Linguistic Problems Based on Text Corpora

Boris Iomdin
V.V. Vinogradov Institute of Russian Language, Russian Academy of Sciences (Moscow)
National Research University Higher School of Economics (Moscow, Russia)
iomdin@ruslang.ru

Alexander Piperski
M.V. Lomonosov Moscow State University (Moscow, Russia)
Russian State University for the Humanities (Moscow, Russia)
apiperski@gmail.com

Anton Somin
Russian State University for the Humanities (Moscow, Russia)
somin@tut.by

Abstract

The paper is focused on self-contained linguistic problems based on text corpora. We argue that corpus-based problems differ from traditional linguistic problems because they make it possible to represent language variation. Furthermore, they often require basic statistical thinking from the students. The practical value of using data obtained from text corpora for teaching linguistics through linguistic problems is shown.

1 Introduction

The genre of self-contained linguistic problems appeared long before the onset of corpus linguistics. The authors of most problems either constructed phrases or sentences on their own, or (much less commonly) used some real texts (e.g. excerpts from ancient manuscripts). Now that text corpora become widespread, they offer new possiblities for problem composing. This paper gives examples of such problems offered to high school students in Russia at recent linguistic olympiads. We comment on the ways such new problems are solved and show how the data obtained from text corpora and linguistic problems based thereon could be used for teaching linguistics to high school students.

We deliberately include some of the original Russian versions of the problems alongside with their English translations (made specially for this paper and never published before), so that (1) those familiar with the Russian language could use the problems for training, and (2) issues of linguistic problem translation could be illustrated, too: translations of linguistic problems are not always equivalent to the originals (cf. Derzhanski et al. 2004).

2 Corpus-based problems and traditional problems: what is the difference?

The most straightforward way of using corpora for composing problems is to find real-world examples of some linguistics phenomena. If the problem deals with a language other than its author's native tongue, it is preferable to constructing phrases or sentences (unfortunately, when experts or native speakers look at constructed data in the problems assigned at some earlier contests, they sometimes find it non-idiomatic, infelicitous or even ungrammatical). If the problem illustrates some phenomenon in the native language of its author, it is also preferable to use corpus examples, because they do not impose the author's introspection upon students.

Some corpus-based problems are quite different from traditional linguistic problems. Corpus data allow to present linguistic variation in a problem, which was difficult to do before the corpora era. It might be diachronic variation, register variation or some other kind of variation.

The traditional linguistic problems require a strictly deterministic way of thinking ("if this, then that"). However, in real life linguists often have to deal with statistical patterns, and this is where corpus linguistics comes into play. Problems based on corpus data may exemplify this approach.

Proceedings of the Fourth Workshop on Teaching Natural Language Processing, pages 9–17,
Sofia, Bulgaria, August 4-9 2013.

3 Some corpus-based problems

3.1 Corpus examples illustrating a linguistic phenomenon in the solvers' native language

Problem #1 (composed by Boris Iomdin)

При изучении фраз с глаголом *предлагать* естественно выделять два типа употреблений этого глагола. Ниже приведены примеры обоих типов употреблений.

I

1. *Дядя Владимир предложил трактирщику его заменить и поторговать за него* (А. Левицкая);
2. *Ходил на вокзал, предлагал пассажирам помочь снести вещи* (А. Пантелеев);
3. *А Плахотников предложил стать моим руководителем и поделиться всем, что знал сам как опытнейший дрессировщик* (В. Запашный);
4. *Узнав, что друг его плохо себя чувствует <…>, Диккенс с трогательной заботливостью предлагает приехать и помочь ему в работе* (М. Шагинян);
5. *Ира оказалась очень чутким и отзывчивым человеком и сразу же предложила Евгению Александровичу переехать к нему и ухаживать за его полупарализованной матерью* (О. Демьянова).

II

6. *Встретивший его адъютант предложил ему располагаться и ждать* (К. Симонов);
7. *Предлагаем читателям разработать, изготовить и испытать такое приспособление* (Б. Синельников);
8. *Но пришла сестра и предложила уйти, дать ему отдохнуть* (Л. Бронтман);
9. *Он предложил мне дать в их издательство книжку стихов* (А. Городницкий);
10. *Ире как человеку чуткому и отзывчивому было тяжело смотреть на страдания любимого человека, и она предложила продать квартиру и переехать жить к ней* (О. Демьянова).

Задание 1. Объясните, чем различаются эти два типа употреблений.

Задание 2. К какому типу употреблений можно отнести следующие примеры:

11. *Он предложил Мижуеву дать денег на это дело, и Мижуев радостно согласился* (М. Арцыбашев);
12. *Через два дня позвонили со студии и предложили приехать и заключить договор* (Л. Вертинская);
13. *И вот дедушка Рахленко предлагает Кусиелу прогнать мерзавца приказчика и вместо него взять моего отца* (А. Рыбаков);
14. *Я предложила Ире помочь деньгами, но она сказала, что у них есть на жизнь* (З. Масленикова).

Если в каких-то случаях Вы считаете, что возможны оба ответа, укажите это. Поясните Ваше решение.

Задание 3. Что Вы можете сказать о следующем примере:

15. *Он даже предложил мне давать Юре уроки французского языка и платить за урок тарелкой супа* (В. Гроссман)?

English translation:

Looking at sentences with the Russian verb *predlagat'* 'to offer, to suggest', one finds out that it can be used in two different ways. Consider some examples for both (the Russian verb in question is replaced by a fictional English verb *to predle*):

I

1. *Uncle Vladimir **predled** the barman to replace him and to trade for him for a while* (A. Levitskaya);
2. *He used to come to the railway station and **predle** the passengers to carry their luggage* (A. Panteleev);
3. *Plakhotnikov **predled** to become my instructor and to share with me everything he knew as a very experienced animal tamer* (V. Zapashny);
4. *As soon as he learned that his friend was sick, Dickens, with a touching affection, **predles** to come and help him in his work* (M. Shaginyan);
5. *Ira appeared to be a very considerate and sympathetic person: right away she **predled** to Evgeny Alexandrovich to settle at his place and to take care of his half-paralyzed mother* (O. Demyanova).

II

6. *The aide-de-camp who met him **predled** him to sit down and wait* (K. Simonov);
7. *We **predle** the readers to work out, make and test such a device* (B. Sinel'nikov);
8. *But a nurse came and **predled** to leave, to give him some rest* (L. Brontman);
9. *He **predled** me to submit a verse book to their publishing house* (A. Gorodnitsky);
10. *Ira, as a considerate and sympathetic person, found it hard to watch how her beloved one was suffering, so she **predled** to sell the apartment and to settle with her* (O. Demyanova).

1. Explain the difference between the two usages of the verb.

2. Consider the following examples. What can you say about the ways the verb *to predle* is used in them?

11. *He **predled** Mizhuev to give money for the cause, and Mizhuev gladly agreed* (M. Artsybashev);
12. *After two days, someone called from the studio and **predled** to come and sign a contract* (L. Vertinskaya);
13. *So, old Rakhlenko **predles** Kusiel to kick out the villain clerk and to employ my father instead* (A. Rybakov);
14. *I **predled** Ira to assist financially, but she told me that they had enough for a living* (Z. Maslenikova).

If in some cases you believe that both answers are possible, give both and explain.

3. What can you say about the following example:

15. *He even **predled** me to give French lessons to Yura and to pay with a plate of soup for each lesson* (V. Grossman)?

Solution of Problem #1

It can be seen that the verb *predlagat'* (or the artificial English verb *to predle*) may govern two types of infinitives: subject infinitives, as in I, and object infinitives, as in II. These two types of usage may represent two different senses of the verb *predlagat'*, which could roughly be translated as 'to offer (to do something)' and 'to suggest (that someone else does something)'. The most interesting thing here is that in many cases it is quite hard to determine whether the infinitive refers to the subject or to the object. In fact, both answers are possible in all examples 11–14. Example (15) seems infelicitous: the author clearly meant that 'me' is to give lessons and 'he' is to pay. See (Iomdin & Iomdin 2011) for further discussion.

Comment

This problem does not differ much from traditional linguistic problems. The solvers are offered textual examples from their native language and are requested to analyze them. However, it is important that these examples are not constructed, which makes the problem more trustworthy. No solver can say "I don't use this verb this way!" because s/he is confronted with real-life examples.

Problem #2 (composed by Boris Iomdin)

Один лингвист попросил знакомого голландца, хорошо знающего русский язык, перевести на голландский несколько отрывков из литературных произведений. Его интересовало, какие глаголы голландец использует для перевода тех русских глаголов, которые в приведённых ниже цитатах подчёркнуты. Варианты, предложенные голландцем, указаны в скобках.

1) *За пустую бутылку охотно отдавали свои огромные тростниковые шляпы. Все у нас наменяли* **(krijgen)** *этих шляп* (И. Гончаров).

2) *И вот работяга, отработав 12-14 часов в смену, моет полы ночью за эти две папиросы табаку. И ещё считает за счастье – ведь на табак он выменяет* **(krijgen)** *хлеб* (В. Шаламов).

3) *Книжек я наменял* **(krijgen)** *у мужиков, на курево хотели, да бумага толстая* (Л. Леонов).

4) *Кондуктор того трамвая отнёс пальто на барахолку и там обменял* **(ruilen)** *на сметану, крупу и помидоры* (Д. Хармс).

5) *На пирожные он выменивал* **(ruilen)** *хлеб, муку, масло, пшено, табак – весь состав своего пайка, за исключением сахара: сахар он оставлял себе* (В. Ходасевич).

6) *Наменяли* **(krijgen)** *пятаков, полчаса дозванивались* (М. Веллер).

7) *Но если портному не нужна груша, а нужен, к примеру сказать, стол, то вы должны пойти к столяру, дать ему грушу за то, что он сделает стол, а потом этот стол выменять* **(ruilen)** *у портного на брюки* (Н. Носов).

8) *Но, как видно, руссы сильно желали выменивать* **(krijgen)** *на свои товары арабские монеты, диргемы, которые везде и во всяком значении имели большую ценность* (С. Соловьёв).

9) *Он бежал ночью из Дырок прямо на квартиру к Учителю, винтовку обменял* **(ruilen)** *на две бутылки самогона и в пьяном виде декламировал "Клеветникам России"* (И. Эренбург).

10) *Этот кальян выменял* (**krijgen**)*, или, правду сказать, выманил я у английского путешественника* (А. Бестужев-Марлинский).

11) *...Толпою окружённый слуг;*
Усердствуя, они в часы вина и драки
И честь и жизнь его не раз спасали: вдруг
На них он выменял **(krijgen)** *борзые три собаки!* (А. Грибоедов).

12) *Я сегодня, гражданин,*
Плохо спал:
Душу я на керосин
Обменял **(ruilen)** (В. Зоргенфрей).

Задание 1. Дано ещё несколько цитат. Определите, какие глаголы использовал голландец для перевода каждого из подчёркнутых глаголов. Если в каких-то случаях Вы не можете выполнить задание с уверенностью, отметьте это. Поясните Ваше решение.

А. *В доме было уже продано, выменяно на продукты всё, что можно. Кроме пианино* (И. Грекова).

Б. *Вот он сидит за большим столом и кладёт резолюции на подносимых бумагах: "От-ка-*

зать!!!" Вы хотите обменять что-то меньшее на что-то большее. Отказать! Вы хотите обменять что-то большее на что-то меньшее. Отказать! (В. Войнович).

В. *Выменять ножик на удочку* (С. Ожегов, Н. Шведова, Толковый словарь русского языка).

Г. *Он два дня не ел хлеба, затем выменял на хлеб большой фибровый чемодан* (В. Шаламов).

Д. *Теперь следует вопрос: как добывали руссы свои северные товары? Конечно, они могли выменивать их у туземцев на какие-нибудь произведения греческой промышленности…; но главным источником приобретения были дани и потом охота* (С. Соловьёв).

Е. *Я, наверное, не зря*
В этот раз ходил в моря,
Наменял я там подарков,
Ждёт их вся моя родня (О. Газманов).

Задание 2. Опираясь на материал задачи, сформулируйте правила употребления голландских глаголов *krijgen* и *ruilen*.

English translation:

A linguist asked a speaker of Dutch familiar with the Russian language to translate several excerpts from Russian books into Dutch. He was interested in the verbs the Dutch speaker would use as equivalents for certain Russian verbs: *namenjat'* (hereafter 'to N'), *obmenjat'* (hereafter 'to O') and *vymenjat'* (hereafter 'to V'), all associated with the idea of exchange. The equivalents used by the Dutch speaker are given in brackets.

1. *They were willing to give away their enormous reed hats for an empty bottle. We all N-ed* **(krijgen)** *these hats* (I. Goncharov).
2. *So the drudge, having worked for 12 to 14 hours in a shift, washes the floors during the night for these two cigarettes. And he even is happy with it, since he will V* **(krijgen)** *bread for the tobacco* (V. Shalamov).
3. *I N-ed* **(krijgen)** *some books from the peasants, they wanted to use them for smoking, but the paper's too thick* (L. Leonov).
4. *The tram conductor took the coat to the flea market and there he O-ed* **(ruilen)** *it for sour cream, groats and tomatoes* (D. Kharms).
5. *For pastry he V-ed* **(ruilen)** *bread, flour, butter, millet, tobacco, – all his ration, except for sugar, which he always kept* (V. Khodasevich).
6. *They N-ed* **(krijgen)** *five kopeck coins and tried to establish a telephone connection for half an hour* (M. Veller).
7. *But if the tailor does not need pears, but needs a table for example, then you have to go to the woodworker, give him a pear for the table he should make, and then V* **(ruilen)** *this table for trousers with the tailor* (N. Nosov).
8. *But apparently the Russians really wished to V* **(krijgen)** *for their goods Arabian coins, dirhem, which had a great value everywhere* (S. Solovyov).
9. *At night, he ran away from the Holes directly to the Teacher's apartment, O-ed* **(ruilen)** *the gun for two bottles of alkie, and being drunk recited "To the slanderers of Russia"* (I. Erenburg).
10. *I V-ed* (**krijgen**)*, or, to say the truth, juggled out this water-pipe from a British traveler* (A. Bestuzhev-Marlinsky).
11. *Had a team of loyal servants / That during fight-and-drinking rounds / Had saved his life and honour, but then once / He suddenly V-ed* **(krijgen)** *them for three hounds.* (A. Griboyedov, translation by A. Vagapov)
12. *Today I slept bad, citizen: I O-ed* **(ruilen)** *my soul for kerosene* (V. Sorgenfrei).

1. Consider some more examples. Determine which verbs the Dutch speaker would use in these examples. If in some cases you believe several solutions are possible, explain.

A. *In the house, everything was already sold, V-ed for food. Except for the piano* (I. Grekova).

B. *Here is he sitting at a large table, writing his decisions on the documents he is given: "Refu-sal!!!" You want to O something smaller for something larger. Refusal! You want to O something larger for something smaller. Refusal!* (V. Voynovich).

C. *V a knife for a fishing rod* (Ozhegov and Shvedova Explanatory Dictionary).

D. *For two days, he had not been eating bread, and then he V-ed for bread a large wooden chest* (V. Shalamov).

E. *Now comes a question: how did the Russians get their Northern goods? Of course they could V them from the locals for some Greek manufactured goods …; but the main source for them were imposts, and then hunting* (S. Solovyev).

F. *Probably not in vain / Was I at sea this time / I N-ed presents there, / And my whole family is waiting for them* (O. Gazmanov).

2. Based on the data in the problem, explain what the Dutch verbs *krijgen* and *ruilen* mean.

Solution of Problem #2

All three verbs in question are used with objects. With N (*namenjat'*), the object always signifies the thing obtained in the exchange; with O (*obmenjat'*), the object always signifies the thing given away in the exchange. As for V (*vymenjat'*), it is used in both ways (which is a very rare case of a double government pattern). The Dutch verb *krijgen*, therefore, means 'to get in return', and *ruilen* means 'to give away'.

Comment

This problem is similar to problem #1 in the way it exploits corpus data., The examples come from real texts (and even the Dutch speaker is real), and the only difference is the reference to a foreign language which makes it easier to see the different senses of the words in the solvers' native language.

3.2 Corpus examples illustrating diachronic variation

Problem #3 (composed by Boris Iomdin)

Слово *параллельно* в литературном языке может управлять как существительными в дательном падеже (конструкция А), так и существительными в творительном падеже с предлогом *с* (конструкция Б). Ниже даны примеры обеих конструкций:

1. *Васич увидел лощинку. Она шла* ***параллельно*** *немцам, преграждала им путь к дивизиону* (Г. Бакланов).
2. *Её путь лежал* ***параллельно*** *маршруту трамвая* (Б. Пастернак).
3. *Мне было неприятно, что какие-то люди* ***параллельно*** *с нами, по обе стороны от нас, пробираются на холм* (Ф. Искандер).
4. *Теперь они шли* ***параллельно*** *насыпи* (А. и Б. Стругацкие).
5. *Наматывая мили на кардан, / Я еду* ***параллельно*** *проводам* (В. Высоцкий, 1971).
6. *Намечались короткие летние, перед отпуском, гастроли в Риге* ***параллельно*** *с работой двух московских сцен* (С. Пилявская).
7. *Новые восьми-девятиэтажные дома стояли разомкнутым строем* ***параллельно*** *бульвару* (Ю. Даниэль).
8. *Отросший ус торчал уже не* ***параллельно*** *земной поверхности, а почти перпендикулярно, как у пожилого кота* (И. Ильф, Е. Петров).
9. *Мы глядели на некоторые беседки и храмы по высотам, любовались длинною, идущею* ***параллельно*** *с берегом кедровою аллеею* (И. Гончаров).
10. ***Параллельно*** *с монтажом идёт и отделка фасада* («Комсомольская правда»).
11. *С каждым годом заводскому населению приходится тяжелее, а* ***параллельно*** *с этим возвышается благосостояние управителей, управляющих, поверенных и целого сонма служащего люда* (Д. Мамин-Сибиряк).

Задание 1. В современном русском языке можно усмотреть некоторую тенденцию, в соответствии с которой в одних случаях употребляется конструкция А, а в других – конструкция Б. Объясните, в чём состоит эта тенденция. Все ли примеры, приведённые выше, ей соответствуют? Если нет, с чем, по Вашему мнению, это может быть связано?

Задание 2. Раскройте скобки, использовав либо конструкцию А, либо конструкцию Б. Если в каких-то случаях выбор конструкции вызывает у Вас сомнения, отметьте это:

А. *Как водится, параллельно* ***(бумажная война)*** *происходила чехарда с собраниями акционеров* («Вечерняя Москва»)

Б. *Здесь, на советской территории, у самой границы и параллельно* ***(она)*** *проходит Августовский канал* (В. Суворов)

В. *Комната Франца выходила на улицу, шедшую параллельно* ***(набережная)*** (В. Набоков)

Г. *Они* [*фабрично-заводские комитеты*] *существовали параллельно* ***(профсоюзы)*** *и объединились с ними в 1918 г.* (Большая советская энциклопедия)

Д. *Скажем, в шекспировском* «*Короле Лире*» *сюжетная линия Лира развивается параллельно* ***(линия Глостера)*** (Т. Шабалина).

Задание 3. В последнее время в публицистике слово *параллельно* стало иногда употребляться ещё в одной конструкции – с предлогом *от*:

Е. *Религия должна существовать* ***параллельно от*** *гражданского общества.*

Ж. *Эта сторона жизни существовала где-то* ***параллельно от*** *меня и меня не затрагивала.*

З. *Наше государство всё ещё живёт* ***параллельно от*** *своих граждан.*

Попробуйте объяснить причины возникновения такой конструкции.

English translation:

The Russian word *parallel'no* 'in parallel' can govern nouns in dative case ('parallel to', construction A) as well as in instrumental case ('parallel with', construction B). Consider examples for both constructions.

1. *Vasich saw a small hollow. It was going* ***in parallel to*** *the Germans, blocking their way to the squadron* (G. Baklanov, 1961).
2. *Her way was to be* ***in parallel to*** *the tram route* (B. Pasternak, 1945–55).
3. *I did not like it that some people were climbing the hill* ***in parallel with*** *us, on both our sides* (F. Iskander, 1990).

4. *Now they were going **in parallel to** the embankment* (A. and B. Strugatsky, 1971).
5. *Winding up miles onto the cardan, I am driving **in parallel to** the wires* (V. Vyssotsky, 1971).
6. *A short summer tour in Riga was planned before the vacation, **in parallel with** the operation of the two theater stages in Moscow* (S. Pilyavskaya, 2001)
7. *New eight to nine-storey houses were standing in a broken line, **in parallel to** the boulevard* (Yu. Daniel, 1962).
8. *His much grown whisker was no longer sticking out **in parallel to** the ground surface, but was almost perpendicular to it, as if he was an old cat* (I. Il'f, E. Petrov, 1927–8).
9. *We were looking at some gazebos and churches above, enjoying a long cedar-tree alley, which went **in parallel with** the shore* (I. Goncharov, 1855–7).
10. *Finishing the façade is progressing **in parallel with** the mounting* («Komsomol'skaya Pravda» newspaper, 2007).
11. *Each year, the plant laborers find their life to be harder and harder, and the well-being of the managers, lawyers and a whole bunch of employees grows **in parallel with** it.* (D. Mamin-Sibiryak, 1874–5).

1. There is a tendency in modern Russian to use construction A and construction B in different cases. What is the difference? Does every example above comply with this tendency? If not, what could be the reason?

2. Choose either Construction A or Construction B for the following examples. In case you have difficulty with some sentences, explain why.

A. *As usual, a cockalorum with share holders meetings was happening in parallel to / with the paperwork war* («Evening Moscow» newspaper, 2007).

B. *Here, on Soviet territory, near the very border and in parallel to / with it, lies the Augustów Canal* (V. Suvorov, 1968–81).

C. *Franz's room had window to the street which was going in parallel to / with the quay* (V. Nabokov, 1927–8).

D. *They* [*factory and plant committees*] *existed in parallel to / with the trade unions and joined them in 1918* (Great Soviet Encyclopedia, 1969–78).

E. *Say, in King Lear by Shakespeare, the plot line of Lear is developing in parallel to / with the line of Gloucester* (T. Shabalina, early 2000s).

3. Lately, the Russian word *parallel'no* has been used in another construction, governing nouns in the genitive case:

F. *The religion should stay **in parallel from** the civil society.*

G. *This side of life was somewhere **in parallel from** myself and did not touch me.*

H. *Our state still lives **in parallel from** its citizens.*

Explain the reasons why this construction is emerging.

Solution of Problem # 3

Construction A (*in parallel to*) is used when referring to the spatial situation (e.g. of two lines being parallel to each other). Construction B (*in parallel with*) is used when referring to the temporal coincidence (e.g. of two simultaneous events). This can be explained by the inheritance of the government pattern from its cohyponyms or synonyms (*perpendikuljarno chemu-libo* 'perpendicular, vertical to smth' vs. *odnovremenno s chem-libo* 'simultaneously with smth'). One example does not comply with this rule, and it is the oldest one which dates back to the mid-19th century. Apparently, the rule is rather new (indeed, searching the corpus shows more counterexamples in the old texts; this can be shown when discussing the problem with the students and talking about corpus annotation). Construction C (*in parallel from*) is much newer, it can only be found in 21st century texts; this type of government is inherited from *nezavisimo* (*avtonomno*, *svobodno*) *ot chego-libo* 'independently from smth'.

Comment

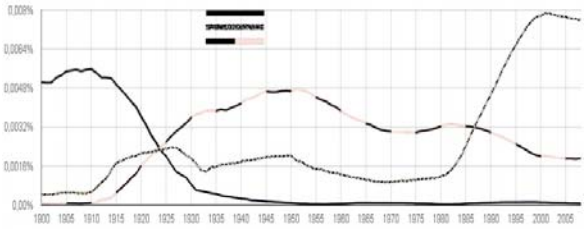
This problem illustrates diachronic variation in Russian. It would be impossible to compose such a problem without using a corpus with rich metadata. Most of the examples come from the Russian National Corpus (http://www.ruscorpora.ru), but it turned out there are not enough examples of these constructions in the RNC. For this reason, some more examples had to be taken from the Internet. This illustrates the concept of Web as corpus (see also problem #6).

This problem was assigned at the Russian Linguistics Olympiad in 2007, and it did not include the dates of the texts. The high school students had to understand themselves that the excerpt from Ivan Goncharov's text is the oldest

one. It was possible because Goncharov's novels are part of the school curriculum in Russia, and students can be expected to know when he lived. This shows that corpus-based problems sometimes require extralinguistic knowledge to account for the variation. Of course, the author of the problem has to be sure that all the solvers are expected to have such knowledge. This makes such problems similar to real-life linguistic research where one can never be confident whether all the information required for explaining the phenomenon is at hand.

Problem #4 (composed by Aleksandrs Berdicevskis)

Consider a Google Books Ngrams frequency diagram for three spellings of the same Russian word throughout the 20th century («Богъ», «Бог», «бог» 'God'). Which line corresponds to which spelling?



Solution of Problem # 4

The sharp frequency changes are caused by historical events which influenced Russian orthography: (1) the October revolution (and the orthography reform of 1918) and (2) the fall of the Soviet Union. Shortly before the revolution, the black line starts subsiding, while the gray line and the dotted line go up. After the revolution, the black line disappears, and the dotted line grows higher than the gray one and stays there until the fall of the Soviet Union. Therefore, the black line is the spelling *Богъ*, ending with the Ъ letter abolished after the Revolution. The dotted line is the spelling *бог*: during the Soviet rule, this word was never capitalized. The gray line stands for *Бог*, which is the present-day norm for the monotheistic deity. The dotted line, however, does not disappear, since the word for 'god' does not always reference such a deity.

Comment

This problem also presents a case of variation, namely diachronically motivated orthographic variation. This problem requires quite a lot of extralinguistic knowledge, but Russian high school students are expected to know about the history of the letter Ъ and the antireligious policy in the Soviet Union. The most important part of the problem is to connect this knowledge with the data represented on the graph.

3.3 Corpus examples illustrating synchronic variation and requiring statistical thinking

Problem #5 (composed by Alexander Piperski)

Russian hypocoristics (diminutive forms of personal names) are most often formed using two classes of suffixes: *-očk-/-ečk-* and *-on'k-/-en'k-*. Below are some names and the hypocoristics with these suffixes derived from them. Each hypocoristic is supplied with the number of texts in the Russian National Corpus (http://www.ruscorpora.ru) in which it occurs:

Base form	***-očk-/-ečk-***	***-on'k-/-en'k-***
Alla	*Alločka* – 48	*Allon'ka* – 1
An'a	*Anečka* – 111	*Anen'ka* – 2
Val'a	*Valečka* – 58	*Valen'ka* – 8
Vas'a	*Vasečka* – 9	*Vasen'ka* – 119
Volod'a	*Volodečka* – 15	*Voloden'ka* – 38
Glaša	*Glašečka* – 0	*Glašen'ka* – 11
Dima	*Dimočka* – 22	*Dimon'ka* – 0
Klava	*Klavočka* – 20	*Klavon'ka* – 0
Kol'a	*Kolečka* – 19	*Kolen'ka* – 73
Nad'a	*Nadečka* – 3	*Naden'ka* – 102
Pet'a	*Petečka* – 11	*Peten'ka* – 70
Saša	*Sašečka* – 4	*Sašen'ka* – 155
Sveta	*Svetočka* – 39	*Sveton'ka* – 0
Sen'a	*Senečka* – 15	*Senen'ka* – 0
Serëža	*Serëžečka* – 6	*Serëžen'ka* – 76
Tan'a	*Tanečka* – 120	*Tanen'ka* – 0
Tol'a	*Tolečka* – 17	*Tolen'ka* – 7
Jul'a	*Julečka* – 22	*Julen'ka* – 27

1. Here are six more pairs of hypocoristics:
Vitečka ~ Viten'ka
Olečka ~ Olen'ka
Lidočka ~ Lidon'ka
Sonečka ~ Sonen'ka
L'ubočka ~ L'ubon'ka
Jašečka ~ Jašen'ka

Try to predict for each of these pairs which hypocoristic occurs in more texts in the Russian National Corpus. If you cannot do that for some names, explain why.

Solution of problem #5

The choice of the suffix depends on the last consonant of the stem. *-očk-*/*-ečk-* is more frequent after non-palatalized ("hard") consonants and after *n'* (dissimilation). *-on'k-*/*-en'k-* is more frequent after hushing sibilants (*š*, *ž*; dissimilation) and after palatalized ("soft") consonants other than *n'*. For *l'* no rule can be stated.

Therefore, the expected more frequent forms are *Viten'ka*, *Lidočka*, *Lúbočka*, *Sonečka* and *Jašen'ka*. For *Olečka* ~ *Olen'ka* no prediction can be made.

Comment

This problem requires the ability to neglect introspection, since all Russian-speaking solvers have some intuitive judgements on the topic. It also shows that a linguist sometimes has to work with tendencies, rather than strict rules and leave some variation unexplained.

3.4 Other corpus-like data

Problem #6 (Composed by Vitaly Pavlenko)

In Turkish, the word *kadın* 'woman' is used when naming women by their profession, occupation, etc. This word can be placed before the noun referring to a profession as well as after it; there are no absolutely precise rules explaining it, but there is a certain tendency, according to which one of the two possible variants is used more often than the other.

Given are some Turkish phrases with the word *kadın* and their English translations. For the first 9 phrases, it is stated how many times they occur on the Internet as *kadın X* and how many times as *X kadın*. For the last 6 phrases the corresponding numbers are given to choose from:

Turkish phrase	**translation**
kadın barmen (40) *barmen kadın (191)*	barwoman
kadın dikişçi (2) *dikişçi kadın (112)*	seamstress
kadın hakem (1,910) *hakem kadın (107)*	female judge
kadın kasiyer (82) *kasiyer kadın (112)*	female cashier
kadın mühendis (3,350) *mühendis kadın (428)*	female engineer
kadın öğretmen (36,200) *öğretmen kadın (6,500)*	female teacher
kadın polis (41,200) *polis kadın (13,700)*	policewoman
kadın gazeteci (23,200) *gazeteci kadın (1,720)*	female journalist
kadın satıcı (400) *satıcı kadın (2,190)*	saleswoman
kadın avukat (...) *avukat kadın (...) 780 / 9,520*	female lawyer
kadın çevirmen (...) *çevirmen kadın (...) 29 / 1,630*	female translator
kadın fırıncı (...) *fırıncı kadın (...)10 / 275*	female baker
kadın ressam (...) *ressam kadın (...) 407 / 15,000*	female artist
kadın sütçü (...) *sütçü kadın (...) 97 / 758*	milkwoman
kadın terzi (...) *terzi kadın (...) 639 / 1,450*	tailoress

1. Put the right numbers into the brackets. Explain your reasoning.

Note. *ğ, ç, ş, ı, ö, ü* are special sounds of Turkish. The numbers given in the problem are Google hit counts as of October 23, 2008.

Solution of Problem #6

When used with the names of skilled occupations *kadın* is put before the noun, and with the names of service sector occupations *kadın* is put after the noun. The answer is as follows:

kadın avukat (9,520) *avukat kadın (780)*	female lawyer
kadın çevirmen (1,630) *çevirmen kadın (29)*	female translator
kadın fırıncı (10) *fırıncı kadın (275)*	female baker
kadın ressam (15,000) *ressam kadın (407)*	female artist
kadın sütçü (97) *sütçü kadın (758)*	milkwoman
kadın terzi (639) *terzi kadın (1,450)*	tailoress

Comment

As well as #5, this problem shows that linguists sometimes have to deal not with precise rules (as it is usually the case in traditional linguistics problems), but only with tendencies. It also demonstrates that search engines can be used in linguistic studies as large corpora. In this problem the exact difference between the two numbers does not matter (e.g., 82 vs. 112 is the same as 2 vs. 112 for the purposes of the problem), but similar problems might be created where the distance between the two numbers is essential.

4 Corpus linguistics problems: Some pitfalls

We have shown that corpus-based problems have some advantages over traditional types of problems. However, some of these can be regarded as weak sides, too.

Corpus linguistics problems illustrate linguistic phenomena with real data. Unfortunately, many sentences present in the corpus are rather large and sometimes even clumsy. If long sentences are used to illustrate the usage of just one word, there will inevitably be a lot of irrelevant information (cf. #1, #2, #3; each of the long sentences is intended to illustrate the behavior of a single word). On the other hand, the solvers might enjoy reading long real-life sentences instead of artificial examples.

Another issue that makes composing and using corpus-based problems difficult is the philosophy underlying such problems. In a typical problem that contains artificial data all phenomena must be explained by the solver. There is no place for unexplained variation. However, problems on corpus linguistics require statistical manner of thinking, rather than strictly deterministic conclusions. For example, problem #5 illustrates tendencies, some of which are more solid than the others. However, less frequent names do exist, which might baffle a solver who is used to explaining everything within a linguistics problem. In problem #4, the word *Богъ* which uses the old spelling did not vanish completely after the spelling reform of 1918. There is some noise at the bottom of the graph, and the solver has to understand that it is not necessary to account for this noise in order to solve the problem.

The authors of problems on corpus linguistics should also be aware that the methodology they demonstrate is not always sound. For instance, in problem #6 Google hit counts are used in spite of the fact that is has been shown many times that they cannot be trusted (cf. Kilgarriff 2007). Ideally, a problem on computational linguistics should be accompanied by an afterword explaining the drawbacks of the methods it uses. Otherwise it might be tempting for students to get a simplistic notion of corpus linguistics and its methodology.

5 Conclusion

The corpus is a valuable data source not only for linguistic studies, but also for composing linguistic problems. Corpus linguistics problems are useful to introduce the study of variation and the basics of statistical thinking in linguistics. However, they also have certain drawbacks, namely their length and unexplained variation within the data (which can however sometimes be an advantage bringing the problem closer to the real life). The main advantage of corpus-based problems is that real data are used, which can be verified and even more thoroughly studied by the student. Moreover, through such problems the students become acquainted with corpus linguistics as a research field that is rapidly gaining importance.

Acknowledgements

This research has been partially financed by a research program of History and Philology Branch of the Russian Academy of Sciences, a grant from the Russian Humanitarian Scientific Foundation No. 13-04-00307a, a President Grant for Leading Scientific Schools of Russia (No. NSh-6577.2012.6) and the Program of Strategic Development of the Russian State University for the Humanities. We would like to thank Aleksandrs Berdicevskis, Leonid Iomdin and Maria Konoshenko for valuable comments.

References

Adam Kilgarriff. 2007. Googleology is Bad Science. *Computational Linguistics* 33 (1): 147–51.

Boris L. Iomdin and Leonid L. Iomdin. 2011. Valency ambiguity interpretation: what can and what cannot be done. In: Proceedings of the 5th International Conference on Meaning-Text Theory. Barcelona, September 8–9, 2011. Ed. by Igor Boguslavsky and Leo Wanner. Barcelona: University Pompeu-Fabra.

Ivan A. Derzhanski, Aleksandr S. Berdichevskij, Ksenia A. Guiliarova, Boris L. Iomdin, Elena V. Muravenko, and Maria L. Rubinstein. 2004. O perevodimosti lingvisticheskix zadach: Uroki pervoj mezhdunarodnoj lingvisticheskoj olimpiady [On translatability of linguistic problems: Lessons of the First International Linguistics Olympiad]. In: *Computational linguistics and intellectual technologies*. Papers from the annual international conference "Dialogue". Moscow: Nauka, pp. 240–5.