Laura A. Janda
USA – Norway, The University of Tromsø – The Arctic University of Norway

Linguistic profiles: A quantitative approach to theoretical questions

Key words: cognitive linguistics, linguistic profiles, statistical analysis, theoretical linguistics
Ключевые слова: когнитивная лингвистика, лингвистические профили, статистический анализ, теоретическая лингвистика

Abstract

A major challenge in linguistics today is to take advantage of the data and sophisticated analytical tools available to us in the service of questions that are both theoretically interesting and practically useful. I offer linguistic profiles as one way to join theoretical insight to empirical research. Linguistic profiles are a more narrowly targeted type of behavioral profiling, focusing on a specific factor and how it is distributed across language forms. I present case studies using Russian data and illustrating three types of linguistic profiling analyses: grammatical profiles, semantic profiles, and constructional profiles. In connection with each case study I identify theoretical issues and show how they can be operationalized by means of profiles. The findings represent real gains in our understanding of Russian grammar that can also be utilized in both pedagogical and computational applications.

1. The quantitative turn in linguistics and the theoretical challenge

I recently conducted a study of articles published since the inauguration of the journal *Cognitive Linguistics* in 1990. At the year 2008, *Cognitive Linguistics* took a quantitative turn; since that point over 50% of the scholarly articles that journal publishes have involved quantitative analysis of linguistic data [Janda 2013: 4–6]. Though my sample was limited to only one journal, this trend is endemic in linguistics, and it is motivated by a confluence of historic factors. Within the last two decades, we have witnessed a) the flourishing of digital corpora and crowdsourcing sites, providing linguists with enormous quantities of data, and b) great advances in the development of open-source statistical software (primarily R), making the analysis of data feasible. Today any linguist with a computer and an internet connection has access both to billions of datapoints as well as the tools to analyze patterns in language data. It is no surprise that linguists are taking advantage of this vast opportunity.

In a sense we linguists are undertaking as a community a big experiment, collectively figuring out what kinds of statistical models can be applied to what kinds of data, how results can be interpreted. In addition to this methodological challenge, we are also facing a theoretical challenge. Counting up datapoints and running them through statistical tests does not guarantee that we will produce meaningful results. It is wonderful to have plenty of data and sophisticated tools, but we also need to address theoretically important research questions.

Combining statistical methods with theoretical insights is a real challenge because it requires creativity on our part. Theories don't specify quantitative methods and quantitative analysis does not guarantee theoretical relevance. Without a theoretical edge, our statistical models will only yield trivial results.

I see this problem in terms of three "levels", where theory occupies a middle ground between what I call "Big Questions" and "operationalization". Big Questions are issues that transcend any given theory, are interesting for all linguists, and have implications beyond linguistics as well. Let's take just two examples of Big Questions:

1) What is the relationship between form and meaning?

2) What is the relationship between lexicon and grammar?

Whereas these Big Questions unite us as linguists, the theoretical approaches we take to these issues vary. We need theory in order to bring Big Questions into focus enough to formulate specific research questions. But the focusing will differ according to the theory, and often it is not possible to view a given Big Question from more than one theoretical perspective simultaneously. For example, while some linguists neatly separate form from meaning, others insist that there is no form without meaning, and these two perspectives lead to different expectations and entailments. A distinction between lexicon and grammar is assumed in theories that assign various phenomena to one or the other, however other theories view lexicon and grammar as parts of a single continuum lacking a clear boundary. Obviously, the very definition of what constitute lexical vs. grammatical categories will differ across such theories. Although some theories presume that linguistic categories are discretely bounded, others come with the expectation that categories may be fuzzy and overlapping, structured around prototypes. Grammatical constructions can be modeled in many ways, for example, as hierarchical structures often diagrammed as trees, or alternatively using a relatively "flat" sequential structure.

Theory can thus be thought of as a kind of lens through which we view the Big Questions. We need a lens in order to gain focus, but there are many lenses, and each lens gives us a clearer view of some part of a Big Question, while at the same time it suppresses other potential views.

Once we have a theoretical focus on a Big Question, we need to operationalize it. We can't directly ask a corpus or participants in an experiment a Big Question, even when we have a clear theoretical perspective. Corpus and experimental data don't specify which language units belong to lexicon vs. grammar or how they are structured. At this point our task is to operationalize our Big Question. We have to ask questions that the corpus (or other data set) can answer. Here we take our theoretical expectations into account, which might tell us that, for example, we expect a difference in form to correlate with a difference in meaning. If it has been asserted that all members of a group of morphemes have the "same" meaning, we can look at their distribution in a corpus and see whether they do indeed behave the same. For example, we could look at the semantic tags of the roots that these morphemes combine with and ask whether the morphemes show any pattern in their distribution, as in section 4.2. A statistically significant pattern would give evidence that the morphemes are distributed according to some semantic cues rather than being in free variation as previously asserted.

Operationalization is not just about methodology and applying techniques. It requires innovative approaches to linguistic data and the patterns therein. One has to find a connection between what we do have (mainly a lot of messy data) and the way it can shed light on our theoretically focused Big Questions. There are certainly many ways to operationalize linguistic research questions. I present one such strategy in this article.

I offer "linguistic profiles" as a suite of methodological ideas bridging the gap between key theoretical issues in linguistics and quantitative models. Linguistic profiling involves probing the frequency distribution of a given factor (or set of factors) in connection with a given linguistic unit (or set of units). Collectively linguistic profiles make it possible to operationalize theoretical questions about the structure of languages. Linguistic profiles can guide the way we collect and analyze data.

Although there are many Big Questions linguists might ask, I will limit myself to the two presented above. In section 2 I present the theoretical perspective that focuses these Big Questions in this article. Linguistic profiling is presented in more detail in section 3, followed by a series of case studies in which different types of linguistic profiles are used to address aspect, prefixation, and synonymy in Russian in section 4. I conclude in section 5.

2. The theoretical perspective of cognitive linguistics

Linguistic profiles can probably be implemented in most, if not all, theoretical frameworks. The case studies I present herein all view Big Questions from the perspective of cognitive linguistics, so I will give a brief overview of the assumptions and expectations that this perspective entails [for more details, see Janda 2015].

Cognitive linguistics makes a **minimal assumption** in that it does not invoke any cognitive mechanism specific to language alone. The assumption is instead that language is not “hard-wired”, but rather a phenomenon that arises due to the general cognitive strategies of the brain. In other words, linguistic cognition is indistinguishable from general cognition and should be accounted for in terms of general cognitive mechanisms and structures that can be independently established. This means that we do not assume the existence of an autonomous language faculty, nor any a priori language universals. This premise does not exclude the existence of language universals, but comes with the expectation that the language universals that do exist are likely to be relatively few, very general, and given by common facts of human experience, and therefore mostly neither very specific nor very interesting. For example, languages of the world tend to have nouns and verbs, corresponding to the fact that human beings experience objects and events, but the behaviors of nouns and verbs can vary greatly across languages. This perspective on language universals is supported by evidence from typological studies (see [Evans, Levinson 2009] and citations therein).

The minimal assumption further entails that there is no strict division between grammar and lexicon. Different languages can distribute meaning in different ways. Evidentiality, for example, is coded by the verbal paradigm in languages like Bulgarian and Macedonian, which choose a relatively grammatical means to mark a distinction that is made in a more lexical way in a language like English with words like *allegedly*. Russian takes a middle ground here, using both more lexical means like *будто* (*бы*) and *якобы* ‘allegedly, as if’ together with the grammatical conditional construction to express evidentiality. Cognitive linguistics views **lexicon and grammar as parts of a single continuum** lacking a clear boundary.

Cognitive linguistics is **usage-based**. This means that generalizations emerge from language data. Cognitive linguistics does not recognize a strict division between “langue” (a speaker’s internal, idealized grammar) and “parole” (a speaker’s language production). Only the latter is observable, and utterances are thus taken as language data that are the object of study. This means that cognitive linguists base their findings on authentic language use, as documented in a corpus or experiment, for example. Cognitive linguists do not invoke underlying forms, but instead focus on observable language phenomena.

Cognitive linguistics presumes that **meaning is central** to all language phenomena. This means that there are no semantically empty forms, and that difference in form necessarily reflects difference in meaning, with the entailment that there are no true synonyms, only near synonyms. Meaning is not the exclusive privilege of the lexicon (or the lexical end of the lexicon-grammar continuum), and grammatical categories such as case, aspect, person, etc. have meaning as well. The central and pervasive role of meaning entails the expectation that differences in behavior of linguistic forms should be motivated by differences in meaning. In other words, if two linguistic forms behave differently (show different associations with various other forms or factors), they should also express different meaning. This does not necessarily mean that we can specifically predict differences in behavior, but it does mean that we expect correlations. Differences in behavior are particularly relevant to linguistic profiles, given that they are a subtype of behavioral profiles, as detailed in section 3.

Since cognitive categories are structured primarily in terms of prototypes and family resemblance, cognitive linguistics also expects most **linguistic categories** to **have a radial category structure**. The behavior of linguistic form and meaning follows from the radial category structure, and, as a result, many linguistic phenomena are gradient rather than absolute, with the strength of an effect scaled according to the prototypicality of the members of a category. For example, a prototypical meaning of the Russian verbal prefix *про-* refers to movement through space, which is characteristic of loud sounds. There are over fifty verbs denoting the production of sounds that use *про-* to form their perfective partners, such as *прогреметь* 'thunder' and *прогрохотать* 'rumble'. Change of state is however, less compatible with the prototypical meaning of *про-*, and there are only four verbs like *прогоркнуть* 'become bitter' and *пропитаться* 'become saturated' (see more about these and similar examples in section 4.2).

For cognitive linguistics the relevant unit of study in language is the construction. Rather than being modeled as hierarchical structures, **constructions are form-meaning pairings**, in keeping with growing evidence that grammatical structure is flat, relying on locally-available sequential cues [Frank et al. 2012].

Let us now return the Big Questions posed in the previous section and see how they can be focused by the theoretical perspective of cognitive linguistics.

1) *What is the relationship between form and meaning?* Given that cognitive linguistics presumes a constant relationship between form and meaning, this question can be restated as: *How does form reflect meaning?* More specifically, we can also ask: *Can we use difference in form as a measure of meaning?*
2) *What is the relationship between lexicon and grammar?* Given that cognitive linguistics does not assume a strict division, our aims include finding evidence for the ways in which lexicon and grammar interact along their continuum. Therefore we ask: *How do we account for meaning in grammar?* Since lexical meaning tends to be more concrete and specific, it is also easier to model. This prompts the question: *Can we use similar models for grammatical meanings?*

Armed with our theoretically focused questions, we can now move on to the next step, namely operationalizing our questions.

3. Operationalization, portability, and multipurposing via linguistic profiles

Linguistic profiles are a way to restate theoretical questions so that we can find patterns in language data that give evidence that support or refute our expectations. Linguistic profiles can be thought of as a suite of methodological ideas that make it possible to approach linguistic research questions from a variety of angles.

Linguistic profiles are focused subtypes of behavioral profiles. Behavioral profiles for words or other linguistic units can include a wide variety of parameters, such as lexical collocations, syntax, morphology, semantics, etc. While behavioral profiles have a long tradition [cf. Firth 1957; Harris 1970; Hanks 1996; Geeraerts et al. 1999; Speelman et al. 2003; Divjak, Gries 2006; Gries, Divjak 2009], they can also have drawbacks. When looking at a wide variety of factors, it is very hard to know how they might be weighted (since some factors are likely to be stronger than others), or whether the factors might overlap with each other (leading to collinearity problems for statistical analysis). For example, in a study of Russian verbs, one might want to include both the aspect of the verbs and the distribution of past vs. non-past finite forms (and possibly several other factors). However, it turns out that aspect and tense are closely associated with each other (see section 4.1), and this means that an analysis that includes both factors might misrepresent their effects. This does not mean that behavioral profiles are not useful – there are certainly situations in which they are the optimal approach, particularly when one is trying to find an overall pattern rather than focusing on a more specific question.

Linguistic profiles restrict the use of factors to give a tighter focus and avoid collinearity problems. Many types of linguistic profiling are possible, though all linguistic profiling methods take the form-meaning relationship as their point of departure. Linguistic profiling methods share the characteristic of selecting observable frequency distributions of forms and measuring their relationship to a given linguistic phenomenon. Here I list only three types, all of which are illustrated in more detail in section 4.

Grammatical profiling examines the relationship between the frequency distribution of grammatical forms and linguistic categories.

Semantic profiling examines the relationship between meanings (semantic tags) and forms.

Constructional profiling examines the relationship between the frequency distribution of grammatical constructions and the meaning of near-synonyms.

Linguistic profiles aim at the Big Questions, but are themselves agnostic about both the theory involved and the statistical methods used. Linguistic profiles are potentially portable across various approaches to linguistics. Although the theoretical perspective of this article is cognitive linguistics, linguistic profiles are not themselves restricted to any theoretical approach, and are thus portable across theories. Linguistic profiles do not specify a statistical model and can thus yield data that is amenable to analysis using various appropriate models. And while all the examples we examine below use Russian data, linguistic profiles are applicable to any language.

Linguistic profiles yield quantitatively measured results that represent real gains in our understanding of languages. This means that in addition to addressing questions of theoretical relevance, linguistic profiles can contribute to multipurpose applications. The results of linguistic profiling can be turned into resources for language learners and users and can inform intelligent (that is, rule-based, as opposed to statistical) machine translation. This is achievable through the creation of disambiguators and parsers that have a sophisticated model of a given language because they include detailed results from linguistic profiling analyses.

4. Examples of linguistic profiles

Three types of linguistic profiling analyses are illustrated by case studies in this section. For each case study, I identify the relevant Big Questions, describe how the questions are focused by theoretical perspective, and then give some details on how the question is operationalized and the type of statistical model used. I also comment on opportunities for portability and multipurpose application. Other types of linguistic profiling not described here include collostructional profiling (examining the relationship between a construction and the words that most frequently fill its slots [Kuznetsova 2013]) and radial category profiling (examining differences in the frequency distribution of uses across two or more near-synonyms; [Nesset et al. 2011; Endresen et al. 2012]).

4.1. Grammatical profiles: TAM in Russian

Tense, aspect, and mood (TAM) are known to interact in Russian, and linguists make many claims about how this works. Janda & Lyashevskaya [Janda, Lyashevskaya 2011] took these claims as hypothesis and tested them against data from the Russian National Corpus (www.ruscorpora.ru).

Very briefly, the facts of Russian TAM categories are as follows (for a fuller description and references, see [Janda, Lyashevskaya 2011]):

Tense. Russian has two tenses: a Past tense and a Non-Past tense that is usually interpreted as Present with Imperfective verbs, but as Future with Perfective verbs.

Aspect. All forms of all verbs express Perfective (marked) vs. Imperfective (unmarked) aspect. Most verbs have partners in “aspectual pairs” of verbs that share the same lexical meaning but differ according to aspect. Aspectual pairs can be formed via both prefixation (as in *писать* ‘write[imperfective]’ vs. ***на****писать* ‘write[perfective]’) and suffixation (*переписать* ‘rewrite[perfective]’ vs. *переписывать* ‘rewrite[imperfective]’). Approximately 1400 imperfective base stems form approximately 2000 perfective aspectual partners using sixteen prefixes, and approximately 20,000 perfective stems form imperfective partners using three suffixes. The suffixes are known to be distributed according to morphological classes of verbs, but there is controversy about the status of the prefixes. Most scholars consider the prefixes to be semantically “empty”, while a minority suggest that they are not empty and that aspectual pairs are formed only by suffixes.

Mood. Russian verbs have imperative forms, but Russian largely lacks modal verbs (except *мочь* ‘be able’). Infinitives often participate in modal constructions.

The Big Questions for this study are: *What is the relationship between form and meaning?* and *What is the relationship between lexicon and grammar?* More specifically, we are asking about the relationship between verb inflection (= form) and the grammatical meaning of aspect, and about the relationship between the lexical meanings of verbs and the grammatical meanings of the TAM categories.

From the perspective of cognitive linguistics, we can reformulate these questions for Russian TAM as follows:

- *Can we measure the expression of aspect according to the distribution of inflected forms?*
- *Can we distinguish between prefixation vs. suffixation in the formation of aspectual pairs?*
- *Can we measure the attraction of lexical classes of verbs to grammatical categories?*

We operationalize these questions by means of grammatical profiles. In order to introduce grammatical profiles, I will give a simple example of what a grammatical profile of a single verb might look like in English, a language with relatively little morphology. Let’s say that we take the verb *eat*, and we find the following distribution of forms in a corpus: *eat* – 749, *eats* – 121, *eating* – 514, *eaten* – 89, *ate* – 258. This distribution, visualized in Figure 1, would then be the grammatical profile of *eat*.

Russian of course has many more verb forms than English, as many as sixty-eight forms for a perfective verb, and 121 for an imperfective verb. But the majority of those forms are participles, where aspect is not in competition (since most types of participles are formed only from verbs of a given aspect). Gerunds are also limited in form by the aspect of the verb. The remaining forms can be gathered into four subparadigms: Non-Past, Past, Infinitive, and Imperative. Further distinctions made within these subparadigms (person, number, gender) are not expected to interact with aspect, so we can count forms for the grammatical profiles at the subparadigm level.

We collected approximately six million verb forms from verbs known to be aspectually “paired” either by means of prefixes or by means of suffixes. Our profiling analysis looks at the overall distribution of verb forms according to

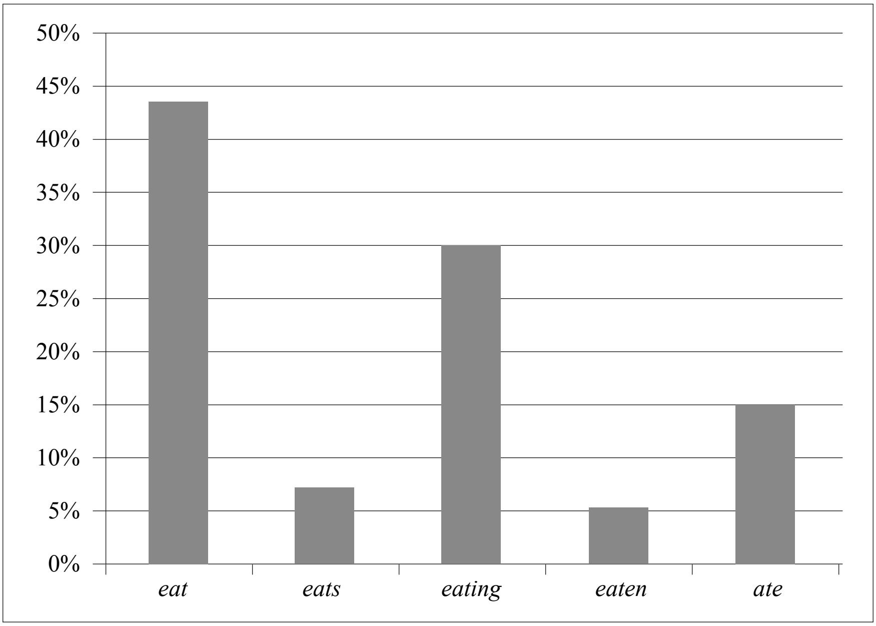

Figure 1. A hypothetical grammatical profile of English ***eat***

the aspect of verbs, and more specifically at the distribution of verb pairs using prefixes vs. suffixes. We furthermore probe the verbs most strongly attracted to given subparadigms for lexical patterns. We analyze the relationship between grammatical profiles and aspect using the chi-square test and Cramer's V effect size. Outliers in distribution plots indicate verbs showing the strongest attraction to given subparadigms.

Table 1 and Figure 2 show the aggregate distribution of grammatical profiles of Russian verbs in our study.

	Nonpast	**Past**	**Infinitive**	**Imperative**
Imperfective	1,330,016	915,374	482,860	75,717
Perfective	375,170	1,972,287	688,317	111,509

Table 1. Grammatical profiles of "paired" verbs in Russian (raw frequencies)

The chi-square test shows a highly significant effect for the relationship between aspect and grammatical profiles (chi-squared = 947756; df = 3; p-value < 2.2e-16). In other words, there is virtually no chance that we could get this distribution if there were no relationship. And the effect size is medium-large (Cramer's V = 0.399).

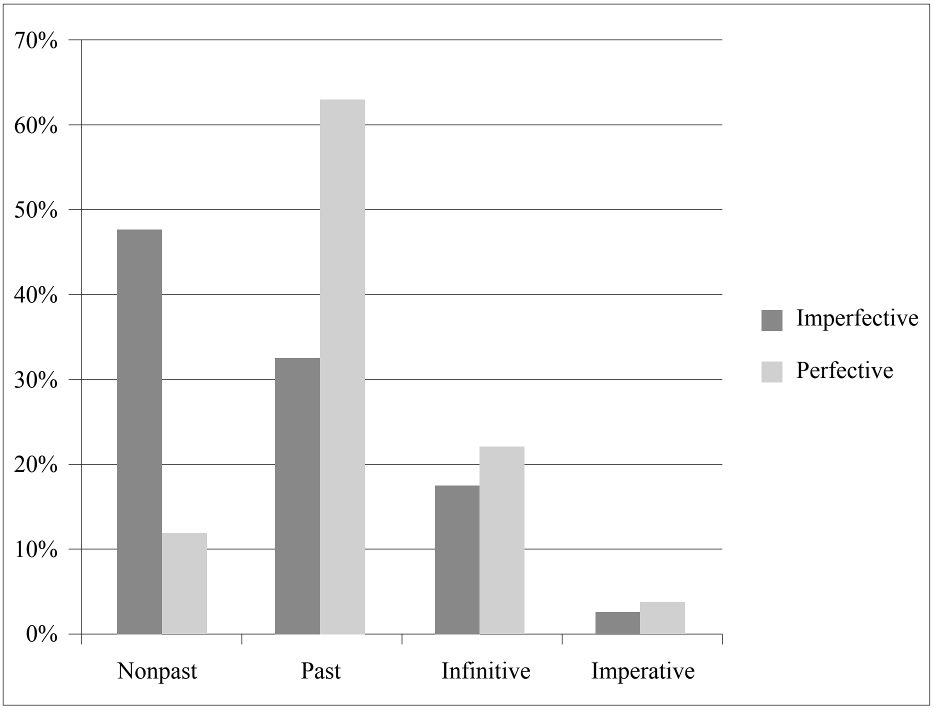

Figure 2. Visualization of grammatical profiles of “paired” verbs in Russian

While the finding that there is a relationship between aspect and grammatical profiles is perhaps not surprising [Comrie 1976: 73–84], we get more information when we disaggregate the profiles according to the type of morphology used to mark the aspectual distinction, namely prefixation vs. suffixation, as visualized in Figure 3.

In Figure 3, the dark bars show the grammatical profiles of verbs that use prefixes (p-partners) to mark aspect, while the light bars show the grammatical profiles of verbs that use suffixes (s-partners) to mark aspect. The top graph shows the profiles of imperfective verbs, while the bottom graph shows the profiles of perfective verbs. While there the slight deviations between prefixation and suffixation are statistically significant, the chi-square test is overwhelmed by the abundance of data. The effect sizes (0.076 and 0.037 for the two graphs, respectively) are more than an order of magnitude too small to make this a reportable effect. In other words, we find no difference in the behavior of prefixation vs. suffixation as a means to mark aspect in Russian. This of course does not mean that we can exclude the possibility that there is some other kind of difference between the two morphological markings, but this test does not reveal any difference.

With grammatical profiles of verbs we can look more deeply into the data and determine which lexical verbs are most attracted to various TAM combinations.

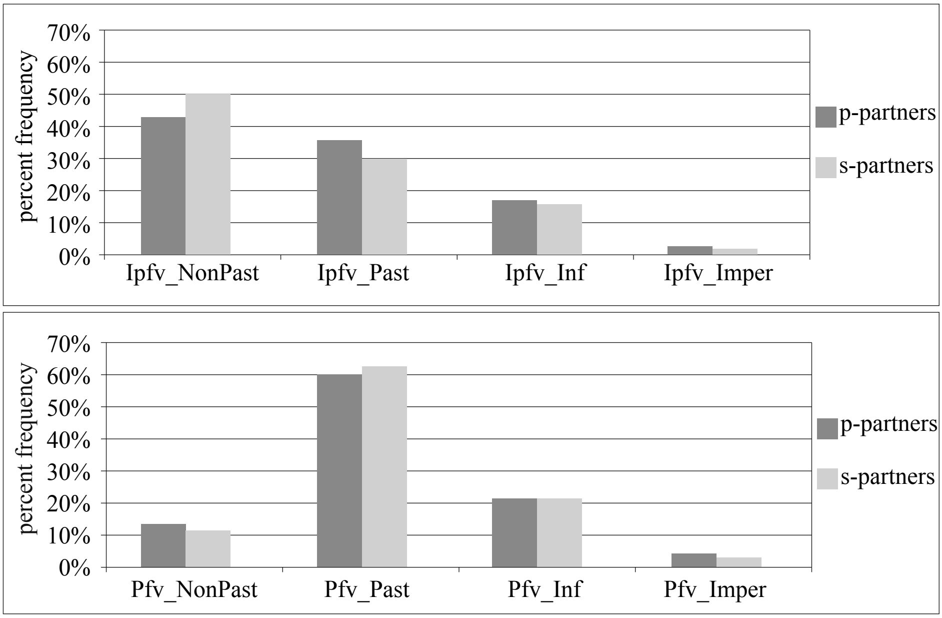

Figure 3. Grammatical profiles of "paired" verbs according to morphology

Let us take just one example, namely the relationship between the imperative mood and imperfective aspect. The scholarly literature on Russian aspect offers various hypotheses concerning the behavior of aspect in imperative forms, especially as concerns uses that are pragmatically marked as either polite or impolite [Падучева 1996: 12–17; Timberlake 2004: 374–375]. In the distribution of grammatical profiles, we find that over 200 imperfective verbs are extremely attracted to the imperative mood (qualifying as "outliers" because they lie more that 1.5 times above the interquartile range at the top of the distribution); some of these verbs are listed here. We find both items that confirm some of the hypotheses and some that could not be predicted from previous scholarship:

- **Polite, used when a guest knows what to expect:** *раздевайтесь* **'take off your coat',** *садитесь* **'sit down'**
- **Insistence, used when a hearer is hesitant:** *ступайте* **'get going',** *глядите* **'look',** *забирайте* **'take'**
- **Insistence, used when a hearer has not behaved properly (often in connection with negation):** *проваливай* **'get out of here',** *кончай* **'stop',** *не перебивай* **'don't interrupt'**
- **Polite requests:** *выручайте* **'help'**
- **Kind wishes:** *выздоравливайте* **'get well'**
- **Idiomatic:** *давайте посмотрим* **'let's take a look'**

- **Idiomatic/culturally anchored:** *прощай*(*те*) **'farewell',** *соединяйтесь* **'unite' (slogan),** *запевай* **'sing' (used as a command in the army)**

The first three uses, taken together, give us a new generalization, namely that the imperfective is used when the speaker assumes that the hearer should know what to do and needs some kind of encouragement rather than information. This generalization is much more informative than the traditional classification according to politeness, which yields contradictory expectations (as polite vs. insistent/ impolite).

Findings. Grammatical profiles show us that perfective verbs behave differently than imperfective verbs. However, "verb pairs" behave the same regardless of which type of morphology (prefixation vs. suffixation) is used to mark aspect. This finding contributes to an ongoing debate about the relative status of prefixes and suffixes in Russian verbal morphology. Furthermore, we can now identify precisely which verbs are most attracted to various TAM combinations in Russian.

Portability. Grammatical profiling is in principle applicable to any language that has grammatical paradigms, and can be used to address a variety of questions. For example, Kuznetsova [Kuznetsova 2013] uses grammatical profiling of past tense forms of Russian verbs (which mark gender) to address gender stereotypes, discovering which verbs are most highly attracted to masculine vs. feminine gender. While some of the results are unsurprising, such as that "masculine" verbs are associated with activities involving physical strength and high prestige, while "feminine" verbs are associated with childbearing and needlework, there are also some surprises, such as that women are often characterized by the sounds and movements of birds, and that men are characterized by vice and criminal activity. Eckhoff & Janda [Eckhoff, Janda 2014] used grammatical profiling of Old Church Slavonic verbs to probe the controversy over whether that language had perfective vs. imperfective verbs, as claimed by Dostál [Dostál 1954] among others. Their profiling analysis gives a resolution of verbs into two groups that agrees 96% with Dostál's designations.

Multipurpose applications. Given the results of grammatical profiling of Russian verbs, we can strategically fine-tune the teaching of Russian by focusing on high-frequency verbs that are highly attracted to each of the possible TAM combinations in Russian. This way students will be exposed to the most prototypical and representative examples of the use of perfective and imperfective aspect with the forms of verbs.

4.2. Semantic profiles: "Empty" prefixes in Russian

As described in 4.1, it is traditionally claimed that Russian prefixes are semantically "empty" when they create aspectual pairs, as in Table 2. Janda & Lyashevskaya [Janda & Lyashevskaya 2013] ask whether that claim is valid.

Imperfective base	Prefixed perfective
советовать ‘advise’	***по**советовать* ‘advise’
варить ‘cook’	***с**варить* ‘cook’
писать ‘write’	***на**писать* ‘write’
твердеть ‘harden’	***за**твердеть* ‘harden’
греметь ‘thunder’	***про**греметь* ‘thunder’

Table 2. Examples of perfectivizing prefixes that create aspectual pairs in Russian

The Big Questions for this study are: *What is the relationship between form and meaning?* and *Are there any semantically “empty” forms?* More specifically, we are asking about the relationship between prefixes (= form) and their meanings. Are the prefixes really as empty as claimed?

From the perspective of cognitive linguistics, we can reformulate the first question as follows: *Can we measure the relationship between prefixes and meanings of verbs?* More specifically, *What is the distribution of prefixes vs. semantic groups of verbs?* If we can show that this distribution is not random, and more importantly, that it shows semantically-motivated patterns, then we will have evidence of a meaningful relationship.

The second question can be reformulated as *How do we show that “empty” forms aren’t really empty?* This is a non-trivial question that is relatively difficult to answer. One way to approach the question of emptiness is to assume that emptiness is equivalent to zero content, and that zero has a unique value. Under these assumptions, if we find that two items are not identical in value, then at least one of them cannot be zero, since it makes no sense to claim that there are two zeroes with different values. In other words, if we can show that the prefixes behave differently, then we also have evidence that they are probably not semantically empty. So our goal is to show that prefixes have different semantic behaviors.

We operationalize these questions by means of semantic profiles.

Semantic profiling examines the relationship between meanings and forms. In this case study we used the semantic tags independently assigned to verbs in the Russian National Corpus and compared their distribution with the distribution of prefixes traditionally assumed to be “empty”.

Our data represents all 382 verbs that fulfilled three criteria: 1) they form perfective partner verbs using one of these five prefixes: *по-*, *с-*, *на-*, *за-*, *про-*; 2) they use only one prefix to form a perfective partner verb; and 3) that verb is assigned only one semantic tag in the Russian National Corpus. The data are visualized in Figure 4, where the bars show the portions of verbs with each prefix that are assigned to each of four semantic classes: “impact”, “change state”, “behavior”, and “sound & speech”.

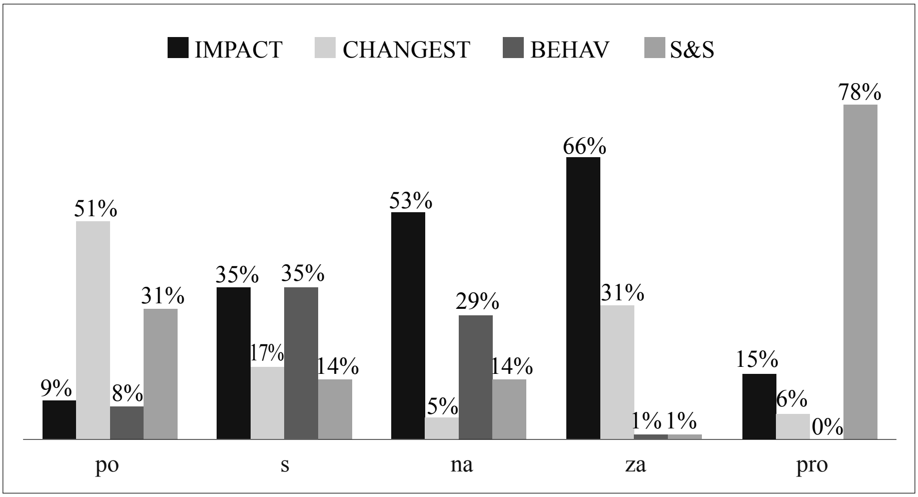

Figure 4. Semantic profiling of five Russian prefixes

This distribution proves to be highly significant with a very large effect size (chi-squared = 248, df = 12, Cramer’s V = 0.81).

Findings. These prefixes do indeed behave differently; in fact one could say that each prefix has a unique semantic profile. *По-* prefers change of state verbs like *побледнеть* ‘become pale’, although it also frequently co-occurs with speech verbs like *поблагодарить* ‘thank’. *С-* is rather evenly divided across impact verbs like *сшить* ‘sew’ and behavior verbs like *схитрить* ‘do something clever’. *На-* prefers impact verbs like *намылить* ‘make soapy’ but is often found also with behavior verbs like *нахулиганить* ‘behave like a hooligan’. *За-* also prefers impact verbs like *заплатать* ‘patch’, but its second choice is change of state verbs like *загрязнить* ‘make dirty’. *Про-* is used almost exclusively by sound and speech verbs like *прогреметь* ‘thunder’ and *прокричать* ‘yell’. Furthermore, the choice of verbs is clearly motivated by meanings of the prefixes that are more obvious when they are used to create specialized perfectives (verbs that are not merely perfective partners of the corresponding imperfective base verbs). For example, the verbs that choose *на-* are compatible with surfaces and acccumulation, *за-* is motivated by covering, etc.

Portability. In this study, semantic profiling was limited by the quantity of data available, as only five prefixes yielded sufficient numbers for a statistical analysis. But in principle semantic profiling could be applied to other prefixes as well as prefixes in other Slavic languages or even to any situation in which a set of markers is used to sort the units they attach to into groups (which might overlap to varying extents).

Multipurpose applications. At present, a second language learner of Russian is faced with the daunting task of mastering the correct combinations of over a dozen supposedly “empty” prefixes that create perfective partner verbs for over 1400 imperfective base verbs. If we can sort out the semantic profiles of the prefixes, this task can be made much more manageable. We can redesign our teaching materials that reduce the burden of memorization, offering instead a system that is at least partly coherent.

4.3. Constructional profiles: ‘Sadness’ in Russian

Janda & Solovyev [Janda, Solovyev 2009] looked at the relationship between sets of near synonyms and the grammatical constructions they appear in. They used constructional profiles to “measure” the distances between near synonyms, thus giving empirical substance to the unsatisfactory and sometimes contradictory claims in dictionaries. Here I will focus on just one of the two synonym sets from the original article, that of the six ‘sadness’ words *печаль*, *тоска*, *хандра*, *меланхолия*, *грусть*, *уныние*, which, unlike English, lack an “umbrella” term like *sadness* that would cover the whole group.

The Big Question here is: *What is the relationship between form and meaning?*, more specifically *What is the relationship between words and the larger linguistic units they participate in?* From the perspective of cognitive linguistics, we would expect that there are no exact synonyms and that each near synonym should show its own unique behavior. One type of behavior would be the grammatical constructions a synonym is found in. Thus our Big Question can be theoretically focused as *What is the relationship between near synonyms and the grammatical constructions they appear in?* This question already leads toward operationalization in terms of a relationship between near synonyms and their constructional profiles, namely the distribution of grammatical constructions the near synonyms are associated with.

For the purposes of this study, a grammatical construction for the six ‘sadness’ nouns was defined as the (*preposition* +) *case* construction that the noun appeared in, where case is obligatory, but the presence of a preposition is not. Russian has six grammatical cases, all of which can appear with a variety of prepositions, and all but one (the Locative) of which can also appear without any preposition. Taken together, the options for bare case and preposition + case yield approximately seventy grammatical constructions, though this count can vary due to differing definitions of what constitutes a preposition and the status of so-called “secondary” prepositions.

Combined data from the Russian National Corpus and the Biblioteka Maksima Moškova yielded 500 sentences for each of the near synonyms, all of which were coded for the (*preposition* +) *case* constructions that they appeared in. Figure 5 visualizes the results of this study, showing the five constructions that appeared most frequently in the data: *в* + Accusative, *в* + Locative, bare Instrumental, *с* + Instrumental, and *от* + Genitive.

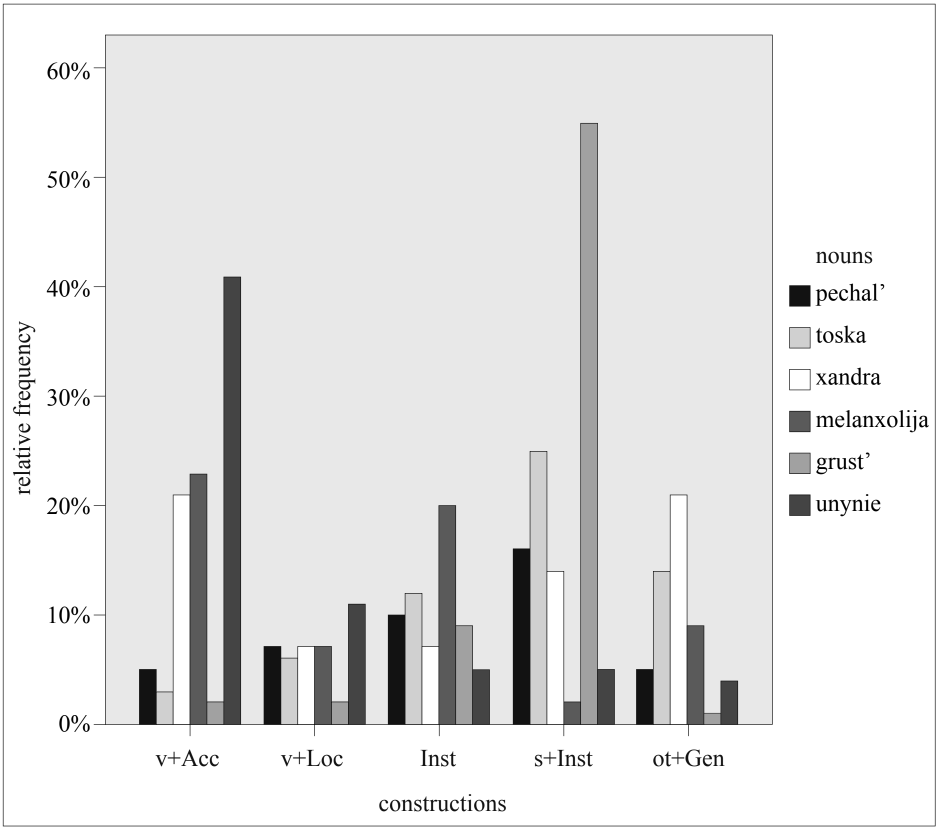

Figure 5. Constructional profiling of six Russian words for ‘sadness’

These results are also statistically significant with a medium-sized effect size (Chi-square = 730, df = 30, p < 0.0001, Cramer’s V = 0.3).

Findings. On the one hand, the near synonyms are in one sense very close to each other. The same five (*preposition* +) *case* constructions are the ones that are most important for each of the near synonyms. This is rather like having six different people choose at random the same five cards out of a deck of seventy. The chances of this happening are infinitesimally small. We compared the ‘sadness’ words with some other unrelated words, showing that other words were attracted to other groups of grammatical constructions. So the synonyms are indeed very close. However, there are also clear differences within this group of synonyms, showing that each noun has its own preferred set of grammatical constructions. For example, *грусть* is strongly attracted to the *с* + Instrumental construction, whereas *уныние* is strongly attracted to the *в* + Accusative construction. A hierarchical cluster analysis reveals that *печаль* and *тоска* are the closest synonyms in this group, followed by *хандра* and *меланхолия*, with *грусть* and *уныние* behaving as outliers by comparison. An additional finding is that the 'sadness' nouns in Russian show certain patterns in their metaphorical behavior: use in the *в* + Accusative and *в* + Locative constructions, for example, often corresponds to a metaphorical understanding of emotions as "holes", as in expressions like *он впал в уныние* 'lit. he fell into sadness'.

Portability. The constructional profiling approach can be applied to any group of synonyms that appear in a range of grammatical constructions, regardless of the language involved. Note also that the level of the grammatical construction can be adjusted as appropriate. In this case we chose a rather fine-grained level (that of the PP or NP), but that both finer (within a word) and coarser (clause-level or discourse-level) levels might be more appropriate for a given study. Some other examples of this type of profiling analysis include:

- constructional profiling of *нагрузить*, *погрузить*, and *загрузить*, all meaning 'load', across the "Locative Alternation" constructions: *грузить ящики на телегу* 'load boxes onto the cart' vs. *грузить телегу ящиками* 'load the cart with boxes' [Sokolova, Lyashevskaya, Janda 2012], showing that the three prefixes behave quite differently;
- constructional profiling of aspectual pairs formed by the prefix *про-* [Kuznetsova 2013], showing that "pairedness" for aspect may be a scalar rather than an absolute phenomenon in Russian.

Multipurpose applications. Distinguishing among near synonyms and knowing when it is most appropriate to use which one rank among the greatest challenges for advanced language learners, and constructional profiling analyses could certainly inform valuable teaching materials targeting this problem. Lexical selection is a parallel problem in machine translation that could be improved by taking the constructional profiles of individual words into account.

5. Conclusions and implications

In sum, linguistic profiles help us to bridge the gap between Big Questions that come into our theoretical focus and operationalization that facilitates addressing such questions. In this way, linguistic profiles make it possible to fruitfully combine theoretical inquiry with quantitative research. And the results can be not only theoretically interesting, but useful as well.

Here again are the two Big Questions that relate to the three case studies presented above:

1) What is the relationship between form and meaning?

2) What is the relationship between lexicon and grammar?

The answers to both questions are certainly that these relationships are both tight and complex and deserve much more research. Within the scope of these questions we offer several specific findings about Russian:

1) Perfective verbs behave differently from imperfective verbs, regardless of the morphology used to signal aspect;
2) We can identify the verbs that are most attracted to certain verbal subparadigms;
3) Verbal prefixes are most likely not semantically "empty" as presumed, instead each prefix has a unique semantic profile;
4) We can identify the grammatical profiles that given nouns are most attracted to.

These results shed light on theoretical issues, and also make it possible to improve both our pedagogical and our computational approaches to Russian.

References

Comrie B., 1976, *Aspect*, Cambridge: Cambridge University Press.

Divjak D., Gries S. Th., 2006, Ways of trying in Russian: Clustering behavioral profiles, *Corpus Linguistics and Linguistic Theory*, 2, p. 23–60.

Dostál A., 1954, *Studie o vidovém systému v staroslověnštině*, Prague: Státní pedagogické nakladatelství.

Eckhoff H. M., Janda L. A., 2014, Grammatical Profiles and Aspect in Old Church Slavonic, *Transactions of the Philological Society*, Vol. 112, Issue 2, p. 231–258.

Endresen A., Janda L. A., Kuznetsova J., Lyashevskaya O., Makarova A., Nesset T., Sokolova S., 2012, Russian 'purely aspectual' prefixes: Not so 'empty' after all?, *Scando-Slavica*, 58:2, p. 231–291.

Evans N., Levinson S., 2009, The myth of language universals: Language diversity and its importance for cognitive science, *Behavioral and Brain Sciences*, 32, p. 429–492.

Firth J. R., 1957, A synopsis of linguistic theory 1930–1955 [in:] J. R. Firth et al. (eds.), *Studies in Linguistic Analysis*, (Philological Society), Oxford: Blackwell, p. 1–32.

Frank S. L., Bod R., Christiansen M. H., (2012), How hierarchical is language use?, *Proceedings of the Royal Society B*, 279, p. 4522–4531.

Geeraerts D., Grondelaers S., Speelman D., 1999, *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen*, Amsterdam.

Gries S. Th., Divjak D., 2009, Behavioral profiles: a corpus-based approach towards cognitive semantic analysis [in:] V. Evans, S. S. Pourcel (eds.), *New Directions in Cognitive Linguistics*, Amsterdam: John Benjamins, p. 57–75.

Hanks P., 1996, Contextual dependency and lexical sets, *International Journal of Corpus Linguistics*, 1, p. 75–98.

Harris Z. S., 1970, *Papers in structural and transformational linguistics*, Dordrecht: Reidel.

Janda L. A., 2013, Quantitative Methods in *Cognitive Linguistics* [in:] L. A. Janda (ed.), *Cognitive Linguistics: The Quantitative Turn. The Essential Reader*, Berlin: De Gruyter Mouton, p. 1–32.

Janda L. A., 2015, Cognitive linguistics in the year 2015, *Cognitive Semantics*, Vol. 1, p. 131–154.

Janda L. A., Lyashevskaya O., 2011, Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian, *Cognitive Linguistics*, 22:4, p. 719–763.

Janda L. A., Lyashevskaya O., 2013, Semantic Profiles of Five Russian Prefixes: po-, s-, za-, na-, pro-, *Journal of Slavic Linguistics*, 21 (2), p. 211–258.

Janda L. A., Solovyev V., 2009, What Constructional Profiles Reveal About Synonymy: A Case Study of Russian Words for SADNESS and HAPPINESS, *Cognitive Linguistics*, 20 (2), p. 367–393.

Kuznetsova J., 2013, *Linguistic Profiles. Correlations between Form and Meaning*. PhD dissertation, University of Tromsø.

Nesset T., Endresen A., Janda L. A., 2011, Two ways to get out: Radial Category Profiling and the Russian Prefixes *vy*- and *iz*-, *Zeitschrift für Slawistik*, 56:4, p. 377–402.

Sokolova S., Janda L. A., Lyashevskaya O., 2012, The Locative Alternation and the Russian 'empty' prefixes: A case study of the verb *gruzit'* 'load' [in:] D. Divjak, St. Th. Gries (eds.), *Frequency effects in language representation*, (Trends in Linguistics. Studies and Monographs, 244.2), 2012, Berlin: Mouton de Gruyter, p. 51–86.

Speelman D., Grondelaers S., Geeraerts D., 2003, Profile-Based Linguistic Uniformity as a Generic Model for Comparing Language Varieties, *Computers and the Humanities*, 37 (3), p. 317–337.

Timberlake A., 2004, *A Reference Grammar of Russian*, Cambridge: Cambridge University Press.

Падучева Е. В., 1996, *Семантические исследования. Семантика времени и вида в русском языке. Семантика нарратива*, Москва: Языки русской культуры.