# Integrating controlled corpus data in the classroom

Caroline Rossi, Cécile Frérot, Achille Falaise



CAROLINE ROSSI, CÉCILE FRÉROT AND ACHILLE FALAISE

# Integrating controlled corpus data in the classroom: a case-study of English NPs for French students in specialised translation

## 1. Introduction

It is widely acknowledged that terms are highly frequent in scientific writing and noun phrases (henceforth NPs), in which a head noun may be modified by an adjective, another noun or a prepositional phrase, are known to be problematic in French-English translation due to their varying and contrasted complexity (Bouscaren et al 1992; Vinay and Darbelnet 2004; Huart and Larreya 2006). Indeed, French students hardly master English NPs in their translations –they tend to overuse the ‘the [Noun1] of [Noun2]’ construction as a loan translation (e.g. 'qualité de l’image' translated as 'quality of the image') where the ‘[Noun2][Noun1]’ construction (e.g. ‘image quality’) may be more appropriate. This remains a pitfall for more advanced translation students, notably in specialized (medical) translation. Indeed, medical English generally follows the principle of economy, so that the use of concise, complex NPs prevails (Maniez 2012). Yet in some contexts, the (the [noun] of [noun]) construction will be preferred, and there is no straightforward rule to help students decide which construction will yield an accurate translation. Based on how challenging English NPs are in French-English translation, we have carried out a corpus-based study in medical English texts, with a view to providing students with controlled corpus data that could be brought to bear on the decision-making process.

## 2. Methods

### *2.1.A constructionist approach*

While constructions were first conceptualised as referring only to those form-meaning pairs in which the construction accounted for non-compositional meaning, psycholinguistic evidence has shown constructions to be based mostly on frequency. As a result, Goldberg's definition of constructions was extended to the following:

> Any linguistic pattern is recognized as a construction as long as some aspectof its form or function is not strictly predictable from its component parts or from other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency(Goldberg 2006:5).

Even though it was seldom studied as such, there is a "noun phrase construction" in Goldberg's theory (see e.g. Goldberg 2003:221). We chose to characterise our complex noun phrases as constructions assuming that they might be learnt and stored as separate units of language, i.e. that they formed coherent categories within speakers' knowledge of language, since they shared a number of common features. In other words, the two patterns of English that are dealt with in this paper, as well as their French translation equivalents, are considered as: "learned pairings of form and function, [since these are characterized as] including words and idioms as well as phrasal linguistic patterns" (Goldberg and Suttle 2010:469). Crucially then, the approach implies that the items under study may not be learnt individually but that generalisations can be achieved for each construction. It also implies that one English construction will not be derived from another, "because different surface patterns are typically associated with differences in meaning or different discourse properties" (Goldberg and Suttle 2010:470). Consequently, this paper seeks to analyse and describe those differences in order to grasp at least some of the generalisations associated with each construction, and foster accurate language use in our students' translations.

One final reason for adopting a constructionist approach to language is that our study has been prompted by the impact of one cross-linguistic difference on French students' productions. As a matter of fact, constructionist approaches areparticularly relevant to cross-linguistic comparisons:

> [C]onstructions are viable descriptive and analytical tools for cross-linguistic comparisons that make it possible to capture both language-specific (idiosyncratic) properties as well as cross-linguistic generalizations." (Boas 2010: 15

By acknowledging the existence of two distinct constructions in English, and trying to capture their specific properties, a degree of syncretism in the French 'le [Noun1] de [Noun2]' construction can be grasped. In what follows, we have tried to capture as many language-specific properties as we could for each English construction.

### *2.2. Introducing controlled corpus data into the classroom*

Over the past fifteen years, the use of corpora has grown increasingly attractive in the translation classroom and a significant number of corpus scholars have advocated the integration of corpora in the curriculum of future translators (e.g. Zanettin 2002; Varantola 2003; Bernardini and Castagnoli 2008).

As teachers involved in the training of future professional translators, we are very much aware of the benefits corpora can have on students' translations: pedagogical uses of corpora have consistently proved relevant to students to help them produce natural-sounding translations based on idiomatic words and phrases. While a considerable number of studies has shown the potential of corpora to extract collocations, retrieve terminology and promote language awareness in student translators (e.g. Bowker 1999; Kübler 2003; Maïa 2003), others have stressed how complementary corpora were, when used with traditional resources, in that they provided accurate and relevant contextual information missing from dictionaries – whether they are monolingual or bilingual and general or

specialized dictionaries. (Pearson 1996; Zanettin 1998; Frankenberg-Garcia 2005; Frérot and Josselin-Leray 2007).

On the whole, corpora are reported to bring an added value to translations and this enhancement is mainly achieved by searching bilingual corpora, i.e. collections of either comparable or parallel texts with concordancers. Comparable corpora are commonly defined as "a collection of texts composed independently in the respective languages and put together on the basis of similarity of content, domain and communicative function" (Zanettin 1998:614) while "components in two or more languages, consisting of original texts and their translations" (Aston 1999:290) are referred to as "parallel". Undoubtedly, concordancers[1] play a major role in helping students navigate through comparable or parallel corpora; as a matter of fact, the vast majority of practice-oriented studies has focused on searching corpora in the classroom through concordancers, with corpora acting as "documentation tools" (Marco and van Lawick 2009).

In our study, we stand quite a different view in that we aim at designing coherent sets of controlled corpus data, or corpus-based 'clues',based on the assumption that helping students tackle a given linguistic translation issue –e.g. grasp the intricacies of English NPs– requires providing them with previously analysed, selected and structured linguistic material (in other words, 'controlled' data by teachers themselves) that goes beyond a list of concordances. In that respect, not only does our pedagogical perspective greatly favour using corpora for translation teaching, it is also in line with scholars such as Marco and van Lawick (2009) who regard corpora as a "source of materials for the translation classroom", thus prompting an emerging –or still under-explored– perspective in corpus-based applied translation studies.

### *2.3. Preliminary study*

[1] They may be freely available web-based tools or stand-alone products and may runwith raw or post-tagged texts, i.e. enriched with grammatical categories and lemmas.

We first conducted a qualitative, item-based study of a set of constructions.

A list of the most frequent head nouns (Noun1) in the recurring construction 'the [Noun1] of [Noun2]' was obtained from small a learner corpus in Nuclear Medicine of about 5,000 words, including 17 student essays. We isolated the first 12 elements, which are ranked by frequency in table 1 below.

| Frequency ranking | Raw frequency | Construction |
|---|---|---|
| 1 | 8 | The use of |
| 2 | 4 | The position of |
| 2 | 4 | The response of |
| 3 | 3 | The effect(s) of |
| 3 | 3 | The implementation of |
| 3 | 3 | The quality of |
| 3 | 3 | The risk of |
| 3 | 3 | The case of |
| 4 | 2 | The choice of |
| 4 | 2 | The development of |
| 4 | 2 | The investigation of |
| 5 | 1 | The study of |

*Table 1. Most frequent head nouns in the learner corpus in Nuclear Medicine*

Our premise was that the corresponding [Noun2] [Noun1] construction may be preferred in at least some of the occurrences (e.g. 'drug use' vs 'the use of drug(s)'). In order to verify our assumption, we searched a new, on-line French and English corpus of scientific texts, which includes over 40 million words: Scientext (Falaise et al 2012).

The Scientext project corpora were collected for the purpose of a linguistic study on reasoning and positioning in scientific writing, mainly focusing on phraseology and syntactic markers of causality. They consist of four distinct corpora (two in English, and two in French): biology and medical articles in English, scientific publications in French (including a biology and medical articles sub-corpus), essays written by French learners of English, and French

reviews of proposals for oral communications. These four corpora have been processed with a syntactic parser: Syntex (Bourigault 2007), providing part-of-speech tagging, lemmatisation, as well as syntactic dependency trees. They have also been manually partitioned into discursive sections (e.g. summary, introduction, development, and conclusion sections for the corpus of English articles). In order to search the corpora, an on-line environment has been designed: ScienQuest.

As a result of our preliminary investigation of Scientext, we were able to link observed frequency differences between a given 'the [Noun1] of [Noun2]' construction and its [Noun2] [Noun1] construction, with relatively stable right and/or left contexts. Table 2 summarises those results.

| Frequent left context | Frequent right context | Freq. [2] | Construction (all nouns are lemmas) |
|---|---|---|---|
| V + PP (according to, related to) | PP (with, for) | [67]93 | [The]use of (ADJ/N) drug |
| Adj (intravenous, psychotropic)/ N (injection) | [3] | 552 | Drug use |
| Lexical V (influence, affect) | PP (to) | [44]120 | [The]response of (ADJ/N) cell |
| Noun (t-, tumour; Nk, host) | | 384 | Cell response |
| Lexical verb (increase, reduce) | PP (in) | [154]513 | [The]risk of (ADJ/N) cancer |
| Adj/N (breast, colorectal) | | 808 | Cancer risk |
| V +PP (precede, involved in) | Sentence final | [143]190 | [The]development of (ADJ/N) cancer |
| Noun (breast, lung) | | 196 | Cancer development |
| Lexical V (examine, assess) | PP (on) | [154]346 | [The]effect of (ADJ/N) treatment |
| Adj (significant, large) | | 424 | Treatment effect |

*Table 2: Results of preliminary study (summary)*

2 Frequencies have not been normalised, they correspond to frequencies in the English, 35-million-word Scientext corpus.

3 Blank cells indicate that no frequent elements (i.e. covering over 20% of all occurrences) could be isolated.

Since frequencies were reasonably low, we started by analysing the concordance outputs, looking for regularities in left and right contexts. Frequent left contexts showed that 'the [Noun1] of [Noun2]' constructions were mostly used as verbal complements, while '[Noun2][Noun1]' constructions were very often modified by an adjective or noun. This was expressed as a first generalisation, labelled generalisation 1: while 'the [Noun1] of [Noun2]' construction is more likely to occur as a verbal complement, '[Noun2][Noun1]' constructions will often have pre-modifiers (which may result in the creation of a semantic subclass).

Frequent right contexts were less contrasted, and mainly showed additional PP expansions to be preferred after 'the [Noun1] of [Noun2]' constructions, hence generalisation 2: 'the [Noun1] of [Noun2]' construction is more frequently followed by an additional prepositional phrase (PP) than the '[Noun2][Noun1]' construction.

Looking at the respective positions of the two constructions in scientific texts, we noticed that the proportion of 'the [Noun1] of [Noun2]' constructions found in introductions was always significantly higher than that of '[Noun2][Noun1]' constructions, which were introduced later. This is in line with diachronic evidence showing gradual lexicalisation of combinations of two nouns (into compound nouns), which is not the case for 'the [Noun1] of [Noun2]' constructions: '[Noun2][Noun1]' constructions would then correspond to more opaque, technical terms that are less likely to be used in an introduction. Hence generalisation 3: while 'the [Noun1] of [Noun2]' construction will be preferred in introductions, '[Noun2][Noun1]' is likely to be preferred later in texts.

The presence or absence of a definite determiner was also quantified so as to assess the relative importance of definite constructions in our data. Indeed, in reference grammars such as Marcelin et al (2007:32), French students are usually taught that when a noun is followed by an 'of'-prepositional phrase, it will almost always be preceded by a definite determiner, with contrasted examples such as:

a. He teaches literature.
b. He teaches the literature of the Middle-Ages.

Although the presence of definite determiners is slightly above average in table 2, frequency counts also exhibit a degree of variation, from 30% of occurrences for ‘the risk of [ADJ/N] cancer’ to 75% for ‘the [ADJ/N] development of cancer’. Constructions in which definite determiners prevailed were therefore hypothesised to form a relatively homogeneous class. We decided to focus on this subclass only in future work, as both the French constructions under consideration and our students’ loan translations contained a definite determiner.

On the whole, even though a number of similarities could be found, the generalisations achieved were fully item-based and relatively limited in scope: we started from a limited number of NPs that had been found in a small, learner corpus More data was needed in order to confirm or modify the above, tentative generalisations.

### *2.4. Present study*

The aim of the present study is to establish a broader picture of the use of each construction in scientific English, with a view to helping our French students in specialized translation decide which English translation equivalent to choose.

In order to assess the relative frequency of each construction in scientific writing,we searchedthe English and French Scientext[4] corpora for occurrences that would match the two English constructions ‘the [Noun1] of [Noun2]’ versus ‘[Noun2][Noun1]’, and then the French construction ‘le [Noun1] de [Noun2]’.

| Construction | Frequency in Scientext (FR or ENGL) | Freq. per million words |
|---|---|---|
| the [Noun1] of [Noun2] | 143,432 | 4,070 |
| [Noun2][Noun1] | 2,290,929 | 65,010 |
| le [Noun1] de [Noun2] | 10,345 | 10,100 |

*Table 3. Relative frequency of each construction in Scientext*

4 In order to maximize comparability, we searched only the Medicine and Biology sections of the French corpus, which contain 1,024,235 words.

As can be seen from the table above, '[Noun2][Noun1]' constructions are almost sixteen times as frequent as 'the [Noun1] of [Noun2]' constructions. The fact that the total frequencies for both constructions should be almost seven times as frequent as the French construction might be at least partly linked with sampling issues: the French and English corpora in Scientext are not comparable corpora, both in terms of total number of words and text types –the French corpus containing research papers but also conference proceedings and PhD theses, with theses alone making up for more than four million words. Within the Medicine and Biology section of the French corpus, which was queried here, the total word count for theses alone is over 600,000. Another reason may be that adjective phrases should also be taken into account, especially as they are good candidates for French translation equivalents for one or the other English constructions (Maniez 2012).

A quick comparison of the frequencies in table 3 with those of both constructions in the British National Corpus (henceforth, BNC) confirms that these high frequencies are a specific feature of scientific English. According to an n-gram search within William H. Fletcher's *Phrases of English* page (http://phrasesinenglish.org/[5]), there are about 16,460 '[Noun2][Noun1]' constructions, and about 1,330 'the [Noun1] of [Noun2]' constructions per million words in the BNC, i.e. about a quarter of the proportions found in Scientext.

We then sought confirmation for the generalisations we had reached in our preliminary study. We started by isolating the most frequent nouns in each English construction, so as to make sure we were dealing with the most entrenched patterns, as well as to maximise the number of occurrences that could be used in our subsequent analyses.

| Construction | the [Noun1] of [Noun2] | Frequency per million words | [Noun2][Noun1] | Frequency per million words |
|---|---|---|---|---|
| Most frequent | The number of | 13.1 | Gene expression | 249.9 |

5 The database (accessed Oct 26, 2014) includes most but not all of the BNC: according to William H. Fletcher (personal communication) the exact number of tokens is 97,098,852.

| | | | | |
|---|---|---|---|---|
| nouns | gene(s) | | | |
| 2nd most freq. nouns | The number of patient(s) | 11.4 | Cell line | 242.1 |
| 3rd most frequent nouns | The proportion of patient(s) | 7.5 | Breast cancer | 223.3 |
| 4th most frequent nouns | The number of cell(s) | 7.1 | Health care | 136.1 |
| 5th most frequent nouns | The majority of patient(s) | 6.6 | Risk factor | 127.4 |

*Table 4.Most frequent nouns in each English construction in Scientext*

As is clear from table 4, looking at the most frequent constructions makes differences in proportioneven more salient. For the first most frequent noun combinations only, the proportion of the '[Noun2][Noun1]' construction is nineteen times as high as that of the most frequent'the [Noun1] of [Noun2]' construction.

For each of the tenpairs of constructions presented in table 5 below, we first tried to verify our assumptions by searching Scientext for the recurrent elements we had isolated in the left and right contexts of previously analysed constructions.

| the [Noun1] of [Noun2] | Frequency per million words | [Noun2][Noun1] | Frequency per million words |
|---|---|---|---|
| **The number(s) of gene(s)** | **13.1** | **Gene number(s)** | **1.6** |
| **The number(s) of patient(s)** | **11.7** | **Patient number(s)** | **1.1** |
| The proportion of patient(s) | 7.5 | Patient proportion(s) | 0 |
| **The number(s) of cell(s)** | **7.1** | **Cell number(s)** | **20.1** |
| The majority of patient(s) | 6.7 | Patient majority | 0 |
| **The expression(s) of gene(s)** | **3.4** | **Gene expression** | **249.9** |
| The line(s) of (ADJ/N) cell | 0.1 | Cell line | 242.1 |
| The cancer(s) of (ADJ/N) breast | 0.03 | Breast cancer | 223.3 |
| The care(s) of (the + ADJ/N) health | 0 | Health care | 136.1 |
| The factor(s) of (the + ADJ/N) risk | 0 | Risk factor | 127.4 |

*Table 5. Ten pairs of constructions from Scientext*

As a result of the above frequency counts, only four pairs (in bold) were kept for further comparisons. The paucity or absence of occurrences for one member in the other six pairs reveals that some constructions have become so entrenched as to block the use of an alternative construction. This is particularly clear for fully lexicalised, compound nouns like *health care* (sometimes spelt as one word) or *risk factor.* As for 'the [Noun1] of [Noun2]' constructions, partitive of-constructions indicating a part-whole relationship can hardly be replaced by an'[Noun2][Noun1]' construction. The first two constructions in which 'number' is used as a head noun form a distinct subset: the semantics of 'number' allows for ambiguity and at least part of the '[Noun2][Noun1]' constructions do not have partitive meaning, as exemplified in the concordances below where 'number' serves to designate an element in a series.

(1) We can see that there is an abrupt change of the smoothed local FDR around gene number 500 which corresponds to a threshold t = 0.15 for the p – value.

(2) The original names for known snRNAs were preserved, following the convention atUx.y, where x indicates the U snRNA type and y the gene number.

(3) Opaque, closed envelopes containing information on the allocated treatment for each patient number were prepared for medical emergencies.

(4) However, patient number five relapsed six months after the end of IFN therapy.

(5) Cardiac myocytes express relatively high levels of M6P / IGF2R and transgenic mice containing a homologous deletion of the M6P / IGF2R gene manifest ventricular hyperplasia due to an increase in cell number, 9 , 10 , suggesting that the M6P / IGF2R normally acts to suppress cardiac myocyte cell growth.

It is worth noting, however, that no such ambiguity is observed in concordances with 'cell number(s)', which seems to have become a fully lexicalised term, almost three times as frequent as 'the number(s) of cell(s)'.In order to see whether our generalisations also obtained with these fully lexicalised terms, the most frequent five

‘[Noun2][Noun1]’ and ‘the [Noun1] of [Noun2]’ constructions –as listed in table 4– were also investigated separately.

The last step in our analyses consisted in using some of the optionsavailable in the Sketch Engine (Kilgarriff et al 2004): as shown by Delcour, Lefer and Maubille(2013), they are particularly helpful in grasping an accurate collocational profile for a given word or pattern. We queried the English TenTen web corpus –a 12-billion-word corpus available in the Sketch Engine– and started from a simple phrase search for each of our analysed constructions. The ‘Sort good dictionary examples’ option (henceforth GDEX) then enabled us to analyse left and right contexts in the 40 best examples for each construction (see Kilgariff et al 2008) and see whether the predictions made by generalisation 1 and 2 were borne out. For generalisation 3, however, only Scientext could be used.

## 3. Results

### *3.1. Pre-modification vs. verbal complementation*

Starting with ‘the [Noun1] of [Noun2]’ constructions, table 5 and 6 below show ScienQuest outputs for queries on frequent left contexts. We sought to verify generalisation 1, as expressed below.When the total number of hits represented less than 20% of the total number of occurrences for the construction, we looked for other, more frequent left contexts.

Generalisation 1: while ‘the [Noun1] of [Noun2]’ construction is more likely to occur as a verbal complement, ‘[Noun2][Noun1]’ constructions will often have pre-modifiers (which may result in the creation of a semantic subclass).

| Frequent left contexts | Frequent ‘the [Noun1] of [Noun2]’ construction | Occurrences (left context + construction) | Percentage of total number of occurrences for the construction |
|---|---|---|---|
| V | the number(s) of genes | 153 | 33,19% |

| | | | |
|---|---|---|---|
| Preposition | the number(s) of genes | 133 | 28,85% |
| V + preposition | the number(s) of genes | 31 | 6,72% |
| V | the number(s) of patient(s) | 89 | 21,55% |
| Preposition | the number(s) of patient(s) | 102 | 24,70% |
| V + preposition | the number(s) of patients | 22 | 5,33% |
| V | the proportion of patient(s) | 43 | 16,23% |
| Preposition | the proportion of patient(s) | 61 | 23,02% |
| V + preposition | the proportion of patient(s) | 14 | 5,28% |
| V | the number(s) of cell(s) | 60 | 23,90% |
| Preposition | the number(s) of cell(s) | 102 | 40,64% |
| V + preposition | the number(s) of cell(s) | 25 | 9,96% |
| V | The majority of patient(s) | 6 | 2,53% |
| Preposition | the majority of patient(s) | 81 | 34,18% |
| V + preposition | the majority of patient(s) | 20 | 8,44% |
| V | the expression(s) of gene(s) | 58 | 47,93% |
| Preposition | the expression(s) of gene(s) | 37 | 30,58% |
| V + preposition | the expression(s) of gene(s) | 5 | 4,13% |

*Table 6. Frequent left contexts for a selection ofsix frequent 'the [Noun1] of [Noun2]' constructions*

Table 5 shows that prepositions represent the most frequent left context for all of our five, most frequent 'the [Noun1] of [Noun2]' constructions. Only the last and much less frequent construction seems to follow the pattern detected in our previous study. It should be reminded, however, that we could only deal with the closest elements in the construction's left context. Because we wanted to analyse constructionsrather than single words, dependency relations in ScienQuest (where only heads are featured) did not enable us to test for verbal complementation as such, and a simple search for verbs followed by prepositional phrases in the construction's left context could not capture syntactic complexity either. Indeed, prepositional complements could occur at a distance from the verb, and they could also complement noun phrases or adjectives, as is the case in the twoconcordances below:

(6) The rapid and continuing rise inthenumberofpatients receiving warfarin has

meant that traditional hospital based clinics are increasingly unable to cope with the throughput of patients.

(7) After every 4 patients the number of patients allocated to splinting is equal tothenumberofpatientsallocated to surgery.

The regularities revealed by the Sketch Engine for each frequent construction are presented in table 6 below, where the most frequent left context is always listed first. For five constructions out of six, verbs are the most frequent left context, so that prediction 1 is indeed borne out.

| Frequent left contexts | Frequent 'the [Noun1] of [Noun2]' construction | Frequency of construction (per million words) in the English TenTen corpus |
|---|---|---|
| V, preposition, V + preposition | the number(s) of genes | 0.02 |
| V, preposition, sentence subject | the number(s) of patient(s) | 0.2 |
| V (+ that clause), sentence subject | the proportion of patient(s) | 0.04 |
| V, preposition, V + preposition | the number(s) of cell(s) | 0.1 |
| Sentence subject, preposition, V (+that clause) | The majority of patient(s) | 0.1 |
| V, V + preposition, preposition | the expression(s) of gene(s) | 0.04 |

*Table 7. Frequent left contexts for a selection of six frequent 'the [Noun1] of [Noun2]' constructions, according to the SketchEngine's GDEX tool*

The likelihood that a construction will occur as sentence subject rather than verbal complement is highin the only two constructions for which no matching '[Noun2] [Noun1]' construction could be found in our data, namely 'the majority of patient(s)' and 'the proportion of patient(s)'. Although most constructions do occur as sentence subjects, this remains marginal (15% on average) except for those two constructions, in which the tendency accounts for 40 % and 30% of the Sketch Engine's good examples, respectively. Taken together, these two elements suggest that they might be part of a distinct subclass of constructions.

As for frequent '[Noun2] [Noun1]' constructions, pre-modification represents on average 42% of all occurrences (with a

relatively high standard deviation at 15.4). The details are given in table 7 below.

| Frequent left contexts | Frequent '[Noun2] [Noun1]' construction | Occurrences (left context + construction) | Percentage of total number of occurrences for the construction |
|---|---|---|---|
| Noun | Gene number | 4 | 7.1% |
| Adjective | Gene number | 13 | 23.2% |
| Noun | Patient number | 1 | 2.6% |
| Adjective | Patient number | 14 | 35.9% |
| Noun | Cell number | 99 | 14% |
| Adjective | Cell number | 203 | 28.7% |
| Noun | Gene expression | 979 | 10.8% |
| Adjective | Gene expression | 2125 | 23.5% |
| Noun | Cell line | 3526 | 41% |
| Adjective | Cell line | 2677 | 31.1% |
| Noun | Breast cancer | 290 | 3.7% |
| Adjective | Breast cancer | 2118 | 26.9% |
| Noun | Health care | 285 | 5.9% |
| Adjective | Health care | 1123 | 23.4% |
| Noun | Risk factor | 317 | 7% |
| Adjective | Risk factor | 2309 | 50.6% |

*Table 8. Frequent left contexts for a selection of eight frequent '[Noun2][Noun1]'constructions*

Noun modifiers are more frequent than adjectives in only one case: 'cell line', with two noun modifiers capturing half of all occurrences ('cancer cell line' and 'tumour cell line').This points to the existence of frequentcombinations of three nounsin our data. In order to analyse these uses we searched Scientext for those combinations. Table 8 below shows the most frequent twenty such combinations. While the statistics in ScienQuest had issued us with a list of twenty-five items, we left out acronyms, as well as combinations with "percent" –as in "percent confidence interval", which was the most frequent combination and occurred over a thousand times–on account that the combination of a figure with the noun "percent" may act more like a quantifier.

| Frequent '[Noun3][Noun2] [Noun1]' construction | Occurrences in Scientext | Frequency per million words |
|---|---|---|
| **Cancer cell line** | 961 | 40.5 |
| Amino-acid sequence | 928 | 27.3 |
| Polymerase chain reaction | 908 | 26.3 |
| Tumour necrosis factor | 889 | 25.8 |
| **Breast cancer cell** | 845 | 25.2 |
| Body mass index | 734 | 24.0 |
| Protein protein interaction | 638 | 18.1 |
| **Gene expression profile** | 600 | 17.0 |
| Case control study | 584 | 16.6 |
| **Breast cancer risk** | 583 | 16.5 |
| **Gene expression datum** | 536 | 15.2 |
| **Breast cancer patient** | 497 | 14.1 |
| **Health care system** | 484 | 13.7 |
| World Health Organisation | 484 | 13.7 |
| **Gene expression pattern** | 472 | 13.4 |
| **Health care provider** | 428 | 14.1 |
| **Gene expression level** | 415 | 13.7 |
| Tumour suppressor gene | 404 | 12.1 |
| Amino-acid residue | 395 | 11.8 |
| Signal transduction pathway | 366 | 11.5 |

*Table 9. Most frequent '[Noun3][Noun2][Noun1]' combinations in Scientext*

The ten combinations including previously analysed constructions– whether in our preliminary study or in the present study– appear in bold in the above list. Strikingly enough, the only construction in the list to occur with a pre-modifier is 'cell line', all other constructions being used as modifiers with a distinct head noun: we are planning to deal with those nouns and their frequent collocates in a new series of analyses. Besides, while the construction 'cancer risk' had been analysed in our preliminary study, the data in table 8 point to 'cancer cell' as a good candidate for further analyses.

### *3.2. Additional prepositional complements*

Generalisation 2: 'the [Noun1] of [Noun2]' construction is more frequently followed by an additional prepositional phrase (PP) than the '[Noun2][Noun1]' construction.

As shown in table 9, testing for generalisation 2 was less convincing, with only slightly higher percentages for prepositions following 'the [Noun1] of [Noun2]' constructions. Further queries also showed verbs to constitute a more frequent right context than prepositions, thus suggesting that occurrences as sentence subjects could be more frequent than previously assumed.

| Frequent 'the [Noun1] of [Noun2]' or '[Noun2] [Noun1]' construction | Frequent right contexts | Occurrences (construction +right context) | Percentage of total number of occurrences for the construction |
|---|---|---|---|
| the number(s) of genes | preposition | 140 | 30.4% |
| the number(s) of genes | V | 157 | 34.1% |
| Gene number | preposition | 16 | 28.6% |
| the number(s) of patient(s) | preposition | 113 | 27.4% |
| the number(s) of patient(s) | V | 197 | 47.7% |
| Patient number(s) | preposition | 11 | 28.2% |
| the proportion of patient(s) | Preposition | 98 | 37% |
| the proportion of patient(s) | V | 110 | 41.5% |
| the number(s) of cell(s) | Preposition | 76 | 30.3% |
| the number(s) of cell(s) | V | 91 | 36.2% |
| Cell number(s) | preposition | 195 | 27.6% |
| The majority of patient(s) | Preposition | 50 | 21.1% |
| the majority of patient(s) | V | 102 | 43% |
| the expression(s) of gene(s) | preposition | 33 | 27.3% |
| the expression(s) of gene(s) | V | 56 | 46.3% |
| Gene expression | preposition | 1940 | 21.5% |

*Table 10.Frequent right contexts for both 'the [Noun1] of [Noun2]' and '[Noun2] [Noun1]' constructions in our selection*

Searching the whole of Scientext for the patterns 'the [Noun1] of [Noun2] preposition' versus '[Noun2] [Noun1] preposition', we obtained more homogeneous results. Table 10 displays those results and gives evidence for the occurrence of prepositional phrases after both constructions, with only a slight preference for 'the [Noun1] of [Noun2]' constructions.

| Construction | Frequency in Scientext | Percentage of prepositions in right context |
|---|---|---|
| the [Noun1] of [Noun2] | 143,432 | - |

| | | |
|---|---|---|
| the [Noun1] of [Noun2] preposition | 33,102 | 23% |
| [Noun2][Noun1] | 2,290,929 | - |
| [Noun2][Noun1] preposition | 387,603 | 16.9% |

*Table 11.Relative frequency of prepositions in right contexts for both 'the [Noun1] of [Noun2]' and '[Noun2] [Noun1]' constructions in Scientext*

In order to assess the importance of verbs following 'the [Noun1] of [Noun2]' constructions, we used the Sketch Engine's GDEX option. The results in table 12 below, like those in table 7, are listed according to frequency ranking, so that the most frequent right context appears first. When frequencies are equivalent, items are separated by slashes.

| Frequent 'the [Noun1] of [Noun2]' or '[Noun2] [Noun1]' construction | Frequent right context |
|---|---|
| the number(s) of genes | Adjective phrase, V, preposition |
| Gene number | V/preposition/sentence final |
| the number(s) of patient(s) | Adjective phrase, preposition |
| Patient number(s) | V, link word, preposition |
| the proportion of patient(s) | Adjective phrase, preposition |
| the number(s) of cell(s) | Preposition, adjective phrase |
| Cell number(s) | V, preposition, sentence final |
| The majority of patient(s) | V, preposition/adjective phrase/sentence final |
| the expression(s) of gene(s) | Adjective phrase, preposition |
| Gene expression | N/sentence final, V/preposition |
| Cell line | N/preposition, adjective phrase |
| Breast cancer | N, sentence final |
| Health care | N, preposition/link word |
| Risk factor | Preposition, V/sentence final |

*Table 12. Frequent right contexts for our selection of 'the [Noun1] of [Noun2]' or '[Noun2] [Noun1]' constructions, according to the Sketch Engine's GDEX tool*

Besides showing more important variation in the right contexts of the '[Noun2] [Noun1]' construction, the results evidence the importance of post-modification in 'the [Noun1] of [Noun2]' constructions' right context, be it by prepositional phrases or adjective phrases. Generalisation 2 could then be amended as follows: 'the [Noun1] of [Noun2]' construction is more frequently followed by an additional post-modifier, in the form of a prepositional phrase (PP) or adjective phrase, than the '[Noun2][Noun1]' construction.

### *3.3 Position in texts*

To the best of our knowledge, ScienQuest is one of the only free online concordancers to compute the relative frequencies of a given item according to their position in texts. Indeed, a fair amount of manual annotation was necessary for the functionality to be fully operational: that has been done on the French corpus, but work on the English data is still under way. The English corpus has already been divided into: Abstract, Introduction, Text body andConclusion. Titles were particularly difficult to isolate and are one element that the team is still working on. We suspect that if '[Noun2][Noun1]' constructions do indeed correspond to specialised terms, they might be more likely to occur in titles, but this hypothesis still awaits verification. Using the present version of ScienQuest, we could already try to verify generalisation 3 with our new data.

Generalisation 3: while 'the [Noun1] of [Noun2]' construction will be preferred in introductions, '[Noun2][Noun1]' is likely to be preferred later in texts.

Table 10 displays the proportions (i.e. normalised[6]percentages of the total number of occurrences) found in introductions vs. text body for each frequent pair of constructions.

| Frequent 'the [Noun1] of [Noun2]' or '[Noun2] [Noun1]' construction | Occurrences in introductions | Occurrences in text body |
|---|---|---|
| the number(s) of genes | 4% | 57% |
| Gene number | 0% | 52% |
| the number(s) of patient(s) | 22% | 32% |
| Patient number(s) | 15% | 18% |
| the number(s) of cell(s) | 16% | 44% |
| Cell number(s) | 16% | 19% |
| the expression(s) of gene(s) | 14% | 17% |
| Gene expression | 23% | 17% |

*Table 13.Position in texts for both 'the [Noun1] of [Noun2]' and '[Noun2] [Noun1]' constructions in our selection*

6 ScienQuest computes normalised frequencies for occurrences in text parts by dividing raw frequencies by total number of words in each part (i.e. frequency of occurrences in introduction / total number of words in introductions, etc.).

Occurrences in introductions are systematically–but probably not significantly– lower for '[Noun2] [Noun1]' constructions, except for highly frequent terms like "gene expression". A less lexicalised term like 'gene number' does not occur at all in introductions. The occurrences in text body are relatively difficult to use: indeed, it is impossible to tell whether occurrences are located e.g. towards the beginning or end of a paper. ScienQuest should soon include frequent article sections, so that position in text body will be easier to track.

Because variation from one item to another made it difficult to decide whether generalisation 3 was borne out, we compared one construction with the other in the whole corpus. The results appear in table 11 below: '[Noun2] [Noun1]' constructions are slightly under-represented in introductions, while the lowest normalised frequency for 'the [Noun1] of [Noun2]' constructions is text body.

| Construction | Frequency in Scientext | Normalised frequencies |
|---|---|---|
| the [Noun1] of [Noun2] | 143,432 | - |
| the [Noun1] of [Noun2] in introduction | 5,877 | 0.005 |
| the [Noun1] of [Noun2] in text body | 126,255 | 0.003 |
| the [Noun1] of [Noun2] in conclusion | 2,641 | 0.005 |
| [Noun2][Noun1] | 2,290,929 | - |
| [Noun2][Noun1] in introduction | 64,283 | 0.059 |
| [Noun2][Noun1] in text body | 2,055,389 | 0.065 |
| [Noun2][Noun1] in conclusion | 28,943 | 0.060 |

*Table 14. Relative frequency of each construction in Scientext, according to position in texts*

On the whole, our results conform to generalisation 3, suggesting that the use of one or the other construction is constrained by discourse factors. Ongoing improvements of the statistics produced by ScienQuest should enable further testing of the hypothesis in a near future.

As a result of our preliminary study, we had startedcreating entries into an online, corpus-cum-dictionary, tailor-made to fit our students' needs. Having verified our working hypotheses, we could start creating new entries for the series of frequent constructions

analysed here. Because of the nature of the tool, access is item-based, butit has been designed for working on the specific contrast between French and English under scrutiny here, with a view to helping students grasp generalisations. Therefore, it is hoped that the more elements students are provided with, and the more frequent and representative these elements, the better their choices are likely to be.

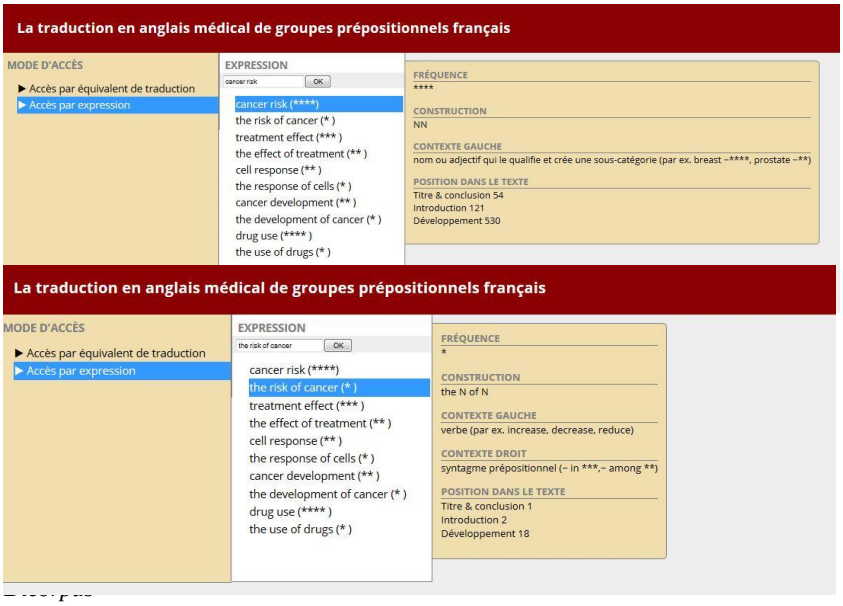
## 4. Integration into a new online tool: Dicorpus

Our classroom-oriented study raises the issue of how corpus data should be integrated in translation classes and questions the search of corpus data by students –a key issue from a pedagogical perspective. The present study aims at providing students with controlled learning material: in particular selected concordances. To this end, we took part in an ongoing experiment which consists in integrating our corpus data in a classroom-friendly version of Scientext, designed for non-native speakers of both French and English (Tutin and Falaise 2013; Hartwell and Jacques 2012).

> ScienQuest is a feature-rich environment designed for linguists to freely search corpora. Using this kind of environment requires linguistic skills, e.g. to discard tagging errors or statistically non-significant results. It also features numerous functionalities which learners do not need. The Dicorpus interface is a lightweight corpus query interface, built upon ScienQuest, which focuses only on learners' needs. With Dicorpus, learners may search the corpus through predefined requests, and consult clean results, previously filtered and validated thanks to numerous analyses, as shown below, and therefore guaranteed to contain only occurrences which would constitute good dictionary examples. The predefined requests are listed in two ways:Grouped under French ‘translation equivalents’ (as displayed below), each leading to two English constructions or phrases.Each English construction can also be accessed directly.

*Figure 1. The Dicorpus interface, displaying a selection of occurrences for "cancer risk"*

Concordances have been selected to match the most frequent left and right contexts revealed by our analyses. Besides, within each entry students can access information about frequent right and left contexts –where appropriate– as well as preferred text position, as illustrated in figure 2 below.




Future work includes continually enriching our entries according to frequent elements in the Scientext corpus, as well as the difficulties encountered by student with a given, French source text. Indeed, our goal is not only to help students on the translation of a given item, but also and more importantly maybe, to have them grasp some of the features of each construction, as captured e.g. by the generalisations tested in this paper. Whether students can and need to reach this level of abstraction is a moot point, but those generalisations were

necessary for the structure of each entry to be clear enough and for contrasts to emerge, as shown e.g. in the two entries in figure 2.

## 5. Conclusion

The present study has enabled us to gain more insight into the contrasted uses of two English constructions, whose respective functions are expressed by one and the same –presumably syncretic– construction in French. This has been achieved by relying on corpus-based evidence, which appeared to be all the more clear as the constructions analysed were frequent. Looking for emergent generalisations in rich corpus data is presented as a key step in designing entries for an online, corpus-cum-dictionary for our students.

Our experiment exemplifies one way in which controlled corpus data can be brought to bear on advanced translation students' understanding of the fine-grained differences between two constructions in the English language. In our view, this enhanced understanding will hardly be achieved when students are left to navigate corpora and sort out corpus data by themselves to solve a given translation problem. The hypothesis is currently being tested in the classroom, and it is hoped that our experiment will bring evidence in support of this claim.

## Bibliography

Aston, G. 1999. Corpus use and learning to translate. *Textus* 12, 289-314.

Bernardini, S. and Castagnoli, S. 2008. Designing a Corpus-based Translation Course for Translation Teaching and Translator Training. *International Journal of Translation*, 21/1-2, 133-147.

Boas, Hans C. 2010. *Contrastive Studies in Construction Grammar*. Amsterdam and Philadelphia: John Benjamins.

Bouscaren, J., Chuquet, J., Danon- Boileau and L., Flintham, R. 1992. *Introduction to a linguistic grammar of English: an utterer-centered approach*, Paris: Ophrys.

Bourigault, D. 2007. Un analyseur syntaxique opérationnel : SYNTEX. Mémoire de HDR. Toulouse, France.

Delcour, M., Lefer, M-A. and Maubille, G. 2013. Lexique et phraséologie dans les rapports de stage en traduction : étude de corpus. *Le Langage et l'Homme*, 48/2, 45-67.

Bowker, L. 1999. Exploiting the potential of corpora for raising language awareness in student translators. *Language Awareness* 8, 160-73.

Falaise, A., Tutin, A. Kraif, O. and Rouquet, D. 2012. ScienQuest: a Treebank Exploitation Tool for Non NLP-specialists. In 24th International Conference on Computational Linguistics (COLING 2012), proceedings. Mumbai, India, 131-140.

Frankenberg-Garcia, A. 2005. A peek into what today's language learners as researchers actually do.*International Journal of Lexicography*, 18/3, 335-355.

Frérot, C. and Josselin-Leray, A. 2007. Enriching Dictionaries with Corpus-Based Data. Towards an Improved Description of Verbs in General Bilingual Dictionaries thanks to a Popular-Science Corpus. In: proceedings of the Corpus Linguistics conference, Birmingham (United-Kingdom), 27-30 July 2007.

Goldberg, A.E. 2006. *Constructions at Work: the nature of generalization in Language*. Oxford: Oxford University Press.

Goldberg, A.E. 2003. Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7/5, 219–224.

Goldberg, A.E. and Suttle, L. 2010. Construction grammar.*WIREs Cognitive Science*, 1: 468–477. doi: 10.1002/wcs.22.

Hartwell, L.M. and Jacques, M.P. 2012. A Corpus-Informed Text Reconstruction Resource for Learning about the Language of Scientific Abstracts. In L. Bradley & S. Thouësny (eds) CALL: Using, Learning, Knowing, EUROCALL Conference, 22-25 August 2012, proceedings, 117-123. Gothenburg, Sweden.

Huart, R. and Larreya, P. 2006. *Les constructions Nom+Nom*, collection « gramvoc », Ophrys.

Kilgarriff A., Rychly, P. Smrz, P. and Tugwell, D. 2004. The Sketch Engine. In Proceedings of Euralex 2004, Lorient (France), 105-116. http://www.sketchengine.co.uk/

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychlý, P. 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus in

Bernal, E. and DeCesaris, J. (eds) Proceedings of the XIII EURALEX International Congress. Barcelona: Universitat Pompeu Fabra: 425-433.

Kübler, N. 2003. Corpora and LSP translation. InF. Zanettin, S. Bernardini and D. Stewart (eds).*Corpora in Translator Education*,) , Manchester: St Jerome, 25-42.

Maia, B. 2003. Some languages are more equal than others. Training translators in terminology and information retrieval using comparable and parallel corpora. InF. Zanettin, S. Bernardini and D. Stewart (eds).*Corpora in Translator Education*, Manchester: St Jerome, 43-53.

Maniez, F. 2012. A corpus-based study of adjectival vs. nominal modification in medical English. In Boulton Alex, Shirley Carter-Thomas & Elizabeth Rowley-Jolivet (eds.), *Corpus-Informed Research and Learning in ESP: Issues and Applications, Studies in Corpus Linguistics*. Amsterdam and Philadelphia: John Benjamins.

Marcelin, J., Faivre, F., Garner, C. and Ratié, M. 2007. *Anglais : Grammaire*. Paris: Nathan.

Marco, J. and van Lawick, H. 2009. Using corpora and retrieval software as a source of materials for the translation classroom. In *Corpus Use and Translating*, ed. by Allison Beeby, Patricia Rodríguez Inés and Pilar Sánchez-Gijón. Amsterdam and Philadelphia: John Benjamins, 9-28.

Pearson, J. 1996. Electronic texts and concordances in the translation classroom.*Teanga*16, 86-96.

Tutin, A. and Falaise, A. 2013. Multiword expressions in scientific discourse: a corpus-driven database. In eLex 2013, proceedings. Tallinn, Estonia.

Varantola, K. 2003. Translators and disposable corpora. InF. Zanettin, S. Bernardini & D. Stewart (eds)*Corpora in Translator Education*,Manchester: St Jerome, 55–70.

Vinay, J.-P. and Darbelnet, J. 2004. *Stylistique comparée du français et de l'anglais*. Paris: Didier.

Zanettin, F. 1998. Bilingual Comparable Corpora and the Training of Translators. *Meta.* 43/4, 613-630.

Zanettin, F. 2002. Corpora in Translation Practice. In Proceedings of the LREC Workshop, Language Resources for Translation Work and Research, 10-14.