Testing usage-based theories with a representative corpus of nineteenth-century French

Angus B. Grieve-Smith (Columbia University, USA)

The researchers who compiled the corpus for the *Trésor de la langue française* dictionary in the 1960s, which became the FRANTEXT (2017) corpus that we use today, aimed to serve an audience drawn from "the upper and middle ranks of society" who desired to produce "careful enunciations, obeying empirical norms as much as they depart from belabored clauses" (Imbs 1971: XVIII). With that aim, they built the corpus around a "principle of authority," examining literary histories of the nineteenth and twentieth centuries and collecting texts that were mentioned multiple times in those histories (ibid: XXIII).

FRANTEXT is large and well-made, which has earned it a place as one of the most popular corpora of the language. It appears to work well for helping its target audience produce their careful enunciations. But other corpus users, such as researchers in usage-based grammar, have needs that diverge widely from those of the audience of the *Trésor de la langue française*, and they would be better served with a different corpus design.

Usage-based theories, such as the theory of grammaticalization (e.g, Bybee and Thompson 1997), posit that the structures of a language have emerged from earlier states of that language. For example, when two constructions are in competition, the one with the higher type frequency - the greater mindshare - tends to increase in type frequency in subsequent years, while the other decreases. But the state of the language is not restricted to well-crafted works of literature as identified by later literary historians. Language users come from all social classes and have been influenced by spontaneous language, vernacular literature, and a kind of writing that is most lacking in literary corpora: bad literature.

To properly test usage-based theories we would need a corpus of every utterance that, for example, Alexandre Dumas fils was exposed to up to the time he wrote each of his plays. Collecting this data is beyond the abilities of science, but we can build a corpus that suits these purposes better than FRANTEXT. The Digital Parisian Stage Corpus aims to be that corpus.

The Digital Parisian Stage builds on an exhaustive catalog compiled by Charles Beaumont Wicks (1950 et seq.) of all plays that premiered in public in Paris in the nineteenth century, totaling over 30,000. The first phase of the project consists of a sample of thirty plays from the period between 1800 and 1815. This is a random one percent sample of the 2980 plays listed in Wicks' first volume (1950).

Of this sample, the scripts for twenty-four plays have been obtained from Google Books, Gallica and other sources. Eighteen of them have been processed with Optical Character Recognition (OCR), and fifteen cleaned. Of those fifteen, three have been determined to be too short for most studies, leaving ten currently available for annotation.

The difference between the two corpora can be seen by brief investigations of well-known variables. For example, on average in the four theatrical texts for this period in FRANTEXT, 49% of negated declarative sentences used *ne … pas*, 21% *ne … point*, and 30% ne alone. In the twelve plays currently available from the Digital Parisian Stage Corpus, we find on average 75% ne … pas, 10% ne … point and 15% ne alone ($p < 0.01$).

A closer look at the texts can offer an explanation for this difference. Here is a quote from the play *Pinto* (Lemercier 1800):

(1) LA DUCHESSE: Ainsi votre esprit s'environne de tous les obstacles qu'il se crée; et si vous ***n'***en aviez de véritables à surmonter, où seroit la gloire de l'entreprise!

In the bolded section, the ne alone construction is used in a subordinate clause activating a presupposition, a pattern found in many earlier texts and one of the most common contexts for this construction. Contrast this with a quote from *le Grenadier de Louis XV* (Dubois 1815):

(2) ANSELME: C'est être bien hardi, après toutes les menaces que vous avez osé me faire si je *ne* vous donnais *pas* ma fille...

In this example, the *ne…pas* construction is used in a very similar context. It is the use of *ne…pas* in these contexts where *ne* alone had predominated that accounts for the much higher percentage of *ne…pas* in *le Grenadier de Louis XV* than in *Pinto*.

The higher proportion of *ne …pas* in the randomly selected plays likely due to the fact that three of the four FRANTEXT plays are dramas featuring aristocratic characters (like the Duchess in *Pinto*), while many of the randomly selected plays are vaudevilles and melodramas featuring characters who are servants, farmers (like Anselme in *le Grenadier de Louis XV*) and artisans. This in turn suggests that the playwrights believed these lower-class characters would sound more realistic with a higher proportion of *ne … pas*.

Since we know that there were far larger numbers of farmer, servants and artisans in early nineteenth-century France than there were duchesses and princes, we can surmise that the speech in *le Grenadier de Louis XV* is closer to what we might have heard in the countryside near Paris in that time - or in the time of Louis XV. But that play is not in FRANTEXT because there was nothing in le Grenadier de Louis XV to draw the attention of literary historians. If our goal is testing linguistic theories, we often have to go beyond the principle of authority.

Building on this success, work is progressing on three fronts. First, the list of thirty plays and the full text of the first fifteen plays were made available to the public on GitHub. Second, a new sample of thirty-one plays was drawn at random from Volume 2 of Wicks' catalog (1953) and the process of obtaining, OCR and cleaning has begun for that sample. Third, additional linguistic variables are being selected and annotated in the corpus.

References

Bybee, J., and S. Thompson. (1997). Three frequency effects in syntax. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*.

Dubois, J.-B. (1815). *Le Grenadier de Louis XV.* Paris: Barba.

FRANTEXT. (2017). Retrieved from http://www.cnrtl.fr/corpus/frantext/

Imbs, P. (1971). *Trésor de la langue française*. Paris: CNRS.

Lemercier, N. (1800). *Pinto*. Paris: Huet.

Wicks, C. B. (1950, 1953). *The Parisian Stage*. Tuscaloosa: University of Alabama.