БЫТЬ ИЛИ НЕ БЫТЬ: КОРПУСА КАК ИНДИКАТОРЫ (НЕ)СУЩЕСТВОВАНИЯ

Пиперски А. Ч. (apiperski@gmail.com)

РГГУ / РАНХиГС, Москва, Россия

В статье обсуждаются понятия приемлемости, встречаемости, грамматичности и существования, в первую очередь — связь между корпусной лингвистикой и вопросом о существовании единиц лексикона. Доказывается, что корпуса не могут свидетельствовать о несуществовании слова, поскольку они обычно являются выборками из некоторой генеральной совокупности, а верхняя граница доверительного интервала для частотности на основе выборки всегда больше 0, вне зависимости от частотности, подсчитанной по выборке. Практическое правило таково: если что-то не встретилось в корпусе, оно могло бы встретиться в корпусе того же размера и состава от 0 до 5 раз. Если же единица присутствует в корпусе, это может служить доказательством её существования в языке, но окончательное решение зависит от того, признаем ли мы корпус репрезентирующим ту разновидность языка, которая нас интересует. Таким образом, корпусное исследование не позволяет доказать несуществование, но позволяет доказать существование; однако второй вид доказательства связан с установлением репрезентативности, которое порой влечёт за собой субъективность и оценочность в суждениях.

Ключевые слова: приемлемость, встречаемость, грамматичность, существование, корпусная лингвистика, выборка, генеральная совокупность, доверительный интервал

TO BE OR NOT TO BE: CORPORA AS INDICATORS OF (NON-)EXISTENCE

Piperski A. Ch. (apiperski@gmail.com)

Russian State University for the Humanities / Russian Academy of National Economy and Public Administration, Moscow, Russia

This paper discusses the notions of acceptability, occurrence, grammaticality and existence, and focuses on the relationship between corpus linguistics and the question of the existence of lexical items. Since corpora are almost exclusively samples from larger populations, it is claimed that they cannot provide evidence for non-existence of words, collocations or constructions.

This is because the upper limit of a confidence interval for frequency based on a sample is always greater than zero regardless of the sample frequency. The rule of thumb goes as follows: anything that does not occur in a corpus might have occurred in a similar same-sized corpus zero to five times. If an item occurs in a corpus, this fact can serve as a proof of its existence in the language, but the final decision depends on whether the relevant contexts from the corpus are judged representative of the language variety of interest. In conclusion, I claim that a corpus-based study cannot prove the non-existence of a linguistic item, although it can be used to prove its existence. However, the latter type of proof includes assessing the representativeness of a corpus, which might lead to subjectivity and value judgments.

Keywords: acceptability, occurrence, grammaticality, existence, corpus linguistics, sample, population, confidence interval

1. Introduction: the notions of acceptability, occurrence, grammaticality and existence

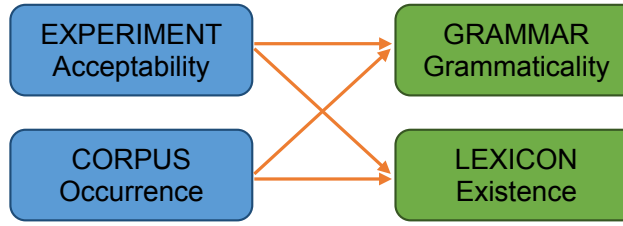
Corpus linguistics has provided linguists with various means of studying frequency-related phenomena. The most radical conceptions of corpus linguistics even state that a corpus is merely a source of information on frequencies (Gries 2009: 11). However, this frequency-based approach is at odds with traditional linguistics which relies heavily on binary distinctions of type: "acceptable vs. unacceptable", "grammatical vs. ungrammatical" and "existent vs. non-existent". Gradient grammaticality has been discussed quite often (cf. Keller 1998, Fanselow et al. (eds.) 2006, Lau et al. 2014; Fedorova 2013 provides a critical survey on this topic and its relation to psycholinguistics), but it is still unwelcome in general linguistics. Scholars are reluctant to accept the idea that grammaticality is gradient rather than categorical, regarding gradience as a matter of performance rather than competence. This raises a question as to whether or not the statistical approach of corpus linguistics generating numerical data and the categorical approach of traditional linguistics can somehow be reconciled. The aim of this paper is to discuss whether statistical data obtained from corpora are transformable into the binary opposition "existence vs. non-existence", which is closely related to grammaticality.

However, in order to do that, we have to make a clear distinction between acceptability, grammaticality and existence, since these terms are sometimes used interchangeably, which might cause confusion. Grammaticality and existence are language-internal, whereas acceptability refers to the speakers' intuitions. As stated by Newmeyer (2007: 398), "no rational linguist would test informants about judgments of grammaticality, since grammaticality is a theoretical construct". In other words, grammaticality is conformity to the rules of the grammar, which cannot be judged without knowledge of these rules. As for existence, it has to do with lexicon rather than grammar. An item is existent if it is listed in the lexicon, and non-existent otherwise. Since lexicon is also a theoretical construct, it is hard to say what kind of items it comprises (cf. Jackendoff 2002 among others), but in general one can say

that grammaticality refers to larger units, e.g., sentences, whereas existence refers to smaller units, e.g., morphemes and words. The position of collocations and constructions on this scale remains debatable.

Corpus frequencies providing a linguist with information about occurrence or non-occurrence are similar to acceptability judgments in that they are both real-life data rather than theoretical constructs. The main source of acceptability data is experiment. I use this term in a broad sense, covering various surveys and tasks as well as introspection, which can be understood as an experiment conducted on a single participant.

Thus, we have two theoretical notions (grammaticality and existence) and two real-life concepts (acceptability and occurrence), and the task of a linguist is to infer information about the former from the latter. This can be summarized in the following scheme:



Scheme 1. The interrelations between acceptability, occurrence, grammaticality and existence

In this paper, I am going to explore only one of the four arrows in Scheme 1, namely the one connecting occurrence and existence. My aim is to answer the following question: Can corpus data tell us whether a word, collocation or construction exists in a given language variety?

2. Absence from corpora as evidence for non-existence?

The absence of an item from a corpus is often taken as evidence of its non-existence. However, this kind of evidence can only be conclusive if a corpus contains the whole population of certain texts rather than a sample. For instance, an absence of a certain word from the Shakespearean canon is sufficient to demonstrate that this word does not occur in the plays of this particular author, but an absence of a word from any corpus of English is not enough to prove that this is not a word of the English language.

Corpus size has a large impact on whether a search returns zero or more results. To illustrate the importance of corpus size, one can compare search results for the same words in two corpora of different sizes. Let us take the main subcorpus of the Russian National Corpus (RNC, www.ruscorpora.ru; 230m words) and the ruTenTen corpus (the.sketchengine.co.uk; 14.5b words), the latter being more than 60 times larger than the former. Table 1 presents a selection of words absent from RNC and their frequencies in ruTenTen:

Table 1: Frequencies of some words absent from RNC in ruTenTen

Word	Absolute frequency	Frequency (ipm)
selfi 'selfie'	4	0.0003
klubneobrazovanie 'tuber formation'	506	0.03
Kuautemok 'Cuauhtémoc (Mexican proper name)'	511	0.03
èkonomist-meždunarodnik 'international economist'	647	0.04
prokrastinacija 'procrastination'	927	0.06
mikruha 'microchip (colloq.)'	2,506	0.17

Clearly, RNC and ruTenTen represent different varieties of Russian, the former including a more standardized language and the language of the 18th, 19th and 20th centuries, but the difference between 0 on the one hand and 506 and 647 on the other hand for such neutral words as *klubneobrazovanie* and *èkonomist-meždunarodnik* is striking. It demonstrates that a word that is absent from a smaller corpus can be quite frequent in a larger corpus. For this reason, the absence of an item from any corpus containing a sample rather than a whole population of some kind cannot be taken as a proof of this item's non-existence in the variety of language represented by this corpus.

3. Confidence intervals instead of binary judgments

Since corpus linguistics is mostly about studying samples, and it is rarely the case that a corpus linguist has to deal with the whole population, standard statistical techniques for estimating population parameters using samples can be applied in this domain. If we are trying to estimate the frequency of a word, we need to construct a confidence interval for the population proportion based on a sample proportion (Baroni & Evert 2009).

Formulae for computing confidence intervals for proportions are given in all basic statistics textbooks (Diez et al. 2012, Field et al. 2012, Rumsey 2010: 77–78, to name just a few recent ones), manuals in statistics for linguists not being an exception (cf. Butler 1985: 62–63, Gries 2013: 129–135). Introductory textbooks usually mention the normal approximation confidence interval:

$$p \pm z\sqrt{\frac{p(1-p)}{n}}$$

where p is the sample proportion, n is the sample size and z is the value of the standard normal distribution corresponding to the desired confidence level (in most cases $z = 1.96$ at the customary confidence level of 95%).

However, this method for computing confidence intervals is inapplicable in the case where p is extremely close or equal to 0 or 1. If $p = 0$, the confidence interval shrinks to [0, 0], which is unsatisfactory: even if we never encounter a phenomenon in our sample, we can never be sure it does not exist at all (cf. Partington 2014). This recalls Laplace's sunrise problem: what is the chance that the sun will rise tomorrow?

Even though the event "the sun does not rise in the morning" has never been observed before, one cannot be sure that its probability is equal to 0. This means that other methods of computing confidence intervals are required.

Fortunately, normal approximation is far from being the only way of estimating confidence intervals for proportions. An extensive list of relevant methods is given in a paper by Newcombe (1998). The most appropriate method for our purposes is Wilson's score method with continuity correction (Newcombe 1998: 859), which is also the default method used by the `prop.test` function in R (R Core Team 2013). The lower limit L and the upper limit U of the confidence interval can be respectively calculated using the following formulae:

$$L = \frac{2np + z^2 - 1 - z\sqrt{z^2 - 2 - \frac{1}{n} + 4p(n(1-p)+1)}}{2(n+z^2)}$$

$$U = \frac{2np + z^2 + 1 + z\sqrt{z^2 + 2 - \frac{1}{n} + 4p(n(1-p)-1)}}{2(n+z^2)}$$

If $p = 0$, the lower limit of the confidence interval L must be taken as 0. If $p = 0$, $z = 1.96$, and we assume that n is much higher than z since sample sizes in corpus linguistics are huge compared to sample sizes in experimental sciences, the expression for U can be simplified:

$$U = \frac{2n \times 0 + z^2 + 1 + z\sqrt{z^2 + 2 - \frac{1}{n} + 4 \times 0 \times (n(1-p)-1)}}{2(n+z^2)} \approx$$

$$\approx \frac{z^2 + 1 + z\sqrt{z^2+2}}{2n} = \frac{4.79}{n}$$

This means that the estimated confidence interval for the proportion given $p = 0$ is $\left[0, \frac{4.79}{n}\right]$. This can be restated as a rule of thumb: if a phenomenon does not occur in a corpus, we can be 95% sure that it will occur 0 to 5 times in a same-sized corpus drawn from the same population.

For this reason, it is not surprising to find a word absent from RNC around 300 times in ruTenTen, which is 60 times larger. This does not even entail the conclusion that RNC and ruTenTen are samples from different populations with respect to the frequency of this particular word. However, the confidence interval approach makes the notion of non-existence virtually non-existent: anything that does not occur in a corpus might have occurred in a similar same-sized corpus 0 to 5 times.

4. Presence in corpora as evidence for existence?

Whether a certain word, collocation or construction is present in a language is often a topic of debate. It is important to answer such questions when writing a text

in the standard variety of a language, composing a dictionary, or creating a grammar. If the judgments of native speakers differ, the presence of an item in a corpus is often quoted as evidence for its existence.

In the Russian-speaking community, it has become increasingly popular to use RNC for investigating the existence of words, collocations or constructions. Since it is the most user-friendly corpus of Russian, accessible even to non-linguists, it is not uncommon to refer to RNC as simply "the corpus", which makes it the sole and ultimate authority for answering questions as to whether something is possible in Russian or not. Discussions of this kind are quite common in social media among educated native speakers, where two positions are clearly identifiable: one group of people might reject some word, collocation or construction because they judge it unacceptable, whereas the other group points to examples drawn from RNC as evidence for the existence of this item.

A typical discussion of this kind took place in 2006 in the blog of the LiveJournal user ormer_fidler(http://ormer-fidler.livejournal.com/22044.html?thread=298268#t298268). A commenter criticized the use of the word *razbudit'sä* instead of *prosnut'sä* 'to wake up' and asked for ormer_fidler's opinion. The latter cited the only example of this word from the RNC as a proof of its existence in Russian, while admitting that this item is very infrequent. Since then, the RNC has experienced a significant increase, and the search now retrieves 5 occurrences of this word. This raises the following question: how many occurrences in a corpus are enough to declare a word existing? Probably the most persuasive answer would be the following: a corpus proves the existence of a word, collocation or construction if it occurs at least once and the retrieved context(s) is (are) judged as relevant to the variety of language in question. The second premise introduces subjectivity into the process of determining what exists in a language and what does not, since contexts from a corpus can be rejected on any grounds, in particular based on value judgments.

A notable feature of RNC is that it incites such value judgments. Since the main subcorpus of RNC contains a high proportion of literary texts (101.8m words / 230m words = 44%), and search results are displayed together with the author's name and the title of the text, users of RNC tend to give more weight to the authors and texts they know and hold in esteem. If a word, collocation or construction was used by a distinguished writer (excluding the writers whose language is unanimously recognized as bizarre, such as Andrei Platonov), users of RNC tend to find it acceptable even if it is very infrequent. However, singular instances can be discarded as "errors" regardless of the status of their author. In other words, if we assume that our corpus is a sample, some parts of it can be claimed to have found their way into the corpus by mistake and not to belong to the population of interest, e.g., "correct standard language". In a recent magazine article, Naberezhnov (2013) provides an instructive example of this kind: when faced with the question of whether *iskrenno sprosit'* 'to ask sincerely' is an acceptable collocation, the organizer of the *Total'nyj diktant* (Total Dictation) project resorts to RNC and finds a single occurrence of this word combination in Boris Pasternak's *Doctor Zhivago*. However, she rejects it as "an unfortunate wording, even though produced by a great writer". Unfortunately, this result is irreproducible. The form *iskrenno* 'sincerely' does not occur all in *Doctor Zhivago*; the

variant form *iskrenne* occurs five times, but it is never used with the word *sprosit'* 'to ask'[1]. Even though unreliable, this example highlights a typical way of using a corpus to prove existence of a collocation.

5. Conclusions

Corpus linguistics is hard to reconcile with the traditional binary distinction "existent vs. non-existent". When doing corpus-based research, linguists need to be more aware of the fact that they are working with samples, which means that they have to apply standard statistical techniques for estimating population parameters from a sample rather than tacitly transfer sample parameters to the whole population. If one bears in mind the nature of a corpus, two conclusions emerge:

a) the absence of a word, collocation or construction from a corpus cannot prove its non-existence, since the upper limit of the confidence interval for its frequency is always above zero.
b) even a single example in a corpus is enough to prove that a word, collocation or construction exists in a language, under the premise that the relevant example(s) can be judged representative of the language variety in question; however, this premise inevitably leads to a certain degree of subjectivity.

References

1. *Baroni, Marco & Stefan Evert.* 2009. Statistical methods for corpus exploitation. In Anke Lüdeling and Merja Kytö (eds.), Corpus linguistics: An international handbook. Vol. 2 (Handbooks of Linguistics and Communication Science 29.2). 777–803.
2. *Butler, Christopher.* 1985. Statistics in linguistics. Oxford: Blackwell.
3. *Diez, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel.* 2012. OpenIntro statistics. [Lexington, KY: CreateSpace].
4. *Fanselow, Gisbert et al.* (eds). 2006. Gradience in grammar: Generative perspectives. Oxford; New York: Oxford University Press.
5. *Fedorova, Olga V.* 2013. Ob èxperimental'nom sintaksise i o sintaksicheskom experimente v jazykoznanii (On experimental syntax and syntactic judgment experiments in linguistics). Voprosy Jazykoznanija 1, 3–21.
6. *Field, Andy P., Jeremy Miles, and Zoë Field.* 2012. Discovering statistics using R. London: Sage.
7. *Gries, Stefan Thomas.* 2009. Quantitative corpus linguistics with R: a practical introduction. New York: Routledge.
8. *Gries, Stefan Thomas.* 2013. Statistics for linguistics with R: A practical introduction: Textbook. Berlin: De Gruyter Mouton.

[1] I am grateful to Vladimir Belikov for pointing this out.

9. *Jackendoff, Ray.* 2002. What's in the lexicon? In S. G. Nooteboom, Fred Weerman and Frank Wijnen. Storage and computation in the language faculty. Dordrecht: Kluwer Academic. 23–58.
10. *Keller, Frank.* 1998. Gradient Grammaticality as an Effect of Selective Constraint Re-ranking In: M. Catherine Gruber, Derrick Higgins, Kenneth S. Olson, and Tamra Wysocki (eds.). Papers from the 34th Meeting of the Chicago Linguistic Society. Vol. 2: The Panels. 95–109.
11. *Lau, Jey Han, Alexander Clark, and Shalom Lappin.* 2014. Measuring gradience in speakers' grammaticality judgements. In: Proceedings of the 36th Annual Meeting of the Cognitive Science Society Québec City, Canada, 23–26 July 2014. 821–826.
12. *Naberezhnov, Grigory.* 2013. Avtorskaja diktatura (Author's dictatorship). Russkij Reporter, 01.04.2013. URL: http://rusrep.ru/article/2013/04/01/totaldiktant
13. *Newcombe, Robert G.* 1998. Two-sided confidence intervals for the single proportion: Comparison of seven methods. Statistics in Medicine 17, 857–872.
14. *Newmeyer, Frederick J.* 2007. Commentary on Sam Featherston, 'Data in generative grammar: The stick and the carrot'. Theoretical Linguistics 33:3, 395–399
15. *Partington, Alan.* 2014. Mind the gaps. International Journal of Corpus Linguistics 19:1, 118–146.
16. *R Core Team.* 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org
17. *Rumsey, Deborah J.* 2010. Statistics essentials for dummies. Indianapolis: Wiley Pub., Inc.