Using conventional sequences in L2 French

FANNY FORSBERG

Abstract

By means of a phraseological identification method, this study provides a general description of the use of conventional sequences (CSs) in interviews at four different levels of spoken L2 French as well as in interviews with native speakers. Use of conventional sequences is studied with regard to overall quantity, category distribution and type frequencies. The most predictive measure is overall quantity, which yields significant differences between several learner levels. It is also found that Lexical CSs are the most difficult to acquire for second language speakers: only the most advanced group use them to the same extent as native speakers. No significant differences are found between the most advanced group of L2 speakers (LOR in France >5 yrs) and native speakers, probably due to the measures and the task investigated. The results are then related to Ellis et al. (2008), suggesting that the sequences’ frequencies of occurrence vs. their MI score in a larger corpus might influence their acquisition and use.

1. Conventional sequences in L2 acquisition and use

1.1. Conventional sequences in language use: Introduction

In 1983, Pawley & Syder launched the concepts of ‘nativelike selection’ and ‘nativelike fluency’, suggesting that native speakers’ productions were characterized to a large extent by combinations of words which were more than just grammatical – they were the preferred choices of native speakers. For in-

---

1. This article is based on a paper presentation at the 2007 AAAL conference in Costa Mesa, California April 21–24, 2007. It constitutes a modified and abridged version of Forsberg (2008). The author would like to thank Britt Erman, Inge Bartning, Emanuel Bylund and two anonymous reviewers for their valuable comments on earlier drafts of this article.
stance, native speakers of English prefer to say ‘I want to marry you’ instead of the awkward, although grammatically correct, ‘I wish to be wedded to you’.

The conventional character of language and the importance of non-generated language have been recognized over the last two decades, mainly due to the advent of corpus linguistics, which has allowed for the patterned character of language to come to light. Sinclair’s (1991) seminal work on the idiom principle and the open choice principle opened up for a whole new way of looking at language production, stressing the idiomatic – or formulaic – character of language, neglected by Chomsky’s generative grammar.

The importance of conventionalized language has also been discussed with regard to pragmatics and conversational interaction, since much of our everyday exchanges are predictable and quite conventionalized. The main idea is that socially routinized situations are also linguistically routinized (cf. Coulmas 1981, Aijmer 1986).

The underlying assumption has often been that these preferred patterns in native speakers’ productions are memorized and stored as formulaic chunks. This is why the field of investigation is frequently called ‘formulaic language’ and the specific sequences under study, ‘formulaic sequences’, but as the reader may notice, the linguistic sequences investigated in the present study are called conventional sequences. The question of terminology will be further discussed in Section 1.3.

1.2. Conventional sequences in L2 acquisition

In the field of second language acquisition, researchers have discussed formulaicity since the 1970’s, especially since the work of Wong-Fillmore (1976), who suggested that acquisition benefits from initial learning of formulaic chunks, as they become later on analyzed and contribute to creative rule development. In a similar vein, Myles et al.’s (1998, 1999) studied the French interlanguage of English school children. They found many instances of unanalyzed chunks such as la Monique j’aime le tennis (the Monique I love tennis), where j’aime was clearly not analyzed as a pronoun and a verb, but perceived as just a verb. Studies like Myles et al’s also show that there is not a total overlap between native speakers’ formulaic chunks and those used by second language learners, which is perfectly congruent with other aspects of interlanguage.

Review articles, such as Weinert (1995) and Wray (1999) have shown that formulaic language in second language acquisition has been studied from several angles. One of them is the approach used by Wong-Fillmore (1976) and Myles et al. (1998, 1999) where unanalyzed chunks of language are mainly seen as a basis for creative rule extraction in early learning. Studies that have investigated more advanced levels of L2 acquisition tend to consider formulaic
language from the point of view of nativelike selection, often as materialized in collocations. As early as in 1989, Yorio investigated the link between idiomaticity and proficiency. Several researchers have since then been interested in the use of collocations and other formulaic sequences in advanced learners, such as Granger (1998), DeCock (2004) and Nesselhauf (2005). Findings indicate that advanced learners’ use of formulaic sequences in L2 is often subject to L1 influence and that advanced learners’ use is characterized by overuse, underuse, misuse or stylistic inappropriateness.

Kecskes (2002) studied a specific kind of formulaic sequences – situation-bound utterances – with regard to L2 acquisition of pragmatics. He means that knowledge and use of these utterances in the L2 are highly dependent on individual differences and that length of residence in the target language community is not a good predictor for mastery of these sequences, but that they are more linked to cultural distance between the L1 and the L2 and willingness to accommodate to the cultural concepts expressed through the situation-bound utterances.

What do the research strands presented above have in common? One is looking at native speakers’ preferred word combinations (and in the case of SLA, how learners cope with these sequences) and the other one is interested in second language learners’ unanalyzed sequences, not always used according to TL norms. Wray (2002) regrouped these phenomena, suggesting that they are both the results of memorization and holistic processing and therefore proposed a working definition of what she calls ‘formulaic sequence’. At the core of Wray’s definition is the assumption that a formulaic sequence is or appears to be generated and stored as a whole, without recourse to grammar or lexicon rules (Wray 2002:9). This is why we find structures such as idioms, collocations, discourse devices, routine formulae and learner specific chunks under one and the same label – formulaic language – although we are dealing with quite different phenomena. What the two research strands seem to have in common is the assumption that they are dealing with sequences that cannot be analyzed using grammatical and lexical rules.

As mentioned in the review above, researchers have to date concentrated either on early learning or on advanced learning and the present study will thus be among the first in offering an overall picture of L2 use of conventional sequences. In Forsberg (2008), an attempt was made to accounting for both nativelike and learner-specific formulaic sequences at different levels of L2 advancement, having as a starting point Wray’s conception i.e. that they all stem from the same processing mechanism. However, as will be argued in the section below, the term ‘formulaic’ is problematic (it is difficult to operationalize and too vast), which is why the present study has a more limited scope than Forsberg (2008). It will not investigate learner internal phenomena (such as Myles et al 1998, 1999, but it will only concentrate on targetlike conventional
sequences and how they are used across L2 learner levels, from beginners to very advanced users of L2 French. This overall aim has, without no doubt, its disadvantages. For instance, quantitative aspects will be in focus, to the detriment of more qualitative aspects, but an overview of second language use is called for, providing a base for more focussed studies in the future.

1.3. Choice of terminology, methodology and theoretical framework

As mentioned earlier, Wray’s (2002) definition is rather unsatisfactory from an operational point of view, especially in a corpus-based study. As Schmitt et al. (2004) state, it is actually quite difficult to verify empirically if a sequence is stored and generated whole. Moreover, even if we make an attempt at testing this claim psycholinguistically, how should this be done? An example of methodologies used is self-paced reading tasks. Most of the time, it seems that ‘holistic storage’ is operationalized as ‘speed of reading’ i.e. processing benefits. Schmitt & Underwood (2004) found mixed evidence when they put corpus derived collocations to test: some had processing benefits, others did not. In another experiment, Conklin & Schmitt (2008) found that opaque idioms, such as ‘grab the bull by its horns’ did show processing benefits, compared to more creative language. Nevertheless, considering that most conventional sequences are not opaque idioms, what we can say with some certainty is that there are strong associative connections between the member of the conventional sequence. This means that a sequence such as ‘make a decision’, although it cannot presumably be generated on the basis of grammatical or lexical rules, this does not necessarily imply that it is stored as a holistic unit in the mental lexicon.

Bardovi-Harlig (subm) and Edmonds (2008) suggest that the term formulaic, especially in Wray’s (2002) definition, implies that the sequence in question is stored and retrieved holistically. However, as they point out, very few studies actually investigate the psycholinguistic reality of these sequences. Therefore, they propose a distinction between formulaic and conventional. Edmonds (2008) analyzed speed of reading for a number of frequent pragmatic expressions in French in her study. Only the sequences showing processing benefits can, according to Edmonds, be labelled as formulaic. Other expressions, as idiomatic and frequent they may appear, are merely called conventional expressions. This latter term is useful for the present study, since no psycholinguistic measures are used: it is a study on how conventional sequences - the preferred choices of combinations of words of a speech community- are used by L2 speakers and native speakers. Bardovi-Harlig (subm) and Edmonds (2008) talk about conventional expressions, but I would like to talk about conventional sequences, since not all of the sequences studied are ‘expressions’, they are often...
Using conventional sequences in L2 French

phrases, with open slots. The definition used in this study of a conventional sequence is the following:

A continuous or discontinuous sequence of at least two words where the combination of the words and/or its function is conventionalized to a particular extent in a speech community.

The concrete procedure of identification, which is not an easy task, will be further discussed in Section 3, but some methodological choices will be discussed below.

There are many ways of identifying conventional sequences. For the identification of conventional expressions, Edmonds (2008) used Discourse Completion Tests (DCT:s) in order to establish the most common expressions, used in a number of given situations by native speakers. This means that conventionality is purely based on participants’ preferences of use. However, since the present study uses recorded spoken production, this method has to be discarded. Furthermore, the method only allows for the study of a reduced number of sequences, while the present study aims at mapping overall use of conventional sequences.

The two most commonly used ways of identifying formulaic sequences in a corpus are the statistical method and the phraseological method (cf. Granger & Pacquod 2008). The statistical method is automatic whereas the phraseological method involves manual identification. The statistical method consists of detecting the most frequently recurring 2-, 3- or 4-words sequences in a corpus, sequences which DeCock et al. (1998) call recurrent sequences. Many scholars (Ellis et al. 2008, Siyanova & Schmitt 2008) are also using measures such as MI (Mutual Information score), which measures the strength of the associative links between word-pairings in the corpus at hand. Sequences obtaining an MI score above 3 are often labelled as collocations.

Although this statistical approach is very attractive from a scientific point of view, it has a number of disadvantages. Pazos Bretaña & Pamies Bertrán (2008) point out that many sequences that are automatically extracted are not interesting from the point of view of phraseology. This means that manual analysis is still required in separating what is actually a sequence of interest or not. Moreover, statistical methods are typically used on large corpora such as the BNC. The present study uses a small-scale specialized corpus, which is nowhere the size of the BNC. The study comprises 36 interviews carried out on different learner levels and native speakers (totaling 72,185 words).

2. The term conventional sequences are used for the sequences in the present study, but terms such as formulaic language, formulaic sequences, prefabs and collocations will be used to talk about the results of other studies, where these terms are used.
3. British National Corpus
This size is not sufficient for applying automatized corpus methods, such as automatic extraction and MI scores. In addition, automatic extraction falls short of taking into account semantic and pragmatic aspects, which can be very important when identifying conventional sequences, since it is not only a question of conventional form, but also of conventional function.

In view of the small size of our corpus and our goal being to describe overall use of conventional sequences, it was decided that a phraseological method would be more appropriate for this study. A phraseological method identifies sequences based on their syntactic, semantic and pragmatic restrictions. Other studies that have used similar methods are Howarth (1998), Nesselhauf (2005) and Bolly (2008). They are all interested in how a sequence’s degree of restriction affects L2 acquisition and use.

Erman & Warren (2000) used the criterion of restricted exchangeability in their much-cited study. Inspired by the work of Sinclair (1991) they wanted to investigate the proportions of the application of the idiom principle and open choice principle, in spoken data and in written data of native speakers of English. They found that between 50–60% of the productions consisted of ‘prefabs’, what is here called conventional sequences. Even if these figures are not exact, they convey the message that conventional sequences are pervasive in language use and it is consequently of great interest to know find out how they are used by second language speakers. Their study thus constitutes the principal source of inspiration for the present study and their categorization and methodology are used with some modifications. The present model of analysis was developed in Erman, Forsberg & Fant (2008).

The results of the empirical study will be interpreted within usage-based theory of language acquisition (Bybee 2008, N. Ellis 2008), which will be described more thoroughly in Section 1.4. Usage-based theory puts a lot of emphasis on the use of formulaic sequences and holds that they are acquired through experience with language and that frequency in the input contributes to the entrenchment of the sequences. However, Ellis (2006) shows that frequency alone cannot explain L2 acquisition and that other factors will also have to be taken into account.

1.4. Conventional sequences and Usage-based perspectives on SLA

During the last decade, usage-based perspectives on language acquisition have become more and more influential. Drawing on research from cognitive psychology, such as connectionist models, Ellis (2008) and Bybee (2008) among others, propose that it is an individual’s experience with language that will model his or her language use, that language per se emerges in use and that an individual will be very sensitive to the frequency with which elements occur.
Using conventional sequences in L2 French

in the input. Grammar, for example, is not a fixed set of pre-existing rules, but for the individual, grammatical regularities are extracted and emerge from the many exemplars that every individual has stocked, through processes of chunking. This is why, according to N. Ellis (2002a, 2002b), the acquisition of formulaic sequences serves two ends: formulaic sequences enable the extraction of creative rules and the use of formulaic sequences is a prerequisite for fluency and idiomaticity. This means that “formulas can break down” (N. Ellis 2002b), but there is, at the same time, a huge processing gain in keeping them as chunks. This idea is quite similar to Wray’s (2002) hypothesis formulated in a needs-only-analysis proposal for formulaic sequences. She suggests that the human mind is capable of double storage, i.e., a sequence such as my name is X can be stored both as a pragmatic chunk and the individual words can be stored separately. When it comes to the learning mechanisms of the sequences, Ellis suggests that there are two different modes of learning: one is initial, holistic chunk learning and the one is incremental learning. In other words, formulaic sequences need gradual strengthening in order to get automatized and entrenched in the mental lexicon, in Ellis’s own words: “nativelike fluency and idiomaticity require an awful lot of figuring out which words go together” (2002a: 157).

Ellis (2006, 2008) develops his theorizing on frequency effects with special regard to second language acquisition. Most people would probably agree that acquiring an L2 and an L1 are two different processes. Ellis (2008) suggests that the child acquiring its’ L1 can be seen as a tabula rasa whereas the L2 learner, already in possession of a linguistic system, can consequently be seen as a tabula repleta. He means that the same mechanisms that can explain the success of L1 acquisition, can also explain why many L2 learners fail in attaining nativelike levels of their L2. Since language learning is based on entrenchment according to Ellis, the entrenchment of the L1 system can sometimes interfere with the emerging L2 system. This may be explained by factors such as blocking and overshadowing (cf. Ellis 2006 for further discussion). In the present study, these factors will not be considered specifically, but it is important to be aware of them. However, an attempt will be made to take into account some factors from the usage-based paradigm, building on recent results from Ellis et al. (2008).

In three different experiments, Ellis et al. (2008) tested advanced non native speakers’ and native speakers’ sensitivity (speed of reading etc) to formulaic sequences with regard to on the one hand, their frequency of occurrence in a corpus and on the other, the sequences’ MI scores in a corpus. A sequence such as ‘is one of the’ has a high frequency, but a low MI score, whereas a sequence such as ‘in other words’ has a high frequency as well as a high MI score. Their results showed that native speakers were the most sensitive to sequences with high MI scores, whereas non-native speakers were not as sensitive to MI scores,
but their processing was linked to the raw frequencies of the sequences. It will be investigated if these findings coincide with results from the present study.

1.5. **Aim of the study**

The present study investigates the use of conventional sequences at different levels of spoken L2 French as well as in native speakers’ spoken productions. Even though the study is largely descriptive, some aspects of the results will be linked to factors affecting acquisition as stated by researchers in usage-based theory.

The design of the study is cross-sectional, i.e., it comprises interviews of different learners at different levels of L2 advancement. It is obviously always desirable to use a longitudinal design, but as most SLA researchers would presumably acknowledge, recordings of the same persons, covering their development from beginners to very advanced L2 users (periods of 5–10 years), are scarce to date. The advantage of using a cross-sectional design is that we can get an idea of how the second language develops, from beginner levels to very advanced levels, even though we only can study selected ‘stations’ on the L2 learners’ journey towards ultimate attainment.

To date, few studies have tried to find correlations between proficiency level and use of formulaic sequences, most studies having investigated only one proficiency level (e.g., Nesselhauf 2005, Granger 1998, Bolly 2008). However, Boers et al. (2006) and Lewis (2008) have found correlations between use of formulaic sequences and proficiency, something that will be further developed in the present study. The research questions are as follows:

1. What are the proportions of conventional sequences at different levels of L2 French and in native speakers’ production?
2. What are the differences between different levels of L2 users and native speakers in terms of distribution in different categories of conventional sequences?
3. What are the differences between different levels of L2 users and native speakers in terms of type variation of conventional sequences?
4. Can the results of the empirical study be linked to the findings of Ellis et al. (2008) regarding frequency of occurrence and MI-scores in a larger corpus?

2. **Participants and task**

2.1. **The participants**

Six groups of speakers are included in the study. Each group contains six participants. Four of the groups are part of the InterFra-corpus (cf. Bartning &
Schlyter 2004), a spoken learner and native speaker corpus of French, which has been collected at Stockholm University since the late 1980s and two additional groups, a very advanced L2 speaker group and another native speaker group, collected between 2003 and 2006.\(^4\) An outline of the informants is shown in Table 1.

The groups from the InterFra-corpus are: Adult beginners (recorded after one month of French studies), High school students (recorded during their fourth year of French in high school), Advanced university students (students having studied French in school for at least four years before entering the university and recorded during their third semester of French at university) and finally the Native speakers Erasmus (French native speakers, coming to Sweden as Erasmus students).

The two remaining groups, the Very advanced L2 users and the Native speakers Paris are part of the recently collected Corpus Forsberg, based on the same principles as the InterFra-corpus. The Very advanced L2 users have lived for at least 4.5 years in France. Some of them have studied French at school and for a couple of semesters in France, whereas others have received little instruction, but none of them are completely naturalistic learners. They are characterized by their long exposure to French on an everyday basis. An additional group of native speakers of the same age living in Paris were also recorded in order to match this group regarding and socio-professional background.

2.2. Tasks

The task investigated in this study is a semi-structured interview, lasting 15–20 minutes, conducted between a native speaker of French and a non-native speaker (Swedish L1). The interviews revolve round subjects such as the university, the French studies, France, travelling, cultural differences, friends, families and hobbies. All in all, 72 185 words were included for analysis and 7826 conventional sequences have been analyzed for the purpose of this study.

---

\(^4\) The lapse of time between the recordings could have been problematic if we were investigating some particular sequences, since the repertoire of conventional sequences develops over time. However, it has not been found, comparing the different sub-corpora, that the general proportion and the category distribution are affected by this difference.
Table 1. Description of the data

<table>
<thead>
<tr>
<th>Group of speakers</th>
<th>Age</th>
<th>N</th>
<th>Years of studying French</th>
<th>Time spent in France</th>
<th>No. words produced</th>
<th>No. of CS produced</th>
<th>No. of interviews/person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginner students</td>
<td>20–25</td>
<td>6</td>
<td>4 weeks – 1 semester</td>
<td>(a week of vacation)</td>
<td>2615</td>
<td>101</td>
<td>1</td>
</tr>
<tr>
<td>High school students</td>
<td>16–17</td>
<td>6</td>
<td>3.5 years – 4.5 years</td>
<td>(a week of vacation)</td>
<td>4060</td>
<td>172</td>
<td>1</td>
</tr>
<tr>
<td>Advanced university students</td>
<td>21–25</td>
<td>6</td>
<td>4-6 years in school, 2 semesters at university</td>
<td>1 month – 1.5 years</td>
<td>11229</td>
<td>1132</td>
<td>1</td>
</tr>
<tr>
<td>Very advanced L2 users</td>
<td>23–30</td>
<td>6</td>
<td>2 semesters – 7 years (school or/and university)</td>
<td>4.5 years – 11 years</td>
<td>20021</td>
<td>1754</td>
<td>1</td>
</tr>
<tr>
<td>Native speakers Paris</td>
<td>23–33</td>
<td>6</td>
<td>—</td>
<td>—</td>
<td>19242</td>
<td>2435</td>
<td>1</td>
</tr>
<tr>
<td>Native speakers Erasmus</td>
<td>19–25</td>
<td>6</td>
<td>—</td>
<td>—</td>
<td>15018</td>
<td>2234</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>—</td>
<td>36</td>
<td>—</td>
<td>—</td>
<td>72185</td>
<td>7826</td>
<td>—</td>
</tr>
</tbody>
</table>
3. Categories of conventional sequences

3.1. Categorization of conventional sequences

As suggested by Conklin & Schmitt (2008:72), “formulaic sequences are used in a wide variety of ways”, and a categorization destined to cover language in use needs to account for these. Nattinger & DeCarrico (1992) mention the transactional and the interactional nature of language. Language use is transactional when we want to convey meaning and interactional when the social interaction is the major goal of our language production. They also suggest that some conventional sequences, or ‘lexical phrases’ in their terminology, are transactional and that others are mainly interactional. Erman & Warren’s (2000) categorization of prefabs coincides quite well with these proposals. They present a categorization that takes into account all the different linguistic categories in which we find conventional sequences: lexical/referential (like prendre le petit déjeuner ‘have breakfast’), grammatical (quantifying/determining such as la plupart ‘the majority’) and discursive (connecting/managing speech such as c’est vrai que ‘as a matter of fact’). At present, there actually seems to exist some sort of consensus in the phraseological literature on how the sequences should be classified. Granger & Pacquot (2008) recently proposed a similar categorization for phraseological units, synthesizing several categorizations proposed earlier. They divide the sequences into Referential, Textual and Communicative sequences.

The present categorization largely follows the one proposed by Erman & Warren (2000), but which has been slightly modified and updated in Erman, Forsberg & Fant (2008):

The categories included in the present study – with examples from the corpus – are:

Lexical CSs: Clausal/Propositional:
- je m’appelle X ‘my name is’
- y a pas de soucis ‘no problem’
- on sait jamais ‘you never know’

Phrasal/Denotative:
- avoir envie ‘to feel like’
- faire du sport ‘practice a sport’
- poser une question ‘pose a question’

Grammatical CSs:
- un peu ‘a little bit’
- en majorité ‘in majority’
- l’un l’autre ‘each other’
Discursive CSs:  
c’est vrai que ‘as a matter of fact’  
au niveau de ‘when it comes to’  
disons que ‘let’s say that’

Each of these categories will now be presented in more detail.

3.2. Lexical CSs

Lexical CSs incorporate at least one content word. They are used for extralinguistic reference and denote actions (such as faire la fête ‘to party’), states (avoir peur ‘to be scared’), objects (pomme de terre ‘potatoe’) and so on. They are divided into Clausal sequences, which are full clausal propositional language-specific sequences, often with pragmatic connotations among which conversational routines are probably the best known, and Denotative sequences, which are primarily being used for their denotative meanings, and which as a rule constitute phrasal sequences, sometimes with open slots, such as X mettre X au courant (‘keep X posted on X’). Below is an example with several Lexical CSs.

(1) E: là je travaille oui. et la dernière emploi /  
c’est-à-dire là où je suis maintenant /  
ça fait deux ans et c’est une / société  
de gestion de portefeuilles.  
I: d’accord.  
E: donc / revirement complet (RIRE) de donc... (Jennifer, VA)

Example (1) above is taken from an interview with the very advanced L2 user Jennifer. It contains several examples of Lexical CSs, namely: ça fait X TIME UNIT ‘since TIME UNIT’, gestion de portefeuilles ‘portfolio management’ and revirement complet ‘complete change’. These certainly appear to be Lexical CSs according to the description given above, but how do we really know that we are dealing with a conventional sequence as opposed to a more creative construction? Erman & Warren (2000) make use of the criterion restricted exchangeability for identification of prefabs (their term). In order for a sequence to qualify as a prefab, it is not possible to exchange one of the words for a synonymous word without change of meaning or loss of idiomaticity (Erman & Warren 2000: 32).

The first step is to find the Lexical sequences that meet the restricted exchangeability criterion. This is then complemented by searches on Google and

5. Henceforth the abbreviations A = Advanced learner, VA = Very advanced L2 user, and NS = Native speaker will be used within the examples.
Using conventional sequences in L2 French

through the judgement of two native speakers. The Google tests are carried out following a specific procedure. To test to what extent restricted exchangeability applies to a sequence, an analogous sequence, which has been subject to one of the following commutations, is constructed:

1. One of the words is exchanged for a synonymous word
2. One of the words is exchanged for an antonymous word (for example "ça marche mal" ‘it works bad’ instead of "ça marche bien" ‘it works well’)
3. Change of article (from definite to indefinite or absence of article)
4. Change of number (from plural to singular or vice versa)
5. Change in word order (for example "égalité femmes/hommes" ‘equality between women and men’ instead of "égalité hommes/femmes" ‘equality between men and women’)

We shall now apply this procedure to some of the sequences present in example (1) above. Let us start with "gestion de portefeuilles" ‘portfolio management’, with ‘administration’ replacing ‘gestion’. A Google search yields 184,000 hits for the first and a mere 119 for the second. Another example is "revirement complet" ‘complete change’ which obtains 9330 hits, compared to 0 for "revirement entier" ‘entire change’. It is suggested that these tests allow us to make predictions concerning conventionalization and native speaker preference. However, the question naturally arises concerning the boundaries of preferences based on the Google searches. Where should we set the threshold for counting a sequence as conventional? In this study, a sequence has to be twice as frequent as the alternative sequence to qualify as a conventional sequence.6

3.3. Grammatical CSs

Grammatical CSs are units which, as opposed to Lexical CSs, have no extralinguistic reference. The grammatical CSs do not denote entities and actions, but serve as quantifiers, determiners and aspebook and temporal markers. They operate at the phrase and sentence level, have restricted semantic intension and are typically quite fixed. Consequently, the restricted exchangeability criterion also applies to these sequences. We all know that sequences such as "une plupart" ‘one majority’ (instead of "la plupart" ‘the majority’) or "un grand peu" ‘a big bit’ (instead of "un petit peu" ‘a little bit’) are inconceivable. These sequences have thus become so fixed that when we try to break them down and exchange one of their components, the construction appears to be ungrammatical. This is

---

6. In addition to Google searches, two native speakers were presented with the two alternative sequences and asked to indicate the sequence they perceived as being the more natural and frequent one in everyday conversation. The native speakers’ preferences almost always coincided with the Google results.
why, although it would have been possible, Google searches or native speaker judges were not used when identifying these sequences. From the original subdivision made by Erman & Warren (2000), the following subcategories are kept:

**Determiners:**
- une sorte de ‘a kind of’, le(la)les même (s) ‘the same’

**Quantity and degree expressions:**
- un petit peu ‘a little bit’, la plupart ‘the majority’

**Proforms:**
- les uns les autres ‘each other’, quelque chose ‘something’

**Aspectual markers:**
- être en train de (expression of simultaneity), venir de (to just have done sth)

Example (2) below from the corpus shows the proform *n’importe quel(le)* ‘whatever’.

(2) E: parce que les gens ont la liberté. // X on / pour
/ (SOUPIR) on doit // pouvoir acheter du vin
*n’importe quel jour n’importe quelle* heure
(I :mhm) / comme toutes les autres e:h produits / . (Christina, A, INT4)

It is important to stress that this category conveys a rather limited repertoire of what is grammatical in a language and therefore should not be considered as a characterization of what grammar is. Instead, they are the subcategories that were kept from Erman & Warren’s original category *Grammatical pre-fabs*, where more sub-categories were included. Some of these have now been placed in the next category to be presented, which will be explained below.

### 3.4. Discursive CSs

Discursive CSs, like the Grammatical CS, do not have either extralinguistic reference or propositional content. The members of this group typically function as information- or argumentation-structuring devices or as means for speech- and interaction-regulation. Included are presentative devices, conjunctions, adverbials and discourse markers. To make a clear distinction between Grammatical and Discursive CSs, one can say that Grammatical CSs are a limited set of structures, mainly determining and quantifying the propositional content in an utterance (*La plupart de ses amis étaient là*, ‘Most of his friends were there’), whereas Discursive CSs serve as sentence builders or links in the organization of discourse (*C’est bien que tu sois venu* ‘It is good that you came’, *À mon avis, il faut l’éviter* ‘In my opinion, one has to avoid it’).
The division between grammar and text is obviously debatable. It is possible to consider connecting and linking devices as Grammatical CSs as well (and not as Discursive CSs), but it was meant to keep the quantifying/determinative sequences apart from the text-structuring devices, since they are quite different. The sequence *c’est* (‘it is’) can be taken as an example of the difficulty in separating Grammatical and Discursive CSs. *C’est* is placed in the Discursive CSs category, since it functions both as an information structuring device and as an own-speech management marker in spoken discourse (mainly as a reformulation device). Some would consider the first, presentative function as clearly grammatical, contributing to the generation of syntax (cf. Bartning & Hammarberg 2007). However, *c’est* can also be considered as a connector (cf. Riegel et al. 2004: 617) and therefore more linked to text than to grammar. There is, without no doubt, more work to be done in fine-tuning the distinctions between Grammatical and Discursive CSs, but for the purpose of this specific study, the present classification will be used. It is hoped that the explanations above and the examples have illustrated what type of sequences are included in each category.

Discursive CSs, like Grammatical CSs, are so fixed in form that they do not require Google searches for ascertaining their prefabricated status. Below are listed the subcategories of Discursive CSs included in the study. Since discourse markers are polyfunctional, they have different functions in different contexts. For instance, *c’est vrai que* ‘as a matter of fact’ can be both a textual marker (connector), but it can also serve as an own-speech management marker and many speakers use it in abundance, natives as well as non-natives. Since this is a quantitative study, a decision was made to place the sequence in the sub-category which corresponds to the most frequent function of the sequence in question.

a. Textual markers (information/text-structuring devices)
   - *c’est* ‘it is’
   - *il y a* ‘there is’
   - *parce que* ‘because’
   - *par contre* ‘on the other hand’

b. Own-speech management markers
   - *je crois (que)* ‘I think that’
   - *je veux ire* ‘I mean’
   - *c’est-à-dire (que)* ‘That is’
   - *disons que* ‘let us say that X’

c. Interactive markers
   - *bien sûr* ‘of course’
   - *d’accord* ‘OK’
   - *tout à fait* ‘absolutely’
   - *ben ouais* ‘sure’
Having presented the analytical categories, it is now time to move on to the results of the empirical study.

4. Results: proportion of CSs, category distribution and type frequencies

First of all, let us have a look at the results regarding quantity of conventional sequences in the productions of the different groups of speakers.

4.1. Quantity of conventional sequences in learner and native speaker productions

Table 2 accounts for the number of CSs/100 words used by different groups of participants.

Table 2 shows an increase in the quantity of CSs used, following the different levels of L2 acquisition included in the study. However, the differences are not significant between every level of L2 advancement, A non-parametric ANOVA (Kruskal-Wallis) shows no significant differences between the two lowest levels, i.e., the Beginners and High school students, which is not very surprising considering results reported on the same groups for other linguistic phenomena (Sanell 2007, Lindström 2008). On the other hand, one would think that 1 month of French at university vs. 3.5 years would yield significant differences, but apparently it does not, at least not with regard to use of CSs on this task. However, both these low level groups differ significantly from the advanced university students ($p < 0.001$) and, obviously, from the Very advanced L2 users and the Native speaker groups. The Advanced learners differ ($p < 0.05$), in turn, from Very advanced L2 users, so the number of CSs does increase with the time spent in the TL community as one would expect. Interestingly, these Very advanced L2 users do not differ significantly from any of the native speaker groups. This suggests that L2 users immersed

<table>
<thead>
<tr>
<th>Group</th>
<th>CS/100 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginners</td>
<td>3.8</td>
</tr>
<tr>
<td>High school students</td>
<td>4.3</td>
</tr>
<tr>
<td>University students</td>
<td>9.9</td>
</tr>
<tr>
<td>Very advanced</td>
<td>12.1</td>
</tr>
<tr>
<td>NS Erasmus</td>
<td>10.1</td>
</tr>
<tr>
<td>NS Paris</td>
<td>11.6</td>
</tr>
</tbody>
</table>
in the TL community use at least the same quantity of CS as the native speakers. These results agree with those presented by Siyanova & Schmitt (2008), who found no differences in terms of number of adequate collocations used in their study comparing advanced learners of L2 English and native speakers. In conclusion, this part of our study suggests that an increased number of CSs is, as Yorio (1989) proposed, an indicator of second language development. These results are fully in line with the predictions of a frequency-based SLA theory – the more a person is exposed and immersed in the TL, the more he/she will perceive, acquire and use the preferred sequences of the TL community.

4.2. Quality of CS use: Category distribution in non-native and native speaker production

Table 3 accounts for the category distribution in the different groups of participants.

It is obvious that the Lexical CSs is the category which distinguishes the most the first three groups of learners from the very advanced L2 users. A One-Way ANOVA, using Tukey-Kramer’s multiple comparisons, yielded the following results. Beginners ($p < 0.001$), High school learners ($p < 0.001$), and University students ($p < 0.001$) all score significantly lower as compared to the Very advanced L2 users and the Native speakers. In contrast, no significant differences were found between Very advanced L2 users and the two Native speakers groups. No significant differences are thus found between the groups in the foreign language setting, but differences are observed between all these groups and the L2 users. In the analysis of the quantity of CSs, no differences were found between L2 users and Native speakers, and the result is the same when it comes to use of Lexical CSs. The results indicate furthermore that the extensive use of this category can be considered as a very late feature in L2, or

<table>
<thead>
<tr>
<th>Category</th>
<th>Lexical</th>
<th>Grammatical</th>
<th>Discursive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult Beginners</td>
<td>1.3</td>
<td>0.7</td>
<td>1.8</td>
</tr>
<tr>
<td>High school</td>
<td>1.3</td>
<td>0.6</td>
<td>2.4</td>
</tr>
<tr>
<td>University students</td>
<td>2.0</td>
<td>1.0</td>
<td>6.9</td>
</tr>
<tr>
<td>Very advanced</td>
<td>4.2</td>
<td>1.3</td>
<td>6.7</td>
</tr>
<tr>
<td>NS Erasmus</td>
<td>3.5</td>
<td>1.2</td>
<td>5.4</td>
</tr>
<tr>
<td>NS Paris</td>
<td>4.4</td>
<td>1.0</td>
<td>6.2</td>
</tr>
</tbody>
</table>
even a nativelike feature. With regard to frequency effects, this category should require more input than the other two, as will become apparent below.

The next category, the Grammatical CSs, do not show any significant development. The lack of development can probably be explained by the fact that this category contains a limited number of structures, which are learned quite early on, even though the repertoire is enlarged over time. This contrasts with the Lexical category, which potentially contains an unlimited number of sequences.

Discursive CSs show more development than the grammatical sequences. Running a One-Way ANOVA (Kruskal-Wallis), using Dunn’s multiple comparisons, the University students use more Discursive CSs than the beginners \((p < 0.01)\) and than the high school students \((p < 0.05)\), which is a sign of development. It could also confirm the tendency for advanced learners to use discourse organization devices extensively, which agrees with Raupach (1984) and Hancock (2000). However, the advanced learners do not differ in their use of Discursive CSs from the remaining groups, even if the figures look slightly higher. Most certainly, the use of discourse organization devices develop above the level of advanced university students, especially in terms of form/function mappings, variation of types and types used. However, these aspects will not be further dealt with in the present study.

4.3. Variation of CS use: Type frequencies

Table 4 accounts for the results regarding type\(^7\) frequencies in three of the categories: Lexical, Grammatical and Discursive CS:

In order to determine statistical differences between the groups, a One-way ANOVA, with Tukey-Kramers multiple comparisons test was used. As regards the first category, we again find no significant differences between the first three groups (foreign language setting), but we do find significant differences between Beginners \((p < 0.001)\), High school students \((p < 0.001)\) and University students \((p < 0.01)\) on the one hand and Very advanced L2 users on the other hand. The first three groups obviously display differences as compared to the Native speaker groups, which, again, are not found between the Very advanced L2 users and the Native speakers.

Grammatical sequences show, as before, very little acquisitional development. The only statistical difference to be found is between the two least advanced groups and the Native speaker groups \((p < 0.01)\).

---

\(^7\) By type frequency, we mean lemmas of CSs. For instance, avoir envie ‘to feel like’ is considered a type.
Table 4. Results in terms of types/100 words (mean values)

<table>
<thead>
<tr>
<th>Group</th>
<th>Lexical types/100 words</th>
<th>Grammatical types/100 words</th>
<th>Discursive types/100 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginners</td>
<td>1.2</td>
<td>0.4</td>
<td>0.8</td>
</tr>
<tr>
<td>High school students</td>
<td>1.1</td>
<td>0.3</td>
<td>0.9</td>
</tr>
<tr>
<td>University students</td>
<td>1.3</td>
<td>0.4</td>
<td>1.6</td>
</tr>
<tr>
<td>Very advanced</td>
<td>2.9</td>
<td>0.5</td>
<td>1.5</td>
</tr>
<tr>
<td>NS Erasmus</td>
<td>3.3</td>
<td>0.5</td>
<td>1.8</td>
</tr>
<tr>
<td>NS Paris</td>
<td>3.3</td>
<td>0.8</td>
<td>1.6</td>
</tr>
</tbody>
</table>

As for Discursive sequences, the pattern is somewhat different compared to the other measures carried out. We find no significant differences between Beginners and High school students, but this time, there are differences between these two groups and the University students \( (p < 0.001) \), which means that University students have a larger variation of discourse organization devices than the two lower proficiency groups. Finally, somewhat surprisingly, no differences are found between the University students and the Very advanced L2 users. Results from Forsberg (2008) indicate that the differences between these groups do not reside in the variation of types but in the types per se, i.e., some sequences are only used by the Very advanced L2 users. When we compare this last group with the two native speaker groups, the result is the same as for all the other measures used, i.e., no significant differences are observed between the Very advanced L2 users and both of the Native speaker groups.

As proposed in Section 4.2, quantity of CSs, category distribution in part, as well as some type frequencies are, sometimes significantly, linked to L2 development. But how can we explain the differences observed?

5. Relating L2 use of conventional sequences to usage-based theory

The most important results of the present analysis is probably the number of CSs increases between levels of L2 advancement, up to the level of very advanced L2 use, where no differences are found as compared to native speakers. As was stated in Section 1.4, there is reason to believe that conventional sequences (or formulaic depending on terminology choices) are acquired through frequency of exposure and repetitive use. Consequently, in the light of this theory, it comes as no surprise that the quantity of CSs increases in parallel with exposure to the TL and that the participants having lived in the TL community for 5-10 years have access to roughly the same quantity of conventional
sequences as native speakers, even though they may differ from a qualitative point of view.

Another important result is that Lexical CSs, such as *avoir affaire à* ‘to deal with’ or *à pied* ‘by foot’, is shown to be the latest acquired category, and that there is, surprisingly, no significant development between the three groups in the foreign language setting. However there are differences between all of these three groups and the Very advanced L2 users, who in turn display no significant differences as compared to the two Native speaker groups. Interestingly, the two ‘functional’ categories, Grammatical and Discursive CSs, do not display the same differences. As regards Grammatical sequences, almost no development can be observed at all the learner groups. When it comes to Discursive sequences, the only development is to be found between the lower levels and the advanced university student level, but then no differences again. It thus seems that the functional categories are more easily acquired than the Lexical category. Can the differences in how these categories are used/acquired also be explained within a usage-based framework?

In order to do be able to do so, we have to take into account not only the frequencies of occurrence of the sequences, but also how strongly the words in the sequence are associated with one another (Mutual information score MI). As was mentioned in Section 1.4, Ellis et al. (2008) test advanced native speakers’ and native speakers’ sensitivity (speed of reading etc) to formulas with regard to, on the one hand, their frequency of occurrence in a corpus and on the other, the sequences’ MI scores. Their results showed that native speakers’ speed of processing was highly dependent on the sequences’ MI scores, whereas non-native speakers were not as sensitive to MI scores, but their processing speed was linked to the raw frequencies of the sequences.

In order to explain why Lexical CSs are harder to acquire than Grammatical and Discursive sequences, one possibility is to check their frequencies and MI scores in a large corpus. A modest control corpus of French native speech has thus been compiled, drawing on different sub-corpora of French native speakers recorded at Stockholm university in 1996-2007 (41 interviews altogether). It includes only interviews on the same topics, in order for it to be comparable to the interviews of the present study. This corpus will henceforth be called *FRECS* (FRENch Control corpus Stockholm). The size, ca. 150,000 (149,493) words, is small compared to other corpora used for this type of measures, but
it was hoped that at least some tendencies could be found. The procedure is as follows: the five most frequent Lexical CSs, Grammatical CSs and Discursive CSs in the Native speakers Paris productions’ were tested for frequency and MI in FRECS. The figures are supposed to give us an idea of how the different categories are represented in French spoken language, i.e., the input to which our informants are exposed. As appears in the tables below, the Lexical CSs require that searches are performed on various forms of one and the same sequence, which is not the case for the other categories. In principle, all present tense forms were searched, and sometimes a couple of more tenses were included, since they appeared to be of importance during the searches.

Tables 5, 6 and 7 account for frequencies of occurrence in one small and one larger corpus as well as MI score in the larger corpus of a limited number of sequences of the three categories under study. The results are quite in line with those presented by Ellis et al. (2008). The table accounting for Lexical CSs frequencies in FRECS and their MI scores, show that the mean frequency is quite low: only 5.4 occurrences for each type, whereas the MI score is quite high viz. 10.3. It should be mentioned here that when it comes to MI scores, a threshold of 3 is often set for the establishment of collocations (cf. Siyanova & Schmitt 2008, Erman et al. forthcoming).9 Comparing the results for Lexical CSs with those of Grammatical and Discursive CSs, the differences, especially in terms of frequencies of occurrence, are quite striking. Grammatical sequences have a mean frequency of occurrence of 189 in FRECS and a mean MI score of 7.3. Discursive CSs have a mean of frequency of occurrence of 514 in FRECS and a mean MI score of 6.8. This means that Lexical CSs are very infrequent compared to the other two categories but they have a higher MI score, i.e., the words included in the Lexical CSs are more strongly associated than those included in the Grammatical CSs and Discursive CSs, which are, however, considerably more frequent than the Lexical CSs. It is also worth taking into consideration that the Lexical CSs are, as the table shows, quite often composed of sequences that need to undergo grammatical manipulations, such as the use of verb inflections. It is quite plausible that these manipulations also contribute to the relative difficulty of acquiring this category. The sequences in this category are not only infrequent – their form also varies depending on person, number and tense, whereas the Grammatical and Discursive CSs contain few open slots and allow for very little manipulation.

It thus seems like the results of our empirical corpus study could be linked to Ellis et al.’s (2008) findings. In their use of CSs, non-native speakers up
Table 5. Frequencies of occurrence and MI scores for Lexical CSs

<table>
<thead>
<tr>
<th>Type</th>
<th>Frequency in Corpus Forsberg</th>
<th>Frequency in FRECS</th>
<th>MI score in FRECS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aimer bien(^a) (lemma)</td>
<td>10</td>
<td>6</td>
<td>6.6</td>
</tr>
<tr>
<td>– aime bien</td>
<td></td>
<td>6</td>
<td>6.7</td>
</tr>
<tr>
<td>– aimes bien</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>– aimons bien</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>– aimez bien</td>
<td></td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>– aiment bien</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Avoir envie de X</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>– ai envie de</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>– as envie de</td>
<td></td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>– a envie de</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>– avons envie de</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>– ont envie de</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>– eu envie de</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>– avais envie(^b)</td>
<td>10</td>
<td>2</td>
<td>9.2</td>
</tr>
<tr>
<td>– aurais envie</td>
<td>2</td>
<td>10</td>
<td>7.8</td>
</tr>
<tr>
<td>Avoir peur de X</td>
<td>6</td>
<td>9.5</td>
<td>9.5</td>
</tr>
<tr>
<td>– ai peur de</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>– as peur de</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>– a peur de</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>– avons peur de</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>– avez peur de</td>
<td></td>
<td>9</td>
<td>15.3</td>
</tr>
<tr>
<td>– ont peur de</td>
<td></td>
<td>3</td>
<td>11.1</td>
</tr>
<tr>
<td>– eu peur de</td>
<td></td>
<td>2</td>
<td>8.4</td>
</tr>
<tr>
<td>Encore une fois</td>
<td>5</td>
<td>7</td>
<td>13.1</td>
</tr>
<tr>
<td>Site internet</td>
<td>5</td>
<td>5</td>
<td>11.6</td>
</tr>
<tr>
<td>Means</td>
<td>7.2</td>
<td>5.4</td>
<td>10.3</td>
</tr>
</tbody>
</table>

\(^a\) In the present study, the occurrences of Lexical CSs composed of verbs with inflections for person, number and tenses were regrouped under the infinitive. Since FRECS is not a lemmatized corpus, it was necessary to search for different forms of one same sequence, since some variants are more frequent than others.

\(^b\) More forms, such as avais envie and aurais envie were included for avoir envie, since they were found to be frequent, probably due to the genre of the interview, where the persons speaks about his wishes and plans.

until the very advanced levels, use more sequences that are highly frequent (Grammatical and Discursive CSs), but with lower MI scores, whereas the low frequent – high MI Lexical CSs, are only used extensively by the very advanced group and the native speakers. Hasty conclusions should not be drawn on such
Using conventional sequences in L2 French

6. Conclusion

By means of a cross-sectional design, the present study set out to investigate the development in overall use of conventional sequences in L2 French. Initially, a distinction was made between targetlike sequences and interlanguage sequences, the present study being concerned with targetlike sequences only. Another distinction was made between the notions of ‘formulaic’ and ‘conventional’. Following Bardovi-Harlig (subm) and Edmonds (2008), it was argued that the term formulaic often implies holistic storage, whereas holistic storage is rarely investigated. Since this study does not use any psycholinguistic measures, it seemed more accurate to label the sequences what they are viz. conventional sequences, nothing more, nothing less.

Furthermore, it was suggested that, among the different identification methods existing in the field of formulaic language, a phraseological identification

Table 6. Frequencies of occurrence and MI scores for Grammatical CSs

<table>
<thead>
<tr>
<th>Type</th>
<th>Frequency in Corpus Forsberg</th>
<th>Frequency in FRECS</th>
<th>MI score in FRECS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Un peu</td>
<td>64</td>
<td>462</td>
<td>5.6</td>
</tr>
<tr>
<td>Beaucoup de</td>
<td>21</td>
<td>147</td>
<td>3.6</td>
</tr>
<tr>
<td>Très très</td>
<td>17</td>
<td>98</td>
<td>5</td>
</tr>
<tr>
<td>Un petit peu</td>
<td>15</td>
<td>143</td>
<td>13.2</td>
</tr>
<tr>
<td>Quelque chose</td>
<td>13</td>
<td>95</td>
<td>9.4</td>
</tr>
<tr>
<td>Means</td>
<td>26</td>
<td>189</td>
<td>7.3</td>
</tr>
</tbody>
</table>

Table 7. Frequencies of occurrence and MI scores for Grammatical CSs

<table>
<thead>
<tr>
<th>Type</th>
<th>Frequency in Corpus Forsberg</th>
<th>Frequency in FRECS</th>
<th>MI score in FRECS</th>
</tr>
</thead>
<tbody>
<tr>
<td>C’est</td>
<td>328</td>
<td>1202</td>
<td>6.2</td>
</tr>
<tr>
<td>Il y a</td>
<td>116</td>
<td>369</td>
<td>12.3</td>
</tr>
<tr>
<td>En fait</td>
<td>81</td>
<td>275</td>
<td>4.4</td>
</tr>
<tr>
<td>Parce que</td>
<td>63</td>
<td>409</td>
<td>5.6</td>
</tr>
<tr>
<td>Et puis</td>
<td>32</td>
<td>315</td>
<td>5.4</td>
</tr>
<tr>
<td>Means</td>
<td>124</td>
<td>514</td>
<td>6.8</td>
</tr>
</tbody>
</table>

as small sample, but it is interesting to see that findings from processing do seem to have a connection with how the sequences are actually used in a corpus of spontaneous spoken production.
method would be the most appropriate for the purpose of this study, since the corpus is too small for statistical identification methods to be applied. It was also argued that a statistical method sometimes falls short in identifying sequences which are pragmatically conventionalized.

The data investigated consisted of interviews, with six informants in six groups of speakers: Adult beginners, High school students, Advanced university students, Very advanced L2 users and two Native speaker groups. Use of conventional sequences was studied from three different aspects: overall quantity of CSs used per 100 words, category distribution per 100 words and finally, types in the different categories per 100 words. It was found that quantity of conventional sequences increases across L2 learning: significant differences were found up to the Very advanced level, who, in turn did not display differences as compared to Native speakers. CSs can thus be considered a valid measure for acquisitional development.

When it comes to category distribution, Lexical CSs were found to be the category posing most problems for L2 users and only Very advanced users, again, displayed no quantitative differences compared to Native speakers. Interestingly, this category showed no development among all the groups in the foreign language setting. Discursive sequences show development up to the Advanced level, but then no quantitative differences were found. By and large, the same development discerned in category distribution, was found with respect to type frequency, where some differences are to be found between the L2 groups, but no differences again between the Very advanced and the native speakers. Naturally, this does not mean that there are no differences between very advanced L2 users and native speakers in how they use conventional sequences, only that these differences were not captured in the present study.

The results were then related to usage-based perspectives on language learning. Following Ellis et al. (2008), we wanted to investigate whether the results regarding category distribution could be linked to the factors of frequency and MI score. To this end, a small control corpus of French (FRECS) was used in order to establish frequencies and MI scores for the five most common sequences in each category. Hence, we were able to have a schematic impression of the characteristics of each category. It turned out that Lexical CSs had very low frequencies, had opens slots for grammatical manipulation, but quite high MI scores, whereas the two other categories, Grammatical CSs and Discursive CSs, who are used to a larger extent by most learner groups, had much higher frequencies, basically no open slots and lower MI scores than the Lexical CSs. These results could thus seem to be in line with Ellis et al.’s (2008): non-native speakers’ processing seems to be more related to raw frequencies of the sequences than to the cohesiveness of the sequences (MI score). In addition, it was suggested that opens slots for grammatical manipulation also constitute a difficulty for L2 speakers when acquiring conventional sequences.
One question raised by the present study is of course: why don’t we find differences between Very advanced L2 users and native speakers in any respect? Many studies show that differences do persist even at very high levels of L2 advancement (Abrahamsson & Hyltenstam 2008) and one would expect that this would be found in our study as well. This is probably due to several aspects of the present study’s design: the measures are quite general and do not take into account specific types and it should also be said that the interview as a task is probably not challenging enough for Very advanced L2 speakers, who need to be tested in a variety of communicative situations, including both processing and socio-pragmatic constraints. A design taking task variation into account would probably reveal interesting differences between these highly proficient, although not nativelike second language speakers. Having this study as a starting point, we now know that Lexical CSs is the most interesting CS category to focus on, especially when we want to investigate the upper limits of L2 use. An in-depth study of specific types of Lexical CSs, linked to variables such as frequency, MI score and L1 influence (Ellis 2006, 2008) could possibly shed some light on the determining factors for the (non)-acquisition of conventional sequences, often called the last stumbling block for L2 speakers.

Stockholm University

References


Bartning, Inge & Björn Hammarberg. 2007. The functions of a high-frequency collocation in native and learner discourse: The case of French c’est and Swedish det är. IRAL 45. 1–43.


