыми простыми — их средняя понятность на контрольном тесте составляет 0.7 и 0.68, соответственно. Если рассматривать результаты по языкам, то для украинского языка разница оказалась значимой на всех тестах; для белорусского — на всех, кроме четвертого; для болгарского — для половины тестов. Польский, чешский, словацкий и словенский показали значимый прирост лишь в одном тесте, сербский — в двух. При этом заметим, что на материале всех тестов польский, словенский и сербский показали статистически значимый прирост. Для чешского языка разница с контрольным тестом была отрицательной два раза (первый и третий тесты). Таким образом, можно утверждать, что в текстах на украинском, белорусском и болгарском языках всегда содержится подсказка для русскоязычного читателя, тогда как тексты на остальных языках необходимо рассматривать индивидуально. Также можно утверждать, что понятность конкретного текста на заданном языке зависит от некоторых параметров, разбору которых посвящен следующий раздел.

Таблица 1

Количество участников и понятность иностранного языка

Table 1

Amount of Participants and Foreign Language Intelligibility

	Контроль	**Укр**	**Бел**	**Болг**	**Пол**	**Чеш**	**Словацк**	**Серб**	**Словенск**
Число участников	283	243	280	233	291	231	243	227	227
Средняя понятность	0.62	0.78	0.74	0.73	0.67	0.64	0.66	0.71	0.68
Доверит. инт., мин	0.60	0.76	0.72	0.70	0.65	0.61	0.63	0.68	0.66
Доверит. инт., макс	0.64	0.80	0.76	0.75	0.70	0.66	0.68	0.73	0.71
Разница с контролем	-	0.16	0.12	0.11	0.05	0.02	0.04	0.09	0.06
Мощность разницы	-	0.00	0.00	0.00	0.00	0.73	0.17	0.00	0.00

Таблица 2

Количество участников и понятность иностранного языка по тестам

Table 2

Amount of Participants and Foreign Language Intelligibility by Tests

	Тест	Контроль	Укр	Бел	Болг	Пол	Чеш	Словацк	Серб	Словенск
Число участников	Тест 1	44	38	38	37	36	37	69	37	37
Средняя понятность		0.56	0.79	0.76	0.65	0.56	0.52	0.60	0.66	0.63
Доверит. инт., мин.		0.51	0.74	0.71	0.58	0.49	0.45	0.56	0.60	0.56
Доверит. инт., макс		0.61	0.83	0.81	0.71	0.64	0.58	0.65	0.71	0.69
Разница с контролем		-	0.23	0.20	0.09	0.00	-0.04	0.04	0.10	0.07
Мощность разницы		-	0.00	0.00	0.14	0.94	0.67	0.43	0.04	0.40
Число участников	Тест 2	46	37	38	36	44	38	59	37	37
Средняя понятность		0.58	0.75	0.76	0.71	0.67	0.61	0.65	0.61	0.65
Доверит. инт., мин.		0.53	0.69	0.71	0.66	0.62	0.55	0.60	0.54	0.69
Доверит. инт., макс		0.63	0.82	0.83	0.76	0.72	0.67	0.71	0.68	0.71
Разница с контролем		-	0.17	0.18	0.13	0.09	0.03	0.07	0.03	0.17
Мощность разницы		-	0.00	0.00	0.00	0.03	0.75	0.24	0.81	0.22
Число участников	Тест 3	37	37	37	37	38	37	39	31	38
Средняя понятность		0.60	0.75	0.73	0.72	0.65	0.57	0.69	0.80	0.72
Доверит. инт., мин.		0.55	0.70	0.70	0.67	0.59	0.49	0.63	0.75	0.66
Доверит. инт., макс		0.63	0.79	0.77	0.76	0.72	0.64	0.74	0.84	0.79
Разница с контролем		-	0.15	0.13	0.14	0.05	-0.03	0.09	0.20	0.12
Мощность разницы		-	0.00	0.00	0.00	0.40	0.84	0.03	0.00	0.01
Число участников	Тест 4	65	57	82	37	79	45	36	34	37
Средняя понятность		0.62	0.73	0.68	0.70	0.66	0.65	0.62	0.67	0.64
Доверит. инт., мин.		0.58	0.68	0.64	0.65	0.62	0.59	0.56	0.61	0.59
Доверит. инт., макс		0.65	0.79	0.72	0.76	0.70	0.70	0.69	0.73	0.69
Разница с контролем		-	0.11	0.06	0.08	0.04	0.04	0.00	0.05	0.02
Мощность разницы		-	0.01	0.07	0.09	0.25	0.71	0.92	0.41	0.76
Число участников	Тест 5	47	37	48	48	57	37	40	36	42
Средняя понятность		0.70	0.85	0.77	0.75	0.74	0.78	0.75	0.76	0.76
Доверит. инт., мин.		0.66	0.81	0.73	0.70	0.70	0.74	0.71	0.71	0.71
Доверит. инт., макс		0.74	0.90	0.81	0.81	0.79	0.82	0.79	0.80	0.81
Разница с контролем		-	0.15	0.07	0.05	0.04	0.08	0.05	0.06	0.06
Мощность разницы		-	0.00	0.02	0.44	0.45	0.02	0.25	0.21	0.19
Число участников	Тест 6	44	37	37	37	37	37		52	36
Средняя понятность		0.68	0.85	0.82	0.82	0.73	0.70		0.75	0.68
Доверит. инт., мин.		0.62	0.82	0.75	0.77	0.68	0.64		0.70	0.63
Доверит. инт., макс		0.73	0.88	0.88	0.86	0.79	0.75		0.80	0.74
Разница с контролем		-	0.17	0.14	0.14	0.05	0.03		0.07	0.00
Мощность разницы		-	0.00	0.00	0.00	0.43	0.84		0.11	0.92
Разница значимых / всех тестов			0.16 0.16	0.14 0.13	0.14 0.11	0.09 0.05	0.08 0.02	0.09 0.05	0.15 0.09	0.12 0.07
Значимых тестов			6	5	3	1	1	1	2	1

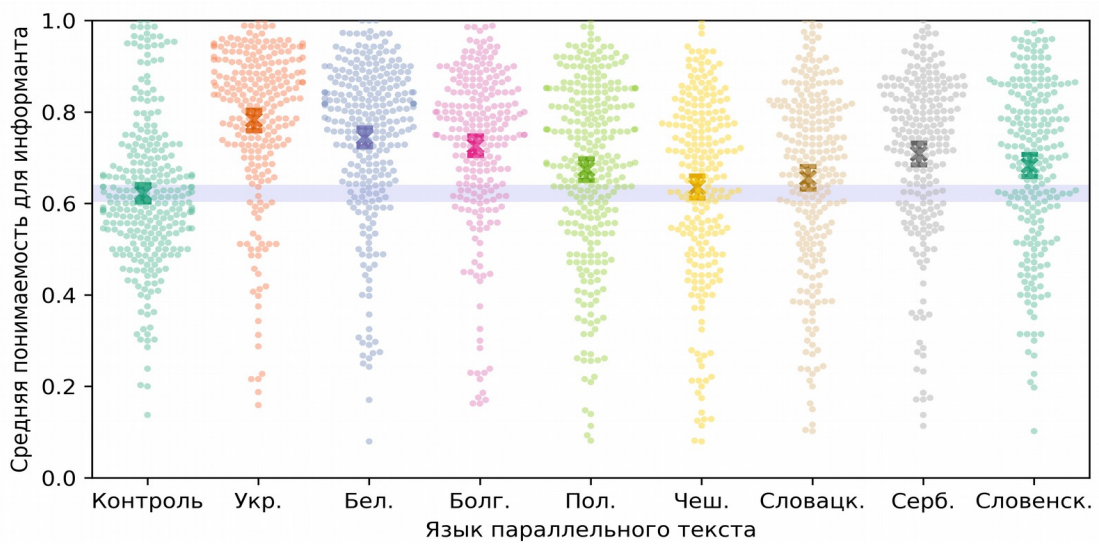

Рис. 1. Средняя понятность ответов пользователей и ее доверительные интервалы
Fig 1. Mean intelligibility of user's answers and its confidence interval

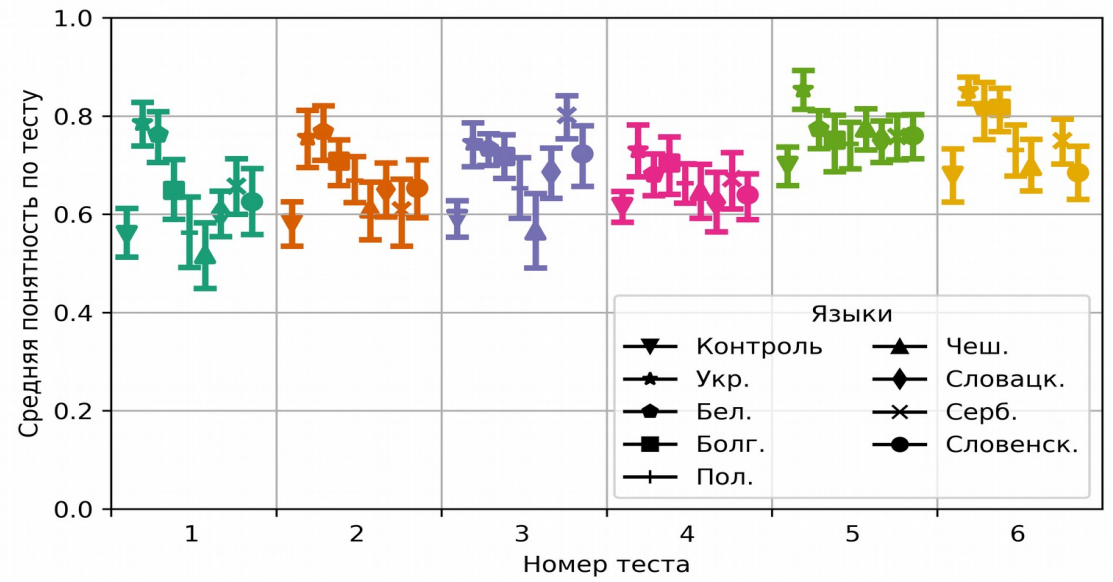

Рис. 2. Средняя понятность ответов пользователей по тестам с усреднением по языкам
Fig 2. Mean intelligibility of user's answers averaged by languages

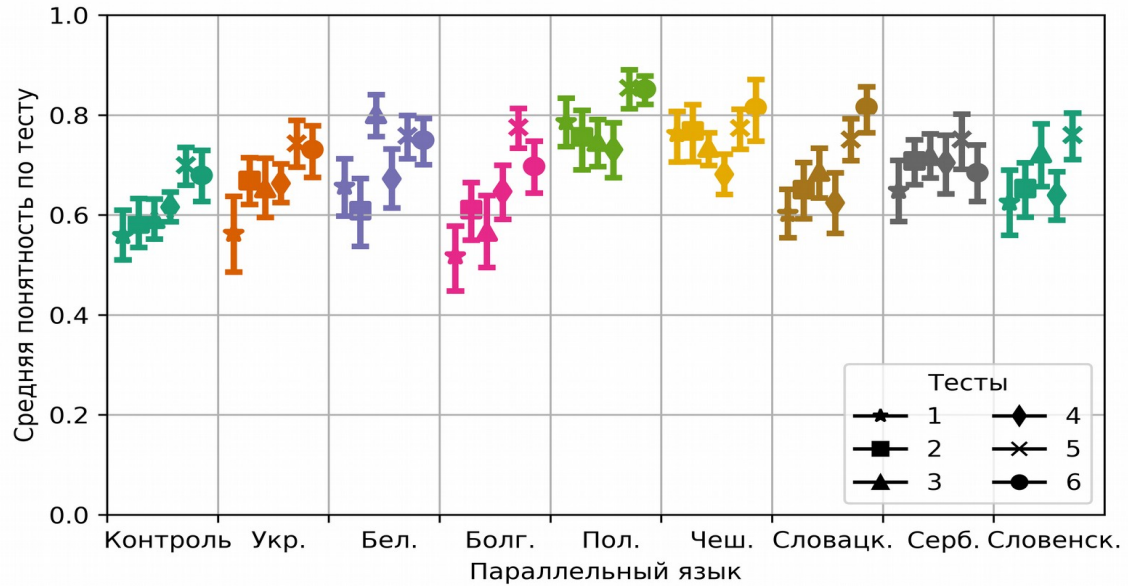

Рис. 3. Средняя понятность ответов пользователей по тестам с усреднением по тестам
Fig 3. Mean intelligibility of user's answers averaged by tests

5. Понятность отдельного текста

Мы предположили, что понятность текста зависит от таких параметров как очевидность восстановления слова по контексту и очевидность слова в параллельном тексте. Очевидность восстановления слова по контексту проверялась при помощи контрольного теста, в котором отсутствовал текст на параллельном языке. В Таблице 3 показано количество слов для каждого контрольного теста, средняя оценка которых превышала 0.8. Заметим, что оценка 0.8 означает, что соотношение числа полностью корректных, частично корректных и полностью некорректных ответов может принимать, например, следующие значения: 60% : 40% : 0%, 70% : 20% : 10% или 80% : 0% : 20%. То есть, средняя оценка для слова, близкая к 1, может служить некоторой оценкой процента пользователей, давших корректный ответ.

Как видно из Таблицы 3, в тесте 5 почти половина слов легко восстановимы; в тесте 6 таких слов примерно треть. В итоге эти тесты оказываются наиболее простыми для информантов, а средние значения на контрольном тесте — самые высокие. Однако самые низкие значения достигаются на тесте 1, в котором очевидных слов (которые корректно отметили 80% информантов) больше, чем в тесте 2, при том, что средняя понятность последнего выше.

Таблица 3

Количество слов теста со средней оценкой выше пороговой

Table 3

Amount of words which intelligibility is higher than threshold

Порог оценки	Тест 1	Тест 2	Тест 3	Тест 4	Тест 5	Тест 6
0.80	6	4	5	6	16	9
0.85	3	4	3	5	13	5
0.90	1	3	3	2	8	1
Всего слов в тесте	39	40	31	44	36	31

С точки зрения понятности иностранного текста, важным является понятие когната. В разных работах авторы выделяют несколько видов когнатов.

- Полные когнаты (или истинные друзья, Vrais Amis) — пары слов разных языков, имеющее сходное написание или звучание и одинаковое значение. При этом совпадение в написании и звучании может быть полным или неполным.
- Генетические когнаты — слова, имеющие единые исторические корни, но не имеющие полного сходства, например, варианты слова *ночь* для разных языков производят от общего праиндоевропейского **nekw-/*nokw-*: чешское, словацкое и польское *noc*, немецкое *Nacht*, английское *night* и греческое *νύχτα* признаются когнатами в той или иной степени [20].
- Частичные когнаты — когнаты, не разделяющие всего множества значений. Так, например, русское *фильм* хотя и происходит от английского *film*, однако не имеет английского значения «пленка».
- Ложные когнаты (или ложные друзья, Faux Amis) — пары слов, имеющих одинаковое или сходное написание или звучание, но не совпадающие значения. [9] Например, русские *пироги* и польские *pierogi* хотя и являются едой, но при этом — разными блюдами (русскому *пироги* соответствует польское *ciasta*, а польскому *pierogi* – русское *вареники*).

Некоторые авторы различают графические и фонетические когнаты. Первые имеют сходное написание и играют важную роль в понимании иностранного текста при чтении, однако требуют одинакового алфавита. Вторые имеют сходное звучание (то есть помогают при слушании речи) и не привязываются к единому алфавиту. При этом если читатель имеет представление о звуках, соответствующих сочетаниям букв в иностранном языке, фонетическое сходство слов может также оказывать помощь при чтении. На данный момент сложно развести влияние двух этих видов когнатов на понимание написанного текста.

Для оценки очевидности слова в параллельном тексте с точки зрения разных видов когнатов, мы разделили все слова параллельных текстов на четыре группы. Первая — однокоренные слова, имеющие сходное звучание в обоих языках (полные и частичные когнаты), например, *повернулся* (рус.) - *повернувся* (укр.) - *завярнуўся* (бел.) или *вырвался* (рус.) - *вырваўся* (бел.) - *wyrwał* (пол.). Вторая группа — слова, имеющие синоним или близкое по смыслу слово с тем же значением (ближе к генетическим когнатам), например, *дорогу* (рус.) - *пътя* (болг.) - *пут* (серб.) или *солдат* (рус.) - v*ojág* (чеш.) - *војник* (серб.). Третья группа — слова не имеющие общих корней или смыслов (не когнаты), например, *выхватил* (рус.) - *выхапіў* (бел.) - *vykřikl* (чеш.) или *трубою* (рус.) - *сурмою* (укр.) - *poľnicou* (словацк.). Четвертая группа — ложные друзья, имеющие сходное звучание с каким-то русским словом, но совершенно иной смысл в параллельном языке, например, *шее* (рус.) - *врата* (болг.) - *hřbetě* (чеш.) или *спиною* (рус.) - *плячыма* (бел.) - *гърба* (болг.).

Мы построили несколько списков слов в зависимости от вида когната и номера теста. Контрольные тесты считались отдельно от параллельных тестов. В каждый список помещалась средняя понятность слова в данном тесте, то есть одно и то же слово приняло участие в процедуре несколько раз. Далее для слов в одном списке вычислялись среднее значение и доверительный интервал. Результаты показаны на рис. 4. Каждая точка означает одно слово в одном тесте; средние значения обозначены крестиком; доверительный интервал отложен вверх и вниз отрезками. На правом рисунке приведены данные для контрольных тестов, на левом — для параллельных тестов. На левом рисунке для сравнения красным цветом приведены средние значения и доверительные интервалы контрольных тестов.

Как видно из рис. 5, слова, имеющие общие корни (полные и частичные когнаты), показывают существенный прирост в точности восстановления; сходные слова (генетические когнаты) также дают прирост, который оказывается статистически значимым лишь в двух тестах из шести. Большинство слов, не имеющих аналога в другом языке, показывают статистически незначимое снижение точности восстановления. Наконец, ложные друзья не показывают определенной тенденции, но точно не выходят за доверительные интервалы.

Для моделирования зависимости понятности текста от составляющих его слов мы построили графики зависимости понятности текста от процента слов той или иной группы (см. рис. 5, Табл. 4 и 5). Из рисунка видно, что понятность текста в целом положительно зависит от доли однокоренных слов (корреляция 0.72) и скорее отрицательно от доли сходных слов, слов без аналогов и ложных друзей (корреляция -0.43, -0.46 и -0.17, соответственно). При этом, если анализировать информацию по отдельным тестам (то есть с разбросом по языкам внутри одного теста), то для однокоренных слов корреляция колеблется от 0.58 до 0.91, для сходных — от -0.67 до 0.34 (с положительным значением для одного теста), слов без аналогов — от -0.84 до -0.00, и для ложных друзей переводчика — от -0.63 до 0.43 (с положительным значением для одного теста).

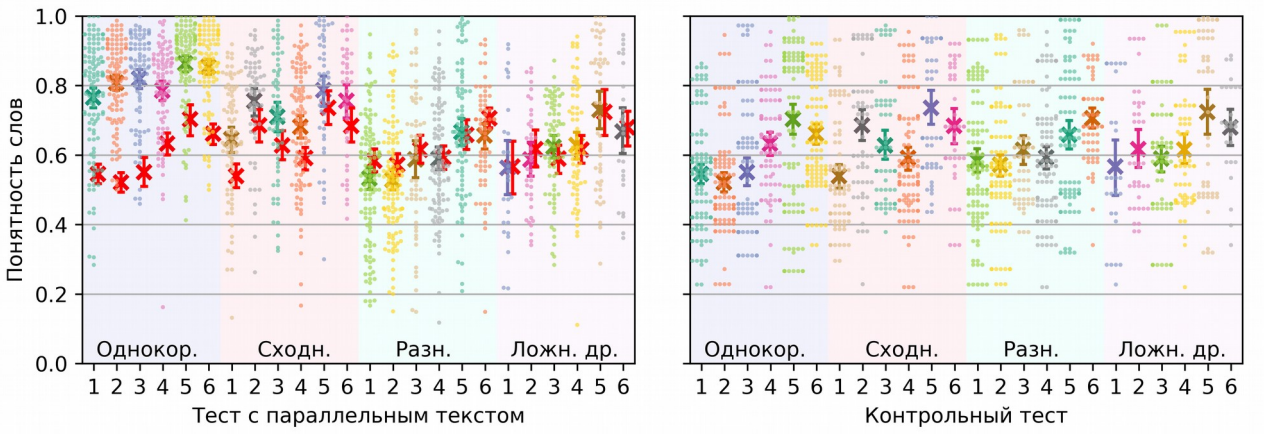

Рис. 4. Понимаемость слова теста от его типа
Fig 4. Intelligibility of a words on its type

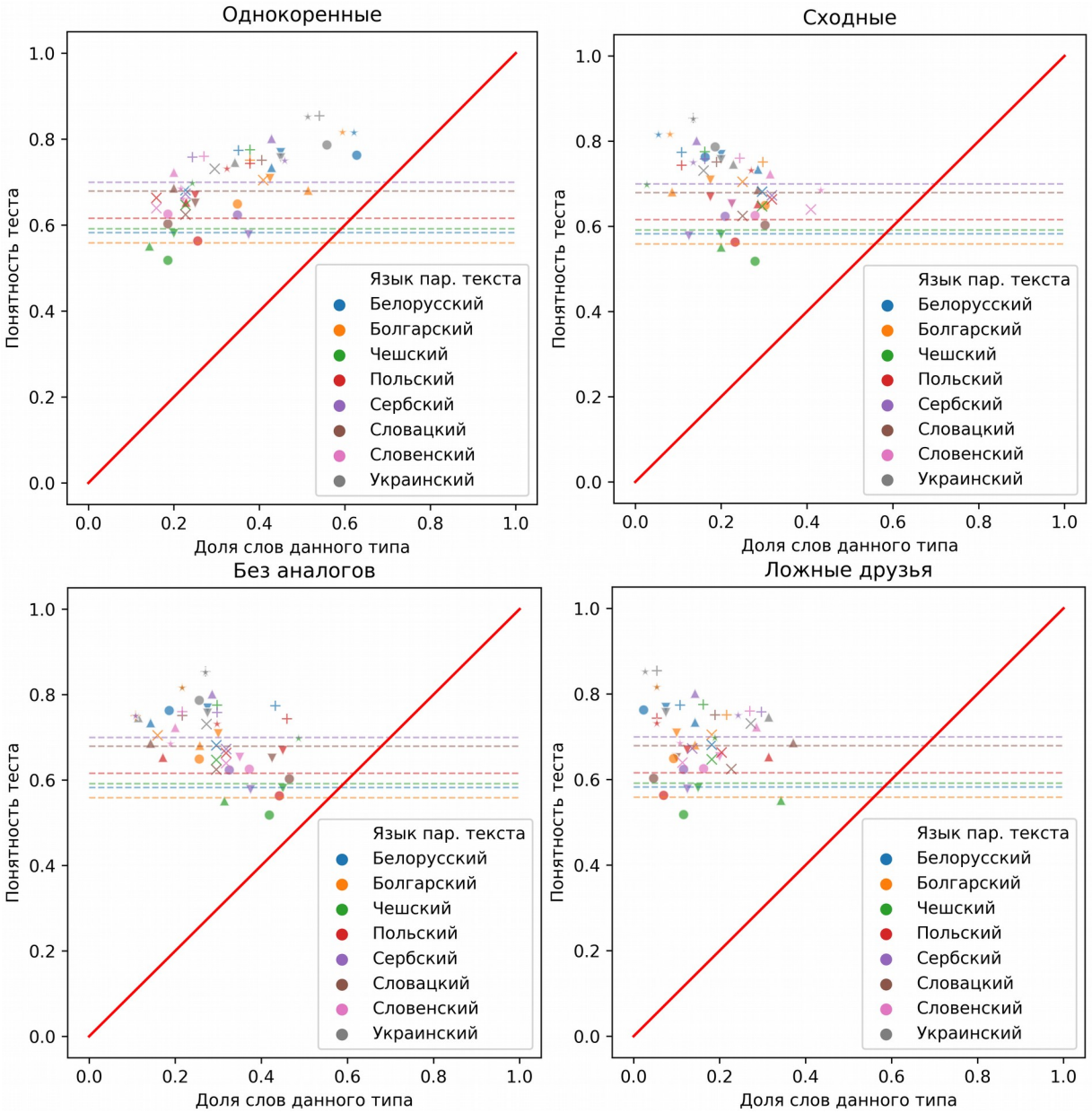

Рис. 5. Понятность текста от доли слов заданного типа сходства
Fig 5. Intelligibility of a test on the percentage of words of a given type

Для отдельных языков (то есть тесты внутри одного языка) корреляция колеблется для однокоренных слов от 0.18 для сербского до 0.91 для чешского (со средним 0.68), для сходных слов от -0.91 для белорусского до -0.20 для сербского (со средним -0.50), для слов без аналогов от -0.82 для словенского до 0.36 для украинского (со средним -0.30), и для ложных друзей переводчика от -0,80 для украинского до 0.59 для сербского (со средним -0.05). Используя подобную информацию, можно было бы вывести регрессионную формулу, предсказывающую понятность текста по долям слов разных видов сходства, однако точная настройка коэффициентов данной формулы потребует гораздо больше фактического материала. Помимо этого, следует вывести еще одну формулу, нелинейную, которая будет предсказывать величину возможной погрешности предсказания.

Исходя из результатов экспериментов, мы можем утверждать, что понятность текста на иностранном славянском языке обеспечивается по большей части за счет полных и частичных когнатов. При этом, мы не можем утверждать, что понятность текста обеспечивается только за счет них. Так, в сербском языке корреляция с ложными друзьями выше, чем с полными когнатами (0.59 против 0.18), а в словенском языке — сопоставима (0.73 и 0.75, соответственно). Для словацкого языка роль ложных друзей также высока (0.84

для полных когнатов и 0.42 для ложных друзей). Для украинского языка важную роль сыграли слова без аналогов (корреляция 0.36). Судя по всему, здесь сказывается общий уровень эрудиции, так как украинские слова более часто встречаются в обыденной жизни, то есть информанты могли узнать их в общем потоке.

Слова, имеющие сходные синонимы (генетические когнаты) для украинского и белорусского языка показали высокую степень антикорреляции (-0.73 и -0.91). Причиной этому может быть тот факт, что информанты, посчитав текст достаточно простым для понимания, ожидали точного совпадения большего количества слов и старались подогнать русский текст под найденные в иностранном тексте аналоги, жертвуя смыслом или точностью повествования. Другим полюсом может являться высокая корреляция числа ложных друзей с понятностью текста. Здесь информанты могли увидеть слишком сложный для понимания текст, который вынудил их более осторожно относиться к увиденным словам и более аккуратно встраивать их в текст на русском языке. Однако эта гипотеза требует более тщательной проверки.

Таблица 4

Корреляция между долей слов каждого типа и понятностью теста

Table 4

Correlation Between the Text Intelligibility and the Percentage of Word Types

Тип слова	**Тест 1**	**Тест 2**	**Тест 3**	**Тест 4**	**Тест 5**	**Тест 6**	**Все тесты**
Однокоренные	0.68	0.90	0.58	0.65	0.71	0.91	0.72
Сходные	0.34	-0.67	-0.03	-0.64	-0.31	-0.50	-0.43
Без аналогов	-0.77	-0.84	-0.35	-0.53	-0.00	-0.26	-0.46
Ложные друзья	-0.63	-0.45	-0.53	0.43	-0.47	-0.58	-0.17

Таблица 5

Корреляция между долей слов каждого типа и понятностью тестов языка

Table 5

Correlation Between the Language Intelligibility and the Percentage of Word Types

Тип слова	**Укр**	**Бел**	**Болг**	**Пол**	**Чеш**	**Словацк**	**Серб**	**Словенск**
Однокоренные	0.79	0.76	0.64	0.58	0.91	0.84	0.18	0.75
Сходные	-0.73	-0.91	-0.40	-0.33	-0.52	-0.69	-0.20	-0.21
Без аналогов	0.36	-0.01	-0.38	-0.05	-0.22	-0.70	-0.58	-0.82
Ложные друзья	-0.80	-0.79	-0.16	-0.22	-0.16	0.42	0.59	0.73

Анализ информации по тестам показывает, что в них также самую главную роль играет доля полных и частичных когнатов — все значения корреляции положительны и превышают 0.5. При этом в первом тесте кореляция доли сходных слов с понятностью текста также является положительной (равна 0.34), но для остальных тестов она отрицательная. В четвертом тесте положительной является корреляция доли ложных друзей (равна 0.43). Все остальные значения корреляции отрицательны, причем варьируются в широких пределах — от 0 до -0.91.

6. Обсуждение результатов и дальнейшие направления исследований

По результатам проведенных экспериментов видно, что понятность текстов одного и того же языка может серьезно отличаться. Как следствие, можно утверждать, что прежде чем говорить о понятности языка в целом, следует исследовать понятность отдельного текста. Разброс понятности отдельных текстов для украинского языка составляет 0.12 (от 0.73 до 0.85), для белорусского — 0.14, для болгарского — 0.17, для польского — 0.18, для чешского — 0.26, для словацкого — 0.15, для сербского — 0.19, для словенского — 0.13. Заметим, что этот разброс в несколько раз выше, чем полученные значения доверительных интервалов для тех же языков.

Одной из очевидных причин для таких отклонений является доля полных и частичных когнатов в иностранном тексте. Однако наши данные позволяют утверждать, что данный фактор не является единственным и полностью определяющим понятность текста. Положительная корреляция доли ложных друзей с понятностью текста для сербского, словацкого и словенского языков, а также слов без аналогов для тестов на украинском языке, показывают, что факторов может быть гораздо больше.

Роль влияния синтаксических особенностей языка в данном вопросе пока не ясна. Как показали наши предварительные исследования [21], русский синтаксис максимально отличается от болгарского и сербского, а наиболее похож на украинский. В такой ситуации возможна обратная корреляция сходства синтаксических структур и средней понятности текстов для языков разных групп. Влияние данного параметра должно быть изучено отдельно в ходе анализа сдвигов, произошедших при переводе. Быстрая проверка показала, что в местах, где пропущенные слова в оригинале и переводе менялись местами (ср., русск. «лица с *весело* оскаленными, *сверкающими* зубами» и польск. «twarze o białych, *połyskliwych*, *wesoło* wyszczerzonych zębach»), заметных изменений в точности ответов не прослеживается.

Исследование также не показало значимых различий между информантами, обладающими языковым образованием, и теми, кто им не обладает. При сравнении результатов, полученных на бакалавриантах, видно значительное улучшение для студентов языкового профиля. Однако уже начиная с магистратуры эта разница исчезает. Данному факту может быть дано два объяснения. Во-первых, в магистратуру часто приходят студенты с других специальностей. То есть, студенты могут отнести себя к людям, получающим языковое образования, не обладая при этом большим опытом специального обучения. С другой стороны, с возрастом, люди накапливают языковой опыт, путешествуя по разным странам и знакомясь с культурой других народов. Межнациональное общение до распада Советского Союза проходило более активно, что могло сказаться на лингвистическом опыте более старшего поколения. С этим может быть связан всплеск в корреляции слов без аналогов в украинском языке. Для украинского языка может быть еще одна причина. Информанты с Толоки атрибутировались системой по номеру телефона как граждане, проживающие на территории Российской Федерации. Но лица, проживающие на юго-западе нашей страны, могут иметь украинские корни и связи. Более того, в данном регионе распространены говоры, основанные на суржике, то есть смеси русского и украинского языков. Эти параметры также должны учитываться как языковой опыт, накапливаемый с годами.

Также, результаты данного исследования нуждаются в дальнейшей проверке с точки зрения когнитивных методов. Так, например, в работе [22] рассматривается метод разрешения полисемии для существительных русского языка, в том числе, и с использованием частотных характеристик. Дальнейшее исследование зависимости способности человека восстанавливать пропущенные слова от частотности их окружения может раскрыть скрытые параметры.

7. Заключение

В данной работе приведены результаты экспериментов по определению степени понятности иностранных славянских языков для неподготовленного русскоязычного читателя. В экспериментах использовались русские тексты с пропущенными словами, и параллельные тексты на украинском, белорусском, польском, чешском, словацком, сербском, словенском и болгарском языках. Контрольной группе параллельный текст не предъявлялся. Наша гипотеза состояла в том, что понятность текста на иностранном языке может быть измерена как превышение среднего числа правильных ответов при наличии параллельного текста над контрольной группой. Эксперименты проводились на шести фрагментах текстов. Всего в экспериментах приняло участие более 2200 информантов.

По результатам экспериментов языки расположились следующим образом (в порядке убывания понятности): украинский, белорусский, болгарский, сербский, словенский, польский, словацкий, чешский. Для чешского и словацкого языков превышение над

контролем не является статистически значимым. Более того, для чешского языка два теста из шести показали значения ниже контрольной группы.

Дальнейшие исследования показали, что понятность текста положительно коррелирует с долей полных и частичных конгнатов (на уровне 0.7), и отрицательно коррелирует с числом генетических когнатов, не когнатов (на уровне -0.4) и ложных друзей (-0.17). При этом разброс прироста понятности отдельных текстов колеблется от 0.12 до 0.26. Таким образом, можно говорить скорее о понятности отдельных текстов, чем о понятности языка в целом.

Результаты отдельных экспериментов показывают, что наличие генетических конгатов в параллельном тексте (не совпадающих слов, имеющих синонимичное или сходное значение в русском языке) также увеличивает понимаемость текста, но увеличение не носит статистически значимого значения. Таким образом, на понимаемость текста влияет не только наличие полных и частичных когнатов, но и некоторые другие параметры. Разбору этих параметров должно быть посвящено отдельное исследование.

Все результаты экспериментов и программное обеспечение для их анализа размещены по адресу https://github.com/klyshinsky/mutual_intelligibility_Russian .

Список литературы

1. Бейкер М. Атомы языка: грамматика в темном поле сознания. М.: Изд-во ЛКИ, 2008. 272 с.
2. Moberg, J., Gooskens, C., Nerbonne, J., Vaillette, N. Conditional Entropy Measures Intelligibility among Related Languages. *Lot Occasional Series*, 2007, Vol. 7, p. 51-66.
3. Gooskens, C. The Contribution of Linguistic Factors to the Intelligibility of Closely Related Languages. *Journal of Multilingual and Multicultural Development*, 2007, Vol. 28, no. 6, p. 445–467.
4. Golubović J., Gooskens, C. Mutual intelligibility between West and South Slavic languages. *Russ Linguist*, 2015, no. 39, p. 351–373.
5. Gooskens, C., Swarte F. Linguistic and extra-linguistic predictors of mutual intelligibility between Germanic languages. *Nordic Journal of Linguistics*, 2017. no. 40(2), p. 123–147. DOI: 10.1017/S0332586517000099
6. Kyjánek, L., Haviger, J. The Measurement of Mutual Intelligibility between West-Slavic Languages. *Journal of Quantitative Linguistics*, 2019, Vol. 26, Iss. 3, p. 205–230. DOI: 10.1080/09296174.2018.1464546
7. Keatley, C. W. History of bilingualism research in cognitive psychology. In R. J. Harris (Ed.), Cognitive processing in bilinguals, 1992, p. 15–49.
8. Grainger, J. Visual word recognition in bilinguals. In R. Schreuder, B. Weltens (Eds.), The bilingual lexicon, 1993, p. 11–26.
9. Lemhöfer, K., Dijkstra, T. Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision. *Memory & Cognition*, 2004, no. 32 (4), p. 533–550. DOI: 10.3758/BF03195845
10. Hammarström, H. Counting Languages in Dialect Continua Using the Criterion of Mutual Intelligibility. *Journal of Quantitative Linguistics*, 2008, Vol. 15, No. 1, p. 34–45
11. Коряков Ю.Б. Проблема «язык или диалект» и попытка лексикостатистического подхода // Вопросы языкознания. 2017. № 6. С. 79—101. DOI: 10.31857/S0373658X0003839-1
12. Heeringa, W. J. Measuring Dialect Pronunciation Differences using Levenshtein Distance. PhD thesis, Groningen, 2004, 315 p.
13. 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
14. Zarei, A. A., 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*, 2013, no. 4(1), p. 7–55.

15. Ackerman, P. L., Beier , M. E., Bowen, K. R. Explorations of crystallized intelligence: Completion tests, cloze tests, and knowledge. In: Learning and Individual Differences, Vol. 12, Issue 1, 2000, p. 105–121.

16. Ageeva, E., Tyers, F. M., Forcada, M. L., 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, 2015, p. 137-144.

17. Ягунова Е.В. Исследование избыточности русского звучащего текста // Труды института лингвистических исследований. 2010. Т. 4, часть 2. СПб: Наука. С. 90-114.

18. Ferrer i Cancho, R. The variation of Zipf's law in human language. *The European Physical Journal B – Condensed Matter and Complex Systems*, 2005, no 44, p. 249-257.

19. Кочеткова Н.А., Клышинский Э.С., Ермаков П.Д. Подчиняются ли составные конструкции закону Ципфа? // Системный администратор. 2016. № 11. С. 89-95.

20. Bouckaert, R., Lemey, P., Dunn, M. et al. Mapping the Origins and Expansionof the Indo-European Language Family. *Science*, Vol. 337, 2012, p. 957-960

21. Klyshinskiy, E., Karpik, O. V. Quantitative Evaluation of Syntax Similarity *Mathematica Montisnigri*, 2019, Vol. XLVI, p. 123–132. DOI: 10.20948/mathmon-2019-46-11

22. Lopukhina, A., Lopukhin, K., Nosyrev. G. Automated word sense frequency estimation for Russian nouns. In: O. Lyashevskaya, M. Kopotev, A. Mustajoki (Eds.), Quantitative approaches to the Russian language. Routledge, 2018, p. 79–94. DOI: 10.4324/9781315105048-4

Reference

1. Backer M. The Atoms of Language: The Mind's Hidden Rules of Grammar. Oxford University Press, 2001, 276 p.

2. Moberg, J., Gooskens, C., Nerbonne, J., Vaillette, N. Conditional Entropy Measures Intelligibility among Related Languages. *Lot Occasional Series*, 2007, Vol. 7, p. 51-66.

3. Gooskens, C. The Contribution of Linguistic Factors to the Intelligibility of Closely Related Languages. *Journal of Multilingual and Multicultural Development*, 2007, Vol. 28, no. 6, p. 445–467.

4. Golubović J., Gooskens, C. Mutual intelligibility between West and South Slavic languages. *Russ Linguist*, 2015, no. 39, p. 351–373.

5. Gooskens, C., Swarte F. Linguistic and extra-linguistic predictors of mutual intelligibility between Germanic languages. *Nordic Journal of Linguistics*, 2017. no. 40(2), p. 123–147. DOI: 10.1017/S0332586517000099

6. Kyjánek, L., Haviger, J. The Measurement of Mutual Intelligibility between West-Slavic Languages. *Journal of Quantitative Linguistics*, 2019, Vol. 26, Iss. 3, p. 205–230. DOI: 10.1080/09296174.2018.1464546

7. Keatley, C. W. History of bilingualism research in cognitive psychology. In R. J. Harris (Ed.), Cognitive processing in bilinguals, 1992, p. 15–49.

8. Grainger, J. Visual word recognition in bilinguals. In R. Schreuder, B. Weltens (Eds.), The bilingual lexicon, 1993, p. 11–26.

9. Lemhöfer, K., Dijkstra, T. Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision. *Memory & Cognition*, 2004, no. 32 (4), p. 533–550. DOI: 10.3758/BF03195845

10. Hammarström, H. Counting Languages in Dialect Continua Using the Criterion of Mutual Intelligibility. *Journal of Quantitative Linguistics*, 2008, Vol. 15, No. 1, p. 34–45

11. Koryakov, Yu. B. Language vs. Dialect: a Lexicostatistic Approach. *Voprosy Jazykoznanija*, 2017, no 6, p. 79—101. (In Russ.) DOI: 10.31857/S0373658X0003839-1

12. Heeringa, W. J. Measuring Dialect Pronunciation Differences using Levenshtein Distance. PhD thesis, Groningen, 2004, 315 p.

13. Klyshinsky, E. S., Logacheva, V. K., Belobokova, Yu. A. Foreign text intelligibility: case of Slavic language group. *Keldysh IAM Preprints*, 2017, no 13, p. 23. DOI:10.20948/prepr-2017-13 (In Russ.)

14. Zarei, A. A., 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*, 2013, no. 4(1), p. 7–55.

15. Ackerman, P. L., Beier , M. E., Bowen, K. R. Explorations of crystallized intelligence: Completion tests, cloze tests, and knowledge. In: Learning and Individual Differences, Vol. 12, Issue 1, 2000, p. 105–121.

16. Ageeva, E., Tyers, F. M., Forcada, M. L., 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, 2015, p. 137-144.

17. Yagunova, E. V., Investigation of redundancy in oral Russian texts. *Acta Linguistica Petropolitania,* 2010, Vol. 4, part 2, p. 90-114. (In Russ.)

18. Ferrer i Cancho, R. The variation of Zipf's law in human language. *The European Physical Journal B – Condensed Matter and Complex Systems*, 2005, no 44, p. 249-257.

19. Kochetkova, N. A., Klyshinsky, E. S., Ermakov, P. D. Do Collocations Meet the Zipf"s Law? (Podchinyayutsya li sostavniye konstrukcii zakonu Zipfa?). *System Administrator*, 2016, no. 11, p. 89-95.

20. Bouckaert, R., Lemey, P., Dunn, M. et al. Mapping the Origins and Expansionof the Indo-European Language Family. *Science*, Vol. 337, 2012, p. 957-960

21. Klyshinskiy, E., Karpik, O. V. Quantitative Evaluation of Syntax Similarity *Mathematica Montisnigri*, 2019, Vol. XLVI, p. 123–132. DOI: 10.20948/mathmon-2019-46-11

22. Lopukhina, A., Lopukhin, K., Nosyrev. G. Automated word sense frequency estimation for Russian nouns. In: O. Lyashevskaya, M. Kopotev, A. Mustajoki (Eds.), Quantitative approaches to the Russian language. Routledge, 2018, p. 79–94. DOI: 10.4324/9781315105048-4

Клышинский Эдуард Станиславович, кандидат технических наук, доцент школы лингвистики Национального исследовательского университета «Высшая школа экономики» (ул. Мясницкая д. 20, Москва, 101000, Россия)

Eduard S. Klyshinsky, National Research University Higher School of Economics (20 Myasnitskaya Str., Moscow, 101000, Russian Federation)

eklyshinsky@hse.ru

ORCID 0000-0002-4020-488X