Corpus linguistics and phraseo-paremiology

Eric Laporte

Corpus linguistics and phraseo-paremiology

Éric Laporte
Universidade federal do Espírito Santo
Université Paris-Est Marne-la-Vallée

You can tell me that the facts
prove the contrary. I will answer:
Too bad for the facts
Nelson Rodrigues

If corpus linguistics and phraseo-paremiology are two fields of linguistics, how are they interlinked and how can they interplay? In this paper, we defend the position that it is the interest of specialists of phraseo-paremiology to practice corpus-linguistic methods, without excluding other methods[1].

Section 1 advocates that the practice of corpus analysis is a criterion of quality. In Section 2, we discuss the merits of various approaches to fact observation in phraseo-paremiology. Section 3 is about forms of interaction between the different methods.

1. Corpus analysis as a criterion of quality

More and more linguists analyse corpora of preexisting texts in order to observe linguistic forms. This happens in several steps of their research, for example:

- during an initial step in which they collect facts;

- or in order to check results or assessing formal representations such as lexicons and grammars: for example, in Laso (2009), grammatical knowledge about support verb constructions point to the phrases *make/provide/produce contribution* as possible paraphrases of *contribute*: the study investigates a corpus to control whether they do actually work as such in texts.

Corpus analysis ensures that linguistic facts in context of use are taken into account. Therefore, practice of corpus analysis during linguistic research has become a criterion of quality. And the advantages of implementing corpus-linguistic methods are at least as valid in phraseo-paremiology as in other fields of linguistics. The increasing practice of corpus analysis belongs to a historical progress, the advent of empiricalness in linguistics.

With all respect due to researchers' methodological freedom, this practice is so common and appreciated that it is now considered as almost indispensable for any linguistic research that involves observation of forms. In language processing, corpus-based assessment of results is practically imposed by journals and conferences.

Corpus linguistics has additional relevance to several notions which are difficult to formalize, such as semantic intensity. For example, the Brazilian Portuguese (BP) phrase *de mão cheia* (lit. with full hand) 'excellent [at some human activity]' expresses intense admiration. Formal approaches to linguistics do not easily account for this scalable element of meaning. In contrast, scalable phenomena are familiar to the world of corpus linguistics, which efficiently collects and interrelates examples of them. Connotation is another example. By 'connotation', we mean impressions and affective judgments which are felt as being only suggested by a linguistic form: the BP idiom *quebrar um galho* (lit. break a branch) 'make

[1] We acknowledge the invitation of the Organizing Committee of the International Congress of Phraseology and Paremiology, and the support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Board for Enhancement of Staff of Superior Level, CAPES).

do', 'help out' implies an idea of relieved frustration[2]. Stylistic preference is a third example: the phrases *in terms of, in regard to, in/with respect to/of, in/with regard to, in/with reference to, concerning, regarding* fit differently in document types and syntactic contexts (Rankin, Schiftner, 2009).

However, corpus-linguistics methods for phraseo-paremiology have limitations which can be surpassed with the aid of grammatical and formal methods. When we investigate variations of idioms, corpus-based studies rely on the actual occurrence of the variants in the corpus. For example, does the BP sentence *As portas da companhia estão abertas* ('The doors of the company are open', 'The company is running') have syntactic variants? As some variants in use may be absent from the corpus being used, a practical alternative for native speakers is to generate forms according to grammatical models, e.g. the form with an agentive subject *abrir as portas de uma companhia* ('open the doors of a company', 'start running a company'), and to check them with the aid of introspection (cf. Section 3.1).

As a matter of fact, corpus linguistics combines successfully with complementary methods through various intermethodological bridges. Seretan (2009) studies collocations collected from a corpus and submits them to a process of formal description before integrating them into a lexicon for automatic translation. Laso (2009) uses introspective observation to select an initial set of paraphrases of *contribute* before examining their occurrences in a corpus. The advantages of a method compensate for the limitations of another.

Such intermethodological bridges are frequently and successfully crossed, but they are not mentioned often by corpus linguists or much valued in scientific discourse, as if to implicitly suggest that their favourite methodology should ideally be exclusive. This rhetorical option can be explained by too reasons. First, corpus linguistics has a prestigious, modern image. Second, it had to fight for its existence in the 1960s and 1970s in the context of its rivalry with generative grammar, and has developped a tradition of pugnacious rhetoric in this context.

From a scientific point of view, the fact of resorting exclusively to corpus linguistics as a methodology of observation is a decision with scientific consequences. Let us return to the example of the variants of *As portas da companhia estão abertas* ('The company is running'). Depending on whether we limit ourselves to forms attested in a corpus, the research will produce divergent outcomes, since the variants in use may occur or not in the corpus.

Summing up our position, all linguists should be corpus linguists; but do they need to give up their other activities, methods and tools?

2. Corpus linguistics and other approaches

Our point of departure in this section are a few expressions frequently used to emphasize merits of corpus linguistics. We discuss whether these merits offer a ground to exclude other approaches: are there any empirical approaches besides corpus linguistics? do they take into account contexts of use? are they based on a real facts? are they good at achieving recall?

2.1. Empiricalness

Corpus linguistics claims, obviously with good cause, to be a discipline for empirical observation. As Garrão (2011) puts it, for example, "o critério estatístico a partir de corpus

[2] By 'connotation' we do not mean the figurative meaning of a lexicalized idiom. One of the effects of lexicalization is that the figurative meaning, here 'make do', 'help out', becomes the strict, explicit meaning of the idiom. In the BP sentence *Dilma tomou posse como presidente* 'Dilma was inaugurated as president', the expression has a connotation of an active part, whereas in English and in the French translation *Dilma a été investie dans ses fonctions de présidente*, it has a connotation of a passive part, though the strict, explicit meaning is the same.

(...), como alternativa a uma abordagem fortemente baseada na intuição do pesquisador, [é] altamente promissor" (the corpus-based statistical criterion (...), as an alternative to an approach strongly based on the researcher's intuition, [is] highly promising).

This naturally leads us to wonder whether approaches based on the researcher's intuition are also promising or not, and if they are valid empirical processes? This opens the long-debated epistemological question of introspective linguistics, i.e. linguistics relying on self-examination of one's linguistic competence. Clearly, the answer depends on how introspective approaches are implemented: there exist bad and good practices. Among good practices (Laporte, 2008), we can cite:

- study of languages by trained native speakers with a regular descriptive activity,
- collective control by peers,
- use of formal criteria,
- assessment of reproducibility of criteria,
- taking into account contexts of use.

Such provisions are typical of empirical investigation. When they are put to practice, they justify at least some trust in a specific form of empirical observation. Condemning any form of introspective linguistics because of bad practices would be a hasty generalization — in other words, a prejudice.

2.2. Consideration of contexts of use

A *locus communis* of corpus linguistics is an opposition between 'contexts of use', or 'language in use', and 'Chomskyan introspection', which is of a 'grammatical' nature. Sure, the observation of corpora provides researchers with contexts of use, and with examples of language in use.

The opposition which has become a *locus communis* raises a question: is introspection also able to take into account contexts of use? Again, this depends on the skill, training and motivation of introspective linguists. Our experience is that the ability to take into account contexts of use depends on memory, because native speakers have to recall linguistic forms which they have been exposed to, and on imagination, since contexts of use correspond to human situations: linguists equipped with memory and imagination can get trained to relate linguistic forms and contexts of use.

The opposition mentioned above, as it mentions Chomsky, raises another question: is introspection exclusively Chomskyan?

The main type of introspective judgment used in generative grammar is grammaticality, which assumes the existence of a speaker-internal grammar. Chomsky (1957) stated that grammar is autonomous and independent of meaning, illustrating this with the meaningless sequence *Colorless green ideas sleep furiously* which is deemed grammatical. This approach tends to consider contexts of use as irrelevant.

Harris (1952) has another view of introspective judgment, and uses another term, acceptability: native speakers can assess to which point a form is in use, taking into account contexts of use, among other constraints; the acceptable forms are those in use. With this definition, **Colorless green ideas sleep furiously* receives the mark of inacceptability, '*'. In this perspective, grammar is not an autonomous entity postulated by the theory, but a target of empirical investigation. Harris' conception of syntax and semantics inspired a large enterprise of description of languages, the Lexicon-Grammar (Gross, 1984), which remained almost only introspective until the 1990s.

In other words, introspection in linguistics is not exclusively a Chomskyan notion, and does not necessarily disregards contexts of use.

2.3. Reality of facts

A massive merit of corpus analysis is that it grounds investigation on real, authentic data: "[com] recursos estatísticos (...), contamos com um valioso aliado (...): dados reais da língua" ([with] statistic resources (...), we have a valuable ally (...): real language data) (Garrão, 2011). What about approaches without statistics? can they take advantage of real data?

One of the well-known motivations for adopting corpus linguistics is the common observation that many dictionaries and grammars contain errors or lack important information, which makes them imperfect sources. However, corpus-based statistical data have well-known limitations and imperfections too. In addition, the existence of errors is not sufficient to invalidate the methods of construction of these dictionaries and grammars: what should be incriminated might also be the way these methods have been implemented, rather than the methods themselves.

Another observation in favour of the use of corpora is that dictionaries (for human readers) made with the aid of corpus-based statistical data are often of a better quality, because such data provide abundant facts to authors, which facilitates their work. But, again, this does not rule out the relevance of other approaches when they bring complementary information.

The question of grammars containing errors is worth looking deeper into. Matuda (2011) warns about "formas permitidas pela gramática mas que não ocorrem" (forms allowed by the grammar, but which do not occur). As a matter of fact, if the grammar does not predict which forms occur, it is invalid or incomplete, since grammars are precisely made to this end[3]. However, this naturally raises another question: is it possible to complete or improve grammars so that they predict which forms occur? Before trying to answer, observe that Matuda's warning above is about "a gramática" ('the grammar'), whereas our own question mentions "grammars". In fact, what does Matuda mean by "a gramática" ('the grammar')?

- a concept preexisting to linguistic research? but then, which concept? if she means the specifically generativist concept of grammar, only generativist approaches are concerned by her criticism;

- a result of investigation? but then, writing "uma gramática" ('a grammar') or "gramáticas" ('grammars') would be more accurate, since a grammar usually does not predict exactly the same forms as another.

Thus, a relevant question is: is it possible to complete or improve grammars so that they predict which forms occur?

Answering this question is equivalent to assessing a whole field of linguistics: manual construction of formal grammars, the research goal of a number of scholars from the 1950s until now. If a grammar is formalized and readable, there is no reason why it would be impossible to correct an error in it. This is what grammar authors do. Among the major challenges they are collectively facing are… phraseological units.

To take on this challenge, they need to rely on real facts. Corpus linguistics is one of the approaches that can massively provide reliable data, but statistical data do not automatically answer all the questions. Grammar authors seek complementarity in introspective sources, and an important part of their activity consists in turning examples into rules, i.e., assessing how general is each piece of information, which often requires introspection too.

[3] The problem of errors in grammars are precisely the forms allowed by grammars, but which do not occur. (And the forms excluded by grammars, but which occur.)

2.4. Recall

Corpus linguistics is reputed to contribute to extending the range of forms taken into account in linguistic research, increasing the recall of descriptions. As a matter of fact, in terminological and phraseological studies, corpus analysis easily provides numerous expressions and collocations, many of which prove to be relevant (Teixeira, Tagnin, 2011). Hence the following question: can methods without statistics achieve high lexical coverage?

The answer is complex and requires reviewing literature and current practices.

As regards inventories of phraseological units, corpus analysis is a time-effective option, widely adopted, for example, for collocations in terminology. However, it is not strictly indispensable: comprehensive phraseological inventories have been compiled for natural language processing lexicons without the aid of any corpus (Gross, 1986). In addition, inventorying syntactic variants of phraseological units is almost as useful as registering the most common form, and this is what is done in lexicology with a commitment to quality. Corpus linguistics, which is more efficient for frequent forms than for rare forms, encounters here an inherent limitation: data sparseness. Technically, data sparseness corresponds to the situation in which numbers of occurrences of forms under study in the corpus are insufficient to ensure statistic significance. Such a situation can stem from variation in form: if an expression occurs in numerous forms, occurrences are scattered over variants, and this reduces numbers of occurrences. The rarest variants are not sufficiently numerous. They can even be absent from the corpus. For example, Fritzinger & Heid (2009) address a problem of data sparseness affecting German collocations:

Verfahren einleiten	'file a sue'
eingeleitetes Verfahren	'sue being filed'
Einleitung des Verfahrens	'filing the sue'
Verfahrenseinleitung	'sue filing'

In their study, corpus linguistics benefits from grammatical and formal knowledge of syntactic and morphological variation: such knowledge allows for aggregating forms, 'undoing' the variation and thus contributing to resolve the problem. In that case, data sparseness is resolved with the aid of data of introspective origin.

Introspection, strangely, does not have an analogous problem with rare forms: it is almost as reliable for rare forms as for frequent forms.

Let us mention another facet of lexical coverage: inventorying senses of words. Corpus analysis, here again, provides valuable help in the form of examples of uses of a given word in various contexts. But, here again, good-quality lexicology always inventoried word senses, with or without the aid of corpora.

3. Interaction with other linguists

The main point of the preceding sections was to highlight the complementarity between corpus linguistics, introspective observation and formalization. More and more researchers take advantage of this complementarity: for example, corpus linguists do use data of introspective origin (e.g. Fritzinger, Heid, 2009; Laso, 2009; Rankin & Schiftner, 2009; Seretan, 2009...).

The contrasts and interactions between grammar and corpus linguistics suggest a duality between them, and situate phraseo-paremiology in the heart of this duality. This duality should be taken as an opportunity of fertile collaboration, rather than of competition between scientific communities. Corpus linguistics is now a respected field which does not need to gain approbation to flourish.

In this section, we discuss forms of effective collaboration. The main force of corpus linguistics is fact observation: can other linguists 'farm out' observation activities to corpus

linguists? Is it better for corpus linguistics to remain separate, or to integrate with other activities?

3.1. Description

Take the example of a descriptive activity: construction of lexicons for language processing. Such lexicons are a crucial resource for machine translation with symbolic approaches, and for other much-demanded computer applications such as search engines and information management (Tolone, Sagot, 2011). In this context, describing phraseological units is one of the problems faced by language-processing lexicologists. They meet huge needs for observation of facts.

For each phraseological unit, meanings must be inventoried. Idioms are less ambiguous than simple words, but some do have several senses, to be represented as independent lexical items (Xatara, 2012):

(1) *O empreendedor **abre as portas** de sua loja*
'The entrepreneur opens the doors of his shop'

(2) *Aquilo me **abriu as portas** para a carreira de atriz*
'This opened doors for me on a career as an actress'

For each item, syntactic variants must be inventoried. In practice, this involves checking all known variation types, because even idioms with similar grammatical structures do not necessarily have the same variants (Xatara, 2012):

O empreendedor abre as portas de sua loja
'The entrepreneur opens the doors of his shop'

(3) *As portas da sua loja vão se abrir*
'The doors of his shop will open'

(4) *Sua loja abre as portas*
'His shop will open its doors'

Dunga abaixou a bola do adversário
'Dunga reduced the pretensions of his opponent'

**A bola do adversário se abaixou*
'The pretensions of his opponent were reduced'

O adversário abaixou a bola
'His opponent reduced his own pretensions'

The inventory of syntactic variants is part of the description of the formal behaviour of an idiom. In addition, it may determine its assignment into a category. For example, BP constructions with the verb *fazer* and a noun may or may not be support verb constructions, depending on whether a variant of the construction with the same meaning without *fazer* may occur (Langer, 2005):

*O Munícipe **fez questão** de parabenizar publicamente a cidade* (Garrão, 2011)
'The citizen was keen on publicly congratulating the town'

*Depois da **questão que fez** o Munícipe de parabenizar publicamente a cidade...*
'After the citizen keenly congratulated the town publicly…'

This expression allows for forming a relative clause, but the reduction of the relative with deletion of *fazer* does not work:

**Depois da questão do Munícipe de parabenizar publicamente a cidade...*
*After the citizen's keen on publicly congratulating the town…

Since no other operation erasing *fazer* is applicable, the phrase cannot be classified as a support verb construction, but rather as an idiom. In contrast, the following support verb construction admits the corresponding variants:

A Igreja Adventista do Sétimo Dia ***fez acordo*** *teológico com a Igreja Católica* (Garrão, 2011)

'The Seventh-day Adventist Church made a theological agreement with the Catholic Church'

Depois do acordo teológico da Igreja Adventista do Sétimo Dia com a Igreja Católica...

'After the theological agreement of the Seventh-day Adventist Church with the Catholic Church…'

The inventory of phraseological units, and for each of them the inventory of senses and of syntactic variants, make up a descriptive work that involves essentially fact observation, which suggests that corpus analysis is relevant. However, the target of this research is a multitude of highly specific pieces of information, and not all of them automatically emerge from raw statistics, for two reasons. First, some pieces of information depend on the identification of a meaning among several ones: for example, distinguishing between examples (1) and (2) above requires an introspective examination of context. Second, some syntactic variants in use, and even some phraseological units at all, may happen to be absent from the corpus used for the research, perhaps examples (3) or (4).

In such complex descriptive work, corpus analysis is welcome or even required, but needs close coordination with the other tasks at stake: prediction of potential syntactic variants (with the aid of formal grammatical knowledge), assignment of occurrences to senses (with the aid of introspection). It is really difficult to believe that such close interaction could be achieved by 'farming out' observational tasks between separate groups of linguists.

3.2. Formalization

When the descriptive activity mentioned in the preceding section produces lexicons and grammars to be used in computer applications with symbolic approaches, the process needs to include a formalization step. Formalization consists in adjusting the linguistic information to a simplifying formal model, which can be handled by software.

Many linguistic phenomena can be situated on a continuous scale, for example that of more or less strict selection restrictions: one end of the scale corresponds to highly frozen phrases, and the other to compositional constructions. For machine translation, in the frozen end, what is recorded is the translation of a word sequence: *lip service/aprovação fingida*; in the free end, it is the translation of a word: *challenge/desafio/défi,* or *bike/bicicleta*. For intermediate cases, there are intermediate solutions, for example recording the translation of a word in presence of others: *ride [a bike]/andar [de bicicleta]*, or *rise to [a challenge]/responder [a um desafio]/relever [un défi]*. But these formal solutions are in finite number. They do not make up a continuous scale, but a discrete one. In such cases, language processing requires the choice of a formal model.

Seretan (2009) addresses this problem by distributing lists of collocations into a finite number of types. Corpus-derived statistical facts are useful for this task, but the transition from raw statistical facts to a model is not automatically provided by corpus linguistics: it belongs rather to formal grammar. Here again, an integration of methods between corpus linguistics and formal grammar seems relevant.

Conclusion

The conclusions of this reflection about the relations between corpus linguistics and phraseo-paremiology are not surprising. The two fields have strong natural connections. There

exists a duality between corpus linguistics and complementary approaches: introspective observation, formalization. It should be viewed as an opportunity of close interaction between the respective linguists and between their methods.

Bibliographical references

Chomsky, Noam, 1957. *Syntactic Structures*, The Hague/Paris: Mouton.

Fritzinger, Fabienne; Heid, Ulrich, 2009. "Automatic grouping of morphologically related collocations", *Proceedings of the Corpus Linguistics Conference*, University of Liverpool.

Garrão, Milena de Uzeda, 2011. "A identificação de expressões fixas verbais com base em *corpora*", *Congresso internacional de Fraseologia e Paremiologia*, Universidade de Brasília, abstract, p. 132-133.

Gross, Maurice, 1984. "A linguistic environment for comparative Romance syntax", *Papers from the XIIth Linguistic Symposium on Romance Languages*, Amsterdam/Philadelphia: John Benjamins, p. 373-446.

Gross, Maurice, 1986. "Lexicon-grammar. The representation of compound words". *Proceedings of COLING*, University of Bonn, p. 1-6.

Harris, Zellig, 1952. "Discourse analysis". *Language* n. 28:1, Linguistic Society of America, p. 1-30.

Langer, Stefan, 2005. "A linguistic test battery for support verb constructions". *Lingvisticae Investigationes* 27(2), p. 171-184.

Laporte, Éric, 2008. "Exemplos atestados e exemplos construídos na prática do léxico-gramática", *Revista (Con)textos Lingüísticos* 2, p. 26-51. http://halshs.archives-ouvertes.fr/halshs-00325926

Laso Martín, Natalia Judith, 2009. "A corpus-based study of the phraseological behaviour of abstract nouns in medical English", *Proceedings of the Corpus Linguistics Conference*, University of Liverpool.

Matuda, Sabrina, 2011. "Extração de unidades fraseológicas especializadas em *corpora* comparáveis", *Congresso internacional de Fraseologia e Paremiologia*, Universidade de Brasília, abstract, p. 94-95.

Rankin, Tom; Schiftner, Barbara, 2009. "The use of marginal and complex prepositions in learner English", *Proceedings of the Corpus Linguistics Conference*, University of Liverpool.

Seretan, Violeta, 2009. "An integrated environment for extracting and translating collocations", *Proceedings of the Corpus Linguistics Conference*, University of Liverpool.

Teixeira, Elisa Duarte; Tagnin, Stella Esther Ortweiler, 2011. *Vocabulário para culinária inglês-português*, Série Mil & um Termos, São Paulo: SBS.

Tolone, Elsa; Sagot, Benoît, 2011. "Using Lexicon-Grammar tables for French verbs in a large-coverage parser". In Zygmunt Vetulani, editor, *Human Language Technology. Challenges for Computer Science and Linguistics. 4th Language and Technology Conference (LTC 2009), Poznań, Poland, November 6-8, 2009, Revised Selected Papers*, volume 6562 of Lecture Notes in Artificial Intelligence (LNAI), p. 183-191, Berlin: Springer.

Xatara, Maria Cláudia, *Dictionnaire multilingue des expressions figées*, available at <http://www.lexicographie.fr/Idioms/>, looked up in May 2012.