3.3. Second language acquisition

The field of second language acquisition is one of those in which the use of corpora has grown exponentially in recent decades. The creation of numerous learner corpora, as well as the development of new methods and annotation tools, has largely contributed to this evolution. While linguists working on the question of second language acquisition have long used learners' productions as a source to build their theories, these data were limited to very small samples or even to single-person studies. Therefore, the generalization potential linked to these data was highly questionable. This led to the creation of real learner corpora, aiming to provide representative samples of this population.

There are different types of corpora containing language produced by learners. The first learner corpora produced at the end of the 1980s were limited to written productions, and these corpora are still the most numerous today (Gilquin 2015), but spoken corpora have also developed recently. Learner corpora often include language produced when performing various kinds of tasks such as essay writing, conversations and descriptions of images. As a result, the status of these productions as samples of natural linguistic productions can be called into question. In fact, classroom writing is an artificial genre, not entirely representative of the way in which students write freely. Similarly, image description is justified for educational purposes but does not correspond to real situations involving spontaneous speech. However, Granger (2008) argues that learner corpora represent "quasi-natural" language, because the situations they represent are typical (albeit not of everyday life situations) of linguistic activities carried out in a language learning situation.

As is the case with other types of corpora, learner corpora have mainly been compiled in English, but there are also resources for other frequently taught languages such as French (see Chapter 5, section 5.5). Some corpora contain language samples produced by learners of different mother tongues, and this information can be found in the metadata (see Chapter 6, section 6.4). These corpora are valuable tools for measuring the role of different mother tongues in the process of acquiring the same foreign language, as we will see below.

Finally, most learner corpora are cross-sectional corpora, including one sample per participant and representing a given moment during the acquisition process, since most of the time learners included in a corpus have a homogeneous level of competence in the foreign language. These corpora are very useful for determining learner competence at a certain level, but considered individually, they do not make it possible to study different acquisition stages. For this, longitudinal corpora are necessary, but these are rare due to the difficulty of sampling the same learners across several years. One way to get around this problem is to compare several cross-sectional corpora including learners from different levels of competence.

The first study we will discuss focused on whether learners at a very advanced proficiency level keep on improving, which would justify the need to define different development stages for learners beyond the so-called advanced stage of acquisition. To do this, in the spoken corpus *InterFra*, Forsberg Lundell *et al*. (2014) defined three groups of French non-native speakers whose mother tongue was Swedish. Each group included 10 speakers. The first group was made up of speakers aged 19 to 34 years who had lived for one to two years in France. The second group included speakers aged 25 to 30 who had lived between 5 and 15 years in France, and the third group included speakers aged from 45 to 60 years who had lived between 15 and 30 years in France. These groups were compared to two groups of 10 native speakers each, who were between 15 and 30 years old and 45 and 60 years old respectively, chosen to match the ages of learners.

The groups of learners were then compared on the basis of five linguistic indicators:

– the number of non-native morphosyntactic forms produced, for example, gender or plural agreement mistakes;

– the number of left dislocations, in sentences such as "*moi*, *si tu me demandes*, *il a tort*" (literally: me, if you ask, he is wrong), which are typical of spoken French;

– the number of formulaic sequences such as collocations;

– lexical richness, calculated based on the number of words with high, middle and low frequency in corpus data that are used;

– fluency, measured by articulation speed and utterance length between two pauses.

The results indicated that the group of learners with 5–15 years of residence differed from the 1–2 years of residence group on the following criteria: use of formulaic sequences, lexical richness and fluency. The group having lived longer than 15 years in France did not differ from the 5–15 years of residence group on any of these indicators. However, speakers who had lived the longest in France managed to pass for natives in a listening discrimination test administered to French native speakers, unlike the speakers from the 5–15 years of residence group. This indicates that some form of progression must have taken place between these two groups, but which could not be measured through the tests chosen for this study. Furthermore, all the groups of learners differed from native speakers (but did not differ from each other) on the assessment of morphosyntax, which seems to indicate that this is an area which can remain beyond the reach of even the most advanced learners. This study thus showed that language continues to develop beyond the so-called advanced acquisition stage and that this progression is not uniform among the different dimensions of language. As the lexicon continues to progress, certain aspects of the language system such as morphosyntax remain at a non-native level, even at truly advanced acquisition stages.

The second study that we will introduce compared the use of two English discourse markers, *in fact* and *actually*, by learners of two different mother tongues and by native speakers. Buysse (2020) compared the oral productions of French-speaking and Dutch-speaking learners of English in the *LINDSEI* corpus with the productions of native English speakers in the *LOCNEC* corpus, a corpus which was compiled to be comparable with the *LINSDEI* (see Chapter 4 for a definition of the concept of comparability). The interest in comparing French and Dutch speakers is that there are different translation equivalents for the English markers in both languages. While Dutch has two markers which closely resemble those in English (*eigenlijk* for *actually* and *in feite* for *in fact*), French only has one close marker, which is *en fait*. So, in French, *actually* has no translation equivalent of its own.

In her study, the author first performed a frequency analysis regarding these two markers in the three sub-corpora. The results indicated that *actually* is significantly more common than *in fact* among Dutch native speakers in comparison to French speakers. Conversely, *in fact* is significantly more common than *actually* in the speech of French speakers compared to Dutch speakers. On the basis of the literature, she then

identified all the possible functions for these markers in English, such as introducing an elaboration or a contrast, and then annotated all the occurrences of these markers according to one of these functions. This analysis allowed her to show that learners use all the possible functions that the markers offer, even if their respective frequency varies a little between French speakers and the other two groups. That being said, the low number of occurrences of the marker *actually* among French speakers (56 in all) prevents a quantitative analysis of the differences between its different functions. In summary, this study demonstrated the influence of speakers' mother tongue on the use of discourse markers in a foreign language, and in particular, the importance of having a similar marker in L1 to help learners use markers appropriately in a foreign language. Indeed, Dutch speakers, who have two very similar markers in their mother tongue, use *in fact* and *actually* in the same way and in the same proportions as natives. On the other hand, French speakers tend to under-use the marker, which has no direct equivalent in their mother tongue (*actually*) and to overuse the other marker (*in fact*), to perform the same functions, as they would do in French. This study thus indicates that negative transfer effects occur even among advanced learners.

3.4. Language teaching

In addition to collecting learner corpora as we discussed earlier, the area of language teaching currently makes an extensive use of corpora produced by native speakers (see, for example, Sinclair (2004) and Cheng (2010) for literature reviews). Corpora including different genres are used for the preparation of teaching materials, in order to present learners with real-life communication examples. Corpora also help to set these examples in a much richer context than traditional dictionaries and grammar textbooks. Finally, using frequently updated corpora makes it possible to provide examples of usage that better match the reality of contemporary speakers than conventional tools, whose examples are aging rapidly and which often represent only a normative usage that is often disconnected from the reality of native speakers.

In the field of vocabulary in particular, the use of corpora makes it possible to empirically provide lists of the most frequent words in a certain field, which should therefore be taught as a priority. Another key point for mastering a foreign language is to know, apart from the meaning of isolated

words, certain elements of phraseology, in other words the typical linguistic sequences in which a word occurs, for example for the word "*knowledge*", "*to acquire knowledge*", "*knowledge gain*" or "*prior knowledge*". Some researchers even think that these elements should be taught as lexical units (Kennedy 2003). On this point, corpora have become very useful resources, because they make it possible to automatically retrieve the most common phraseological elements for a given word.

In addition, corpora provide examples of spoken language, which are clearly more realistic than the constructed dialogues contained in most language methods. Given that they include natural interactions, corpora include reformulations, hesitation markers, turn-taking devices, etc., which are not reproduced in artificial dialogues, but which are important for learners to master since they are an integral part of language uses among native speakers. Finally, the creation of learner corpora has also made it possible to bring a new dimension to language teaching, by allowing learners to consult non-native productions and to compare them with native productions. Access to such data enables learners to become aware of the differences between their productions and those of natives. In addition, these corpora often contain an annotation of errors, which favors an explicit learning process and enables learners to become conscious of typical errors and to avoid them.

An important question for language teaching is to determine to what extent the corpora developed for linguistic research can be reused as such in the classroom. On the one hand, there are many advantages to letting learners use corpora by themselves, for example, by teaching them to search for word occurrences using a concordancer. This practice induces active reflection on the language, when it comes to determining what to look for and how to look for it, which enhances learner autonomy and stimulates students to become involved in the learning process (Bernardini 2004). On the other hand, the use of corpora built for research purposes implies both the need to learn how to use corpus searching tools, as well as the ability to interpret the numerous concordance lines retrieved by such operation, and in particular, how to sort relevant occurrences from noise. In certain teaching contexts, these barriers prevent the use of corpora.

Furthermore, Braun (2005) argues that several features of corpora have been created for research purposes and renders them unsuitable for classroom use. While it is true that their large size is essential for answering many research questions in linguistics, this makes them both impractical and of little use for learners. For the latter users, in fact, a more limited number of well-chosen illustrations is better than hundreds of concordance lines. In addition, for corpora to be useful for learners, they should contain annotations of many linguistic phenomena such as syntactic structure and speech acts, in order to make them suitable for later search. In research corpora, these phenomena are still not frequently annotated at a large scale. Conversely, the sets of tags used by part-of-speech taggers are often too detailed to be understandable by learners (see Chapter 7). For all these reasons, smaller and more specific corpora are produced by teachers to better meet the needs of their class. Aston (2001) suggests that learners should start by using these small corpora specifically built for classroom needs (see Reppen 2010a) before moving on to larger, more general corpora when they reach a more advanced learning stage.

Limitations on the use of corpora created for research also apply to the use of raw corpus data for creating language methods. In order to base a language method on corpus data, it is imperative that the corpus chosen is adapted to the target audience, in particular from the point of view of the variety of the language represented, discourse genres, the age of the speakers, etc. According to McCarten (2010), frequency data drawn from corpora should not always be implemented as such in language methods. Indeed, certain infrequent words in corpora are still part of the basic vocabulary of a language. A case in point are the words for naming the days of the week. Conversely, some frequent words in corpora, such as prepositions, sometimes involve concepts or linguistic structures that are too complex to be included at a beginner level. The designers of a language method strike a balance between the frequency information provided by corpora, the perceived usefulness of the word and its learning difficulty. In the same way, trying to include raw spoken data in teaching materials can be troublesome. First, real conversations are often too long to be studied in their entirety, and cutting them poses consistency problems. Second, spontaneous conversations sometimes refer to topics that are uninteresting, inappropriate or simply difficult to understand without the conversational context. For all these reasons, spoken data should often be prepared (choice of theme, predetermined length, etc.) in order to avoid these problems.

In this section, we will introduce two studies which show the usefulness of corpora for language teaching. Each of them compared corpus data with the presentation of the same phenomenon using different language methods. Biber and Reppen (2002) looked at three aspects of English grammar teaching in six widely used language methods for different levels. They identified three elements in these methods:

– the grammar points discussed;

– the order in which they were presented;

– the vocabulary used in the examples to introduce such points.

Then, they compared the examples with frequency data drawn from a 20 million word corpus, corresponding to four different language registers.

In each of the three areas studied, significant differences were observed between corpus data and the presentation of the same phenomenon in language methods. For example, in the section introducing the forms that noun phrases may take in English, most of the methods only indicate a pre-nominal modifiers can be an adjective (*a nice man*), a present participle (*an exciting game*) or a past participle (*stolen goods*). However, in written corpora, nouns are also common modifiers of other nouns (e.g. *metal seat* and *tomato sauce*) and the relationships they express are diverse and complex. This syntactic pattern should also be included in language methods. In addition, the order of presentation for the different grammatical features does not correspond to the uses observed in corpora, especially in the case of verbal tenses. Most methods strongly emphasize progressive forms and represent them as the default form in conversations. However, corpus analysis shows that the most frequent case in many language registers is, on the contrary, the simple aspect. Finally, the authors observed that the verbs used for illustrating different grammatical properties in language methods are not necessarily the most common verbs in real-life language. Although introducing less frequent verbs may be useful for broadening learner vocabulary, it is nevertheless surprising that the most common words are not used, at least for beginner learners. This study showed that the intuitions of language method designers often do not reflect actual language uses. Corpus data make it possible to produce better-suited educational materials to match the realities encountered by learners.

Racine and Detey (2017) also compared the information given in language methods with corpus data, focusing on the question of *liaisons* in French. Producing liaisons is a particularly difficult aspect of spoken French for learners. In fact, producing liaisons correctly requires mastering production constraints (such as identifying the consonant involved in the liaison, taking into account possible modifications to the phonological environment, etc.), as well as the syntactic environment making the liaison either required, optional or forbidden. Most methods of French as a foreign language focus only on the compulsory, optional or forbidden nature of the liaison, which they present in a normative and simplified manner (Racine 2014). However, recent corpus studies have revealed that learners have many problems in producing liaisons, which are little addressed in language methods, or not addressed at all (Chevrot *et al*. 2013). Conversely, corpus studies with native speakers indicate that the contexts in which liaisons are produced may vary from speaker to speaker, different language registers and even ages (see Chapter 2, section 2.1), and contradict the normative data introduced in language methods. Indeed, when learner language is compared to that of native speakers in contemporary corpora, no difference in the production of compulsory liaisons can be observed between the groups, because the standard represented in language methods often does not correspond to the reality of the productions of native speakers. The study thus highlighted the importance of using native speaker corpora in order to compare their production with that of learners. These comparisons may help to better identify the elements to be taught.

3.5. Lexicography

Writing a dictionary requires the use of textual data in order to identify the words that should be included and to illustrate their contexts of use. Since the beginnings of dictionaries, lexicographers have manually collected examples from various sources, mainly literary ones. This focus on literary texts is particularly visible in the case of French, a language for which lexicographers have focused on a formal register of the language. For example, the *Trésor de la Langue Française* has drawn its 430,000 examples from "two centuries of French literary productions" (Pierrel *et al.* 2004). In other languages such as English, however, the use of other genres like journalistic texts is much more widespread.