Preferred sequences of words in NS and NNS speech

Sylvie De Cock

Université catholique de Louvain

Abstract

As human beings, we are all creatures of habit. Most, if not all, aspects of our everyday lives, including language use, are in greater or lesser measure marked by routine and recurrence. This paper reports on some of the major findings of a large-scale corpus-driven analysis of the recurrent sequences of more than two single words that native speakers of English and advanced EFL learners tend to use as their routinized building blocks or preferred ways of saying things in spoken and written discourse (De Cock 2003). The aim of this paper is to explore the use of recurrent sequences of words in NS and NNS speech both from a quantitative and a more qualitative point of view. In the quantitative analysis I set out to test the validity of Kjellmer's often quoted but largely unproven assumption that learners'"building material is individual bricks rather than prefabricated sections" (1991: 124). The more qualitative analysis concentrates on some of the major functional differences between native speakers' and learners' preferred ways of saying things.

1 Introduction

Most aspects of human existence are marked by habit and recurrence. The way we use language is no exception to this: we have a tendency to recurrently say again what we have said before. As was noted by Kjellmer (1994: ix), "[t]here is no doubt that natural language has a certain block-like character. Words tend to occur in the same clusters again and again." The prevalence of these recurrent clusters of words has been brought to light in an unprecedented manner by recent corpus linguistic studies of combinations of words (Sinclair 1991, Kjellmer 1994, Altenberg 1998, Altenberg and Olofsson 1990, Moon 1998, Biber *et al.* 1999, Biber and Conrad 1999, Cortes 2002a, Cortes 2002b, Biber 2003). These studies have been particularly instrumental in widening the scope of phraseology in that they have demonstrated that beside the psychologically salient but comparatively rare 'classical' idioms with figurative meanings such as *kick the bucket*, which used to lie at the heart of traditional phraseology, there is a large number of recurrent sequences of words, which, even though they have not traditionally been labelled as phraseological and tend to go virtually unnoticed in everyday language because they are not very salient psychologically, can however not be dismissed as uninteresting from a broader phraseological point of view. Although linguists such as Cowie (1999) have convincingly shown that frequency of recurrence is not a criterion for strict phraseological status, it can nevertheless be seen to give us some guarantee that the strings are current in the language variety under study (Fernando 1996). In other words, recurrent sequences of words give us access to what is typical (Stubbs 2002), or in Béjoint's words (2000: 216) to the "tendencies in the encoding of text by native

speakers." This is particularly significant because "[t]hese tendencies are part of the mastery of the language, (…) there are preferred sequences" (Béjoint 2000: 216) or in Schmitt and Carter's words (2004: 10) "they are the preferred choice." Recurrent sequences of words as typical or preferred sequences, preferred ways of saying/putting things or basic building blocks can directly be related to one of the four components making up Hymes's communicative competence, viz. 'Whether (and to what degree) something is done' or, to put it differently, what is actually performed.[1] As Hymes (1972: 286) points out "[s]omething may be possible, feasible, and appropriate and not occur." He also believes that "[t]he capabilities of language users do include some (perhaps unconscious) knowledge of probabilities." Recurrent sequence of words can thus also be seen to reflect and to play a major role in 'idiomaticity' taken in the wide sense of Pawley and Syder's (1983: 191) 'native-like selection', i.e. "the ability of the native speaker routinely to convey his meaning by an expression that is not only grammatical but also nativelike (…) he selects a sentence that is natural and idiomatic from among a range of grammatically correct paraphrases."

This paper reports on some of the major findings of a large-scale corpus-driven analysis of the recurrent sequences of two or more single words that native speakers (NS) of English and advanced EFL learners (NNS) tend to use as their routinized building blocks or preferred ways of saying things in spoken and written discourse (De Cock 2003). The focus of this paper is on spoken discourse as most studies of native speaker and learner recurrent sequences have mainly been concerned with written discourse (e.g. Milton and Freeman 1996, Kjellmer 1994, Cortes 2002a and 2002b, Sugiura 2002, Jones and Haywood 2004, Schmitt et al. 2004).

The aim of this paper is to explore the use of recurrent sequences of words in NS and NNS speech both from a quantitative and a more qualitative point of view. One of the aims of the quantitative analysis is to test the validity of Kjellmer's often quoted but largely unproven assumption that learners' "building material is individual bricks rather than prefabricated sections" (1991: 124). The more qualitative analysis concentrates on some of the major functional differences between native speakers' and learners' preferred ways of saying things.

2 Data and method

The spoken data used in the study consists of a corpus of informal interviews with EFL learners (henceforth NNS corpus) and a comparable control native speaker corpus (henceforth NS corpus). Each corpus totals approximately 100,000 words of interviewee speech: the NS corpus is made up of 117,417 words and the NNS corpus of 90,300 words. The learner spoken corpus is the French component of the Louvain International Database of Spoken English Interlanguage (LINDSEI). The LINDSEI project was launched in 1995 at the Centre for English Corpus Linguistics, Université catholique de Louvain, as the spoken counterpart of the International Corpus of Learner English (ICLE, Granger 1998a). A number of other LINDSEI components have been or are currently being complied: Chinese, Italian, Japanese, Spanish, Swedish, German and Bulgarian to date. The native speaker corpus – the Louvain Corpus of Native English Conversation (LOCNEC), which was compiled within the framework of De Cock 2003 – is actually something of a misnomer as it is made up of informal interviews and not (spontaneous) conversations.

The informal interviews, which last about fifteen minutes each, were recorded with the consent of the participants. The interviews are of similar length (approximately 2,000 words of interviewee speech each) and follow the same set pattern: the main body of the interviews took the form of an informal and open discussion mainly centred around topics

such as university life, hobbies, foreign travel or plans for the future, although many different subjects were touched upon when the interviewees introduced them into the conversation. Each interview starts with one of three topics (an experience that taught them a lesson, a film or play they liked/disliked, a country that impressed them), which the students were given a few minutes to choose and think about. This was designed to make the interviewees, and especially the learners, feel at ease. The students were, however, specifically asked not to make any notes for the sake of spontaneity as it was intended for the spoken productions to be as spontaneous as possible. Each interview concludes with a short picture-based story-telling activity. The interviews were transcribed using a broad orthographic transcription scheme.

The 50 non-native interviewees are labelled as 'advanced' on the basis of an external criterion: they are third and fourth year students of English. The students, who were all students at the Université catholique de Louvain, were native speakers of French aged between twenty and twenty-six. The ratio of males to females is 2: 3. The higher proportion of female interviewees in the corpus is a direct reflection of the higher number of female than male students of English. A detailed biographical profile is available for each learner in the form of a questionnaire they were asked to fill in at the time of the interview.

The 50 native speaker interviewees were all students at Lancaster University, Great Britain. While the majority were undergraduates, mainly in their first or second year but also in their third or fourth year, some of the interviewees were postgraduates. Although the vast majority of the native informants were either Linguistics or English Language students, some of them were reading subjects such as French, Chemistry or Management. The students were aged between eighteen and thirty and all of them British. The ratio of males to females is identical to that in the NNS corpus: 2: 3. The students who took part in the data collection were all volunteers who gave up their own time to attend the interview. The native speaker interviewees were also asked to fill in biographical profiles.

Although ensuring comparability with the NNS corpus was a major consideration when building the NS corpus, there are differences between the two corpora other than the native/non-native distinction. The degree scheme and year of study (and age) of the native and non-native interviewees do not correspond exactly. This is, however, not considered as a major drawback because, whereas it is essential for the non-native speakers to study English (as part of a degree in Languages) and to be in their third or fourth year to qualify as advanced learners of English, these requirements are not deemed relevant when it comes to native speakers.

The method used to investigate frequently recurring sequences of words in NS and NNS speech in De Cock 2003 is the corpus-driven 'recurrent word combination' method used by Altenberg (1998) in his work on the phraseology of spoken NS English in the London-Lund Corpus of spoken English. The 'recurrent word combination' method involves the automatic extraction of sequences of word forms of length *n* which recur in identical form with frequency greater than *m* from a corpus using specialised software. Both the sequence length (two, three, four, etc. words) and the frequency threshold, below which sequences are not reported, are specified by the user.

The 'recurrent word combination' method is an illustration of corpus linguistic methodology at its most heuristic, i.e. as a raw discovery procedure. The method does not presuppose any linguistic category or pre-established list of sequences. This type of raw discovery procedure can be regarded as particularly well-suited to the study of native speakers' and learners' routinized building blocks, not least because of their familiar, common and psychologically non-salient character, and because there are as yet no

comprehensive widely agreed upon lists of preferred ways of saying things. The results yielded by the automatic extraction are a useful and powerful starting point as they arguably lead the researcher to take into consideration a series of frequently used clusters he or she may otherwise have overlooked because of their lack of psychological salience. Using any list of prefabs drawn up on the basis of dictionaries and/or previous studies of the phrasicon would inevitably have limited this study to these very sequences. What is more, when it comes to uncovering preferred ways of saying things in learner language, the corpus-driven discovery procedure is absolutely essential as there simply is no pre-established list of NNS prefabs. Using a pre-established NS list approach would give us access to only an extremely limited part of learners' recurrent phrasicon. In addition, as was suggested by Raupach (1984), the observed recurrent use of a string of words in learner language can often be taken as a useful indicator of routinized status.

The investigation is limited to two-, three-, four-, five- and six-word sequences that occur at least 12, 6, 4, 3 and 3 times respectively in the NS or NNS corpus. A different frequency threshold was set for each sequence size bearing in mind that the length of recurrent word combinations is inversely related to their frequency (Altenberg 1990). The thresholds were also scaled so that approximately 10% - 12% of recurrent sequence types are taken into consideration for each length. Frequency thresholds were adopted mainly because the focus of this thesis is on routinized building blocks. Following Altenberg (1998), the thresholds are regarded as giving us at least some guarantee that the sequences have some currency in NS and NNS speech. They also go some way towards ensuring that the sequences extracted are not the result of local textual repetition.

It is important to point out that studies of learners' phrasicon that make use of the 'recurrent word combination' method are rather few and far between (Milton and Freeman 1996, Sugiura 2002, Adolphs and Durow 2004). Unlike most investigations of recurrent sequences in NS and learner speech and writing, our study is not restricted to one specific sequence length (e.g. Cortes 2002a and b and Biber 2003 focus on four-word sequences; Adolphs and Durow 2004 concentrate on three-word sequences) and two-word sequences, which have usually been left out because of their sheer number, are included in the analysis.

Before setting off on our recurrent sequence expedition, let us briefly turn our attention to the (hopefully reader-friendly) system that will be used to give an indication of the frequencies with which the sequences can be seen to recur in the corpora.

Symbol	**Frequency**
-	not recurrent at or above frequency threshold
△	recurrent sequences occurring less than 10 times per 100,000 words **(NB:** 3-word sequences recur at least 6 times, 4-word sequences recur at least 4 times, 5- and 6-word sequences recur at least 3 times)
▲	recurrent sequences occurring 10 to 19 times per 100,000 words
▲▲	recurrent sequences occurring 20 to 49 times per 100,000 words
▲▲▲	recurrent sequences occurring 50 to 74 times per 100,000 words
▲▲▲▲	recurrent sequences occurring 75 to 99 times per 100,000 words
▲▲▲▲▲	recurrent sequences occurring over 100 times per 100,000 words

Table 1. Frequency of recurrence of preferred sequences

Word-sequence-oriented notions of type and token should also be defined at this stage: each different sequence of words is considered a different type and each occurrence of a sequence of words a different token.

3 Analysis

Our exploration of the use of recurrent sequences of words in NS and NNS speech is divided into two parts. In a first quantitative part I set out to draw a general picture of recurrence in NS and NNS speech. Kjellmer's 'learner individual bricks vs. NS prefabricated sections' hypothesis will also be tested. The second part is more qualitative as it deals with functional aspects of recurrent sequences of words. The focus is on one group of sequences that are markedly underused by the learners in our corpus.

3.1 Quantitative analysis of recurrent sequences in NS and NNS speech

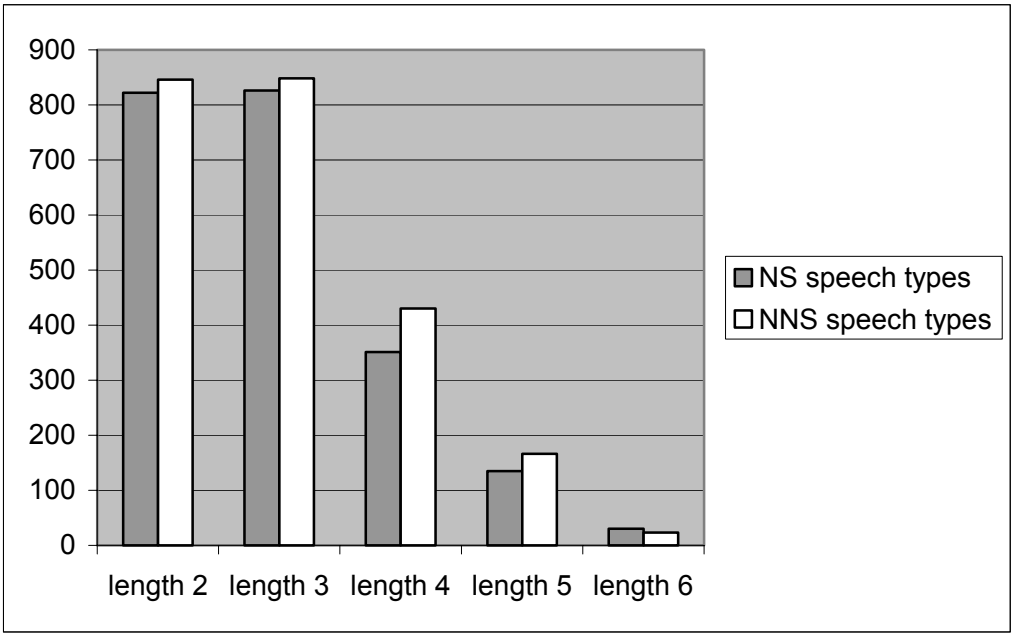
It is important to emphasise that Kjellmer's hypothesis that learners use fewer prefabricated sections than native speakers is only partly tested: the focus is on one specific group of learners, i.e. advanced EFL learners of French mother tongue, and on one particular set of sequences of words, i.e. **highly** recurrent **continuous** sequences of words of all kinds, which most probably only loosely correspond to the 'prefabricated sections' Kjellmer had in mind when he formulated his hypothesis.

3.1.1 NS speech vs. NNS speech

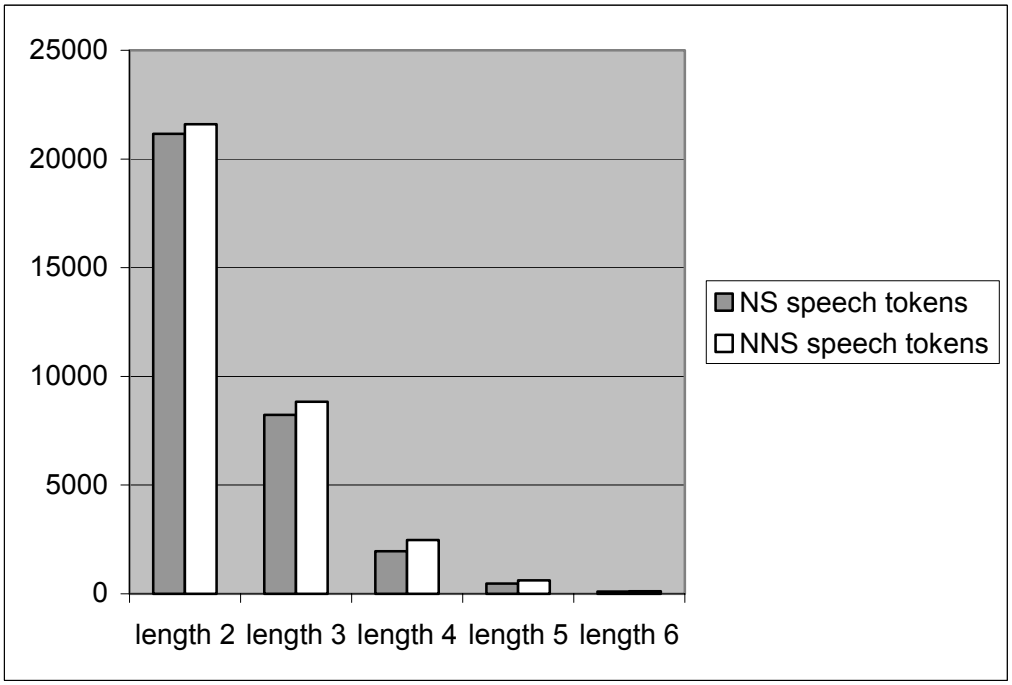
Length	NNS vs. NS speech types	NNS vs. NS speech tokens
2	NNS speech **=** NS speech	NNS speech = NS speech
3	NNS speech **=** NS speech	NNS speech >>> NS speech
4	NNS speech > NS speech	NNS speech >>> NS speech
5	NNS speech **=** NS speech	NNS speech >>> NS speech
6	NNS speech = NS speech	NNS speech = NS speech

Table 2. Number of highly recurrent sequence types and tokens in NS and NNS speech (based on relative frequencies per 100,000 word sequences) [2]

The results displayed in Table 2 and in Graphs 1 and 2 do not, on the whole, lend support to Kjellmer's hypothesis. There are slightly more 2-, 3- and 5-word sequence types in NNS speech but the differences are not statistically significant. There are significantly more 4-word sequence types (at $p \leq 0.05$) in the NNS corpus. There are however slightly fewer 6-word sequence types in the learner corpus but the difference is not statistically significant. The results for tokens point to a highly significant overuse of 3-, 4- and 5-word sequence tokens (at $p \leq 0.005$) and to a slight overuse of 2- and 6-word sequence tokens in the learner corpus. It is noteworthy that the results for NS and NNS 3-word sequence types run counter to Altenberg's observation (1990) (see also Milton and Freeman 1996 and Biber *et al.* 1999) that length and frequency of combinations of words are inversely related: there are slightly more 3-word than 2-word sequence types (NS 2-word types = 822 vs. NS 3-word types = 826; NNS 2-word types = 846 vs. NNS 3-word types = 848, these are relative frequencies per 100,000 word sequences).



Graph 1. NS speech vs. NNS speech types



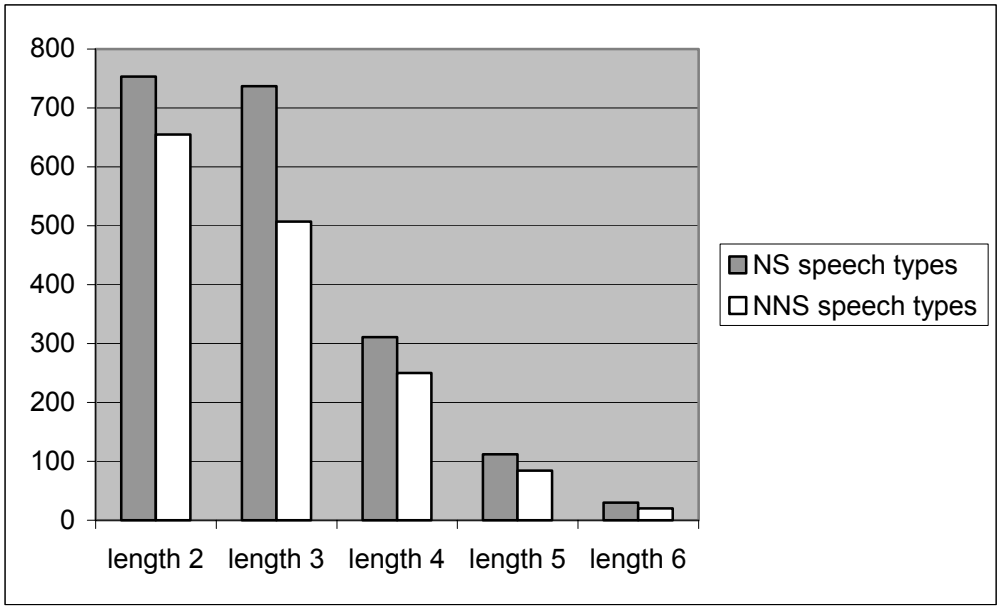
Graph 2. NS speech vs. NNS speech tokens

A closer examination of the sequences in NS and NNS speech provides us with some interesting insights into the types of sequences used in the two varieties. The lists of sequences from the NNS corpus appear to be made up of a higher proportion of sequences containing repeats (e.g. *I I* or *the the*) and hesitations items (e.g. *er* or *erm*) than those from the NS corpus. Comparing NS and NNS spoken sequences excluding repeats and hesitation sequences will enable us to evaluate the extent to which the use of such combinations affects learners' unexpected overuse of sequences of words.

Length	NNS vs. NS speech types	NNS vs. NS speech tokens
2	NNS speech **<** NS speech	NNS speech **<<<** NS speech
3	NNS speech <<< NS speech	NNS speech **<<<** NS speech
4	NNS speech **<** NS speech	NNS speech **<<<** NS speech
5	NNS speech = NS speech	NNS speech = NS speech
6	NNS speech = NS speech	NNS speech = NS speech

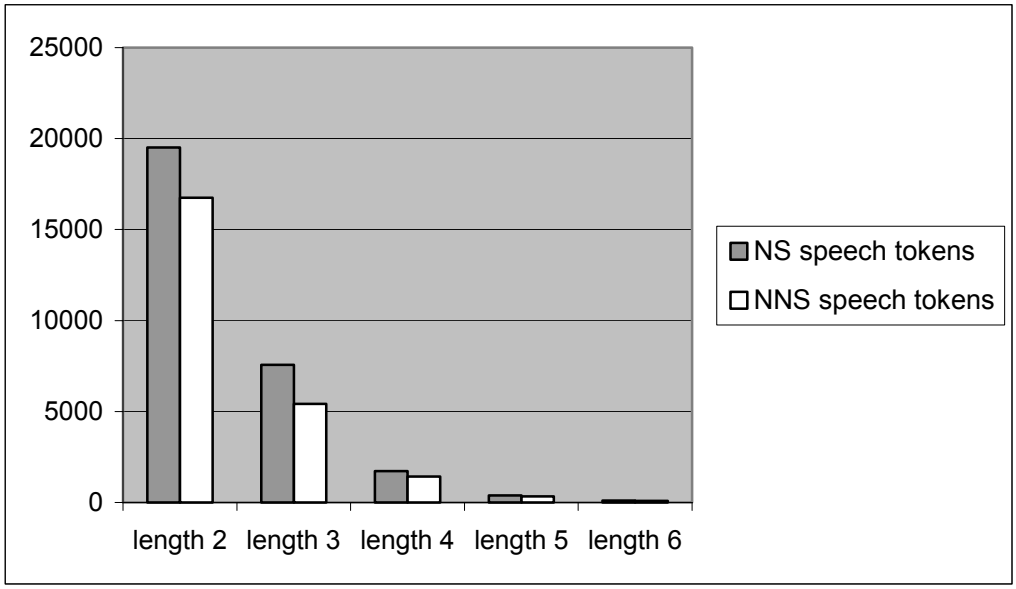
Table 3. NS and NNS speech types and tokens **(excluding sequences containing repeats and hesitation items)** – based on relative frequencies per 100,000 word sequences

Graphs 3 and 4 and Table 3 paint a very different picture from the one above as it reveals that learners tend to underuse rather than overuse those 2-, 3-, 4-, 5- and 6 recurrent sequences that do not contain repeats and/or hesitation items both in terms of types and tokens. Whereas the differences are statistically not significant for lengths 5 and 6 (for both types and tokens), they are significant at $p \leq 0.05$ for 2- and 4-word sequence types and highly significant at $p \leq 0.005$ for 3-word sequence types and 2-, 3- and 4-word sequence tokens. These results confirm rather than infirm Kjellmer's hypothesis.



Graph 3. NS speech vs. NNS speech types **(revisited)**

It is worth noting that the higher number of NS and NNS 3-word sequence than 2-word sequence types observed above appears to have been due to high proportions of 3-word sequences containing repeats and hesitation items. The figures for NS and NNS 2- and 3-word sequence types are now in line with Altenbergs' observation that length and frequency of sequences of words are inversely related: there are slightly more 3-word than 2-word sequence types (NS 2-word types = 753 vs. NS 3-word types = 737; NNS 2-word types = 655 vs. NNS 3-word types = 507, these are again relative frequencies). Table 4, which lists the top 20 NNS highly recurrent 3-word sequences, illustrates the types of strings containing hesitation items and/or repeats that can be found in lists of 3-word sequences.



Graph 4. NS speech vs. NNS speech tokens **(revisited)**

Rank	3-word sequence	Rank	3-word sequence
1	I don't know	11	and er we
2	111	12	and so on
3	and it was	13	no no no
4	and er well	14	but 11
5	the the the	15	to to to
6	and er I	16	11 was
7	and er the	17	yes yes yes
8	it was really	18	a lot of
9	it was er	19	I would say
10	it was a	20	I went to

Table 4. Top twenty 3-word sequences in NNS speech

Length	NS speech types	NNS speech types	NS speech tokens	NNS speech tokens
2	8.39%	24.68%	9.93%	32.03%
3	12.43%	44.74%	11.58%	46.89%
4	12.30%	44.79%	13.02%	46.72%
5	16.96%	50.43%	17.25%	49.29%
6	0%	13.33%	0%	22.97%

Table 5. Proportions (percentages) of word sequence types and tokens containing repeats and/or hesitation items in NS and NNS speech

Tables 5 and 6 display the frequencies and proportions of word sequence types and tokens that contain repeats and/or hesitation items in NS and NNS speech. As the Tables show, the learners in our corpus use approximately 3 to 4 times as many sequences that contain repeats and/or hesitation items as native speakers both in relative and absolute terms. The differences between the native speakers and the learners are statistically highly significant. The top 20 NNS 3-word sequences (Table 4) incidentally also illustrate the high proportion of sequences containing repeats and/or hesitation items: 12 out of the 20 sequences are of this type! It is interesting that there are no sequences containing repeats and/or hesitation items in the top 20 NS 3-word sequences and that the first sequences of this kind are *and I I* (rank 37) and *it it was* (rank 38).

Length	**NNS vs. NS speech types**	**NNS vs. NS speech tokens**
2	NNS speech >>> NS speech	NNS speech >>> NS speech
3	NNS speech >>> NS speech	NNS speech >>> NS speech
4	NNS speech >>> NS speech	NNS speech >>> NS speech
5	NNS speech >>> NS speech	NNS speech >>> NS speech
6	NNS speech > NS speech	NNS speech > NS speech

Table 6. Number of sequence types and tokens **containing repeats and/or hesitation items** in NS and NNS speech: (based on relative frequencies per 100,000 word combinations)

3.1.2 Wobbly thematic springboards

Most of the NS and NNS recurrent sequences containing repeats and/or hesitation items (+/-60%) are actually clause beginnings or 'thematic springboards' (Altenberg 1998). This is line with the assumption that clause beginnings are a major planning point (Biber et al. 1999). The actual number of these sequences is significantly higher in the NNS corpus than in the NS corpus. These results seem to suggest that, rather unsurprisingly, having to plan a clause in a language other than one's mother tongue increases the planning pressure speakers face at the beginning of a clause. Setting off on a clause is something of a challenge for learners, which leads them to use more 'wobbly thematic springboards' than native speakers. Typical examples of wobbly thematic springboards include:

(1) <B> er yeah definitely too there are a lot of activities and er well we have the C S A <?> card and with that we can have a a lot of activities for free <\B>

(2) <B> when you when you know English there you can be you can go everywhere

(3) <B> it's it's easier to learn that <?> than if you go . alone and there are always . eh men who are there and . you could dance with him but <\B>

(4) <B> we landed er there .. em but we already had a problems with em our plane because er . er the the plane was cancelled .. so: we had to change <laughs> <\B>

Interestingly, there is a larger proportion and a significantly higher number of what can be referred to as 'phrasal' sequences containing repeats and/or hesitation items such as *er the the*, *er on the*, *of the the* or *in in the*:

(5) what's your dissertation on <\A>

<B> eh I'm I'm er in Dutch literature . *er on the* em . how did <?> you say that in English em ... the tales <\B>

(6) <B> we landed er there .. em but we already had a problems with em our plane because er . *er the the* plane was cancelled .. so: we had to change <laughs> <\B>

There are six times as many such sequences (in absolute terms) in NNS than NS speech. This seems to indicate that learners appear to have more encoding problems than native speakers at phrase level, presumably either because they have problems finding the words they need to encode their messages, or simply because the difficulty of expressing themselves in a foreign language interferes with the encoding process as a whole.

A series of recurrent sequences can be seen to 'explicitly' bear witness to this difficulty: *I don't know how to say …* (▲); *I don't know how you say…* (△) and *how do you say that* (△) act as communication strategies, and more specifically as a direct 'appeal for assistance' (Tarone *et al.* 1983) to the interlocutor (who was in this case a native speaker of English):

(7) <B> it's eh . it's supposed to be a realist story a real story about . eh strange events erm . erm *OVNIs .. I don't know how to say* it in English <\B>

<A> I suppose that's U F Os [is it <\A>

<B> [yes U F Os <laughs> <\B>

(8) and there are tribes <?> and people are are so different eh they . they live *in er . in* <sighs> *I don't know how to say* this eh houses but made of . er [of wood and eh <\B>

<A> [like clay or wood yes <\A>

<B> yes . and they are so different from from the[i:]

(9) even in the United States but sometimes you . you realise that eh like for instance human rights are also em .. *how do you say that* eh <\B>

<A> violated <\A>

<B> violated in in in in Europe as well <\B>

Two learner idiosyncratic sequences (i.e. recurrent sequences that are exclusively used by the learners) containing a French word, namely *enfin I* (△) and *enfin I I* (△), are well worth mentioning here in connection with wobbly thematic springboards as they can be regarded as further evidence of learners' encoding problems when embarking on clauses. *Enfin*, which actually occurs 74 times (per 100,000 words) in the NNS corpus, is a frequently used discourse item in spoken French, which is roughly equivalent to the English *well* or *I mean*. Consider the following examples, where it can be regarded to act as a repair signal or as an anchorage point from where the learner can set off on a clause:

(10) <B> [yes .. it's i= *enfin I* find it a little bit er *bourgeois* when when you see the: . the shops and <sighs> but but er but it's really nice to live in <\B>

(11) <B> [yes yes yes yes and they they shouted in your ears and er .. yes <laughs> <\B>

<A> how strange <\A>

<B> yes but it was *enfin I* thought it was really wonderful

(12) because th= . they wanted us to react and *enfin I I* think they they really em were successful in in their . <\B>

Unlike the use of the French word *OVNI* in example (7) above, the use of *enfin* and the sequences *enfin I(I)* can arguably not be regarded as cases of code switching (a communication strategy which involves the use of a word from one's first language when experiencing lexical gaps in one's interlanguage) as their use is largely unconscious. Learners do not deliberately choose to use *enfin* or *enfin I(I)* to attempt to overcome a lexical or a grammatical problem. Because of the basic pragmatic functions they are used to fulfil in spoken discourse, their use appears to have become highly proceduralized and as a result largely unconscious. Evidence for the highly proceduralized character of *enfin* comes from cases where even very proficient advanced learners can be heard to use the odd *enfin* in unplanned spontaneous interactions.

3.2 Qualitative aspects of preferred sequences of words in NS and NNS speech

A more qualitative analysis uncovers the wide structural and functional variety of the recurrent sequences in our corpora. From the point of view of structure, a major distinction can be made between clausal sequences and phrasal sequences and within these categories between complete (It's not too bad; at the moment) and incomplete sequences (I really enjoy; a couple of). The structural variety of sequences lies outside the scope of this paper. For more details see De Cock (2003); see also Altenberg (1998) and Biber et al. (1999) for a thorough structural description of recurrent sequences in NS speech and in NS speech and writing respectively.

The functional diversity displayed by the preferred sequences in our corpora is similar to that described in Biber 2003 and Biber *et al.* 2003. The recurrent sequences in our study can broadly be classified into three main categories: referential sequences (e.g. markers of time/place: *at night, during the day, in front of*; quantifying sequences: *loads of, one of the, an awful lot of*; topic-dependent sequences: *a film*; etc.), interactional/interpersonal sequences (e.g. markers of attitudinal stance: *I really enjoyed, which is good, it was very, I'm hoping to*; markers of epistemic stance: *but I think, I don't know if, I can't remember*; responses: *yeah definitely, that's it*; markers of vagueness: *sort of, and things like that*; etc.), and discourse-organizing sequences (e.g. markers of speech/thought reporting: *so I thought, and I was like oh*; markers of contrast: *on the other hand*; makers of cause: *due to the fact* ; exemplifiers: *for example, for instance*; etc.).

For lack of space I have decided to confine this report to an in-depth discussion of one type of interactional/interpersonal sequences in NS and NNS speech, namely markers of vagueness. A functional investigation of the recurrent sequences used by the native speakers in the corpus (based on Chafe 1982, 1987; Chafe and Danielewicz 1987; Biber 1988) reveals that, on the whole, their preferred sequences are interactional and involved in nature. A very large proportion of NS preferred ways of saying things display features characteristically associated with the speaker's involvement with his/her audience and with him/herself. Many NS recurrent sequences contain response items (e.g. *yeah, oh, well*), discourse items (*you know, I mean, like*), first and second person pronouns, private verbs (*think, know, remember*) and/or are used as vagueness markers or to convey attitudinal stance or epistemic stance. Learners' preferred sequences are, on the other hand, less interactional and involved in nature than native speakers'.

Markers of vagueness are of particular interest as they are significantly underused by the learners. This is all the more significant since according to linguists such as Crystal and

Davy (1975), Aijmer (2002), Channell (1994) and Drave (2000), lack of precision is one of the most important features of informal interaction. Whereas formal situations such as debates require speakers to be explicit and precise, in informal interactions, where the emphasis is more on establishing and maintaining interpersonal contacts than conveying detailed information, speakers usually express themselves less clearly and accurately. There are two major sets of markers of vagueness in our data, namely a set of sequences that have commonly been referred to as 'vagueness tags' (Altenberg 1998), 'vague category identifiers' (Channell 1994) or 'general extenders' (Overstreet and Yule 1997), and a set of sequences containing the discourse items *sort of* and *kind of*.

3.2.1 Vagueness tags (VTs)

VTs such as *or something*, *and things* or *or anything* are indicators of intersubjectivity. They can be seen to play a significant role in informal spoken interactions on an interpersonal level: they signal an assumption of shared experience and social closeness. According to Overstreet and Yule (1997: 256) (see also Aijmer 2002), by using VTs speakers convey the following message to their interlocutors. This could be paraphrased as *More could be said but it is not explicitly said because you already know what I mean. I don't have to spell it out because you can fill in the details yourself because we share the same experience of the world and the same background knowledge*.

Consider the following examples from the NS corpus:

(13) <B> . and there the rooms are dead small and on the whole corridor you've got between ten people *or something* it's probably the same here <\B>

(14) <B> one by one the children they they don't die but you know they they get injured *or whatever* <\B>

(15) <B> that's right everybody went back home and .. we we were a bit sad and er .. we we'd exchanged addresses *and everything so* we could write to each other <\B>

(16) <B> so we're trying to get together . some sort of comedy with little sketches *and things* <\B>

(17) <B> [erm and they have a once a year they have a big award where they say which has been their best shop this year *sort of thing* <\B>

(18) the film is basically about . erm the life of William Wallis who was a: a leader erm he was erm a fairly simple person he wasn't a clansman_*or anything* <\B>

Table 7 reveals that, overall, the learners in the corpus significantly underuse VTs. Native speakers use over twice as many VTs as the learners. Learners' underuse of VTs may have a significant impact on how they are perceived by native speakers in informal situations. As stated above, vagueness or lack of precision is one of the most important characteristics of informal interaction. Learners' underuse of VTs may thus go some way towards explaining why, as was noted by Channell (1994: 21), "while grammatically, phonologically, and lexically correct, [they] may sound rather bookish and pedantic to a native speaker." It is worth pointing out that the limited number of VTs learners tend to use with high frequencies, i.e. VTs with *and so on* and *et cetera*, have been found to be mainly used in formal talk (Overstreet and Yule 1997), which only adds to the impression of detachment and formality they may well give in informal situations. Learners' use of the recurrent sequences *for*

example and *for instance* can also be seen to contribute to this impression of formality. These two sequences, which are not highly recurrent in NS speech (*for example* NNS speech 61 vs. NS speech 7; *for instance* NNS speech 47 vs. NS speech 2, these are relative frequencies per 100,000 word sequences), can actually be regarded as more typical of writing.

Vagueness tags	**NNS**	**NS**
(and er) and so on	33	(2)
and so on and so on	5	0
and so on and/and er/it was	21	0
et cetera	24	(2)
(or) something like that	29	22
or something (and)	14	57
(and) things like that (and/so/but)	18	52
or things like that	4	(1)
and so forth and so	3	0
or whatever	(9)	21
and everything (and/so)	(3)	45
all that kind of thing	0	3
(and) that kind of thing	4	9
and things (and)	(1)	52
and stuff	0	34
and stuff like that (and)	0	17
(and) that sort of thing	0	10
sort of thing (but/so)	0	33
or anything	0	25
and places like that	0	4
all the rest of it	0	3
Total	168 ***	392

Table 7. Vagueness tags in NNS vs. NS speech (absolute frequencies)[3]

Learners' use of the VTs or *something (like that)* is a good illustration of learner idiosyncratic misuse of a target language sequence. According to Channell (1994), the VTs *or something(like that)* and *or anything(like that)* are found in complementary distribution between assertive and non-assertive contexts. As is shown in Table 7, learners do not use the VT *or anything* at all. What is more, the NNS corpus includes a few instances of the VT *or something (like that)* in inappropriate, i.e. non-assertive, contexts:

(19) <B> yes yes and er not er . **I don't like** er . novels *or something like that* .. mm only magazines newspapers <\B>

(20) <B> .. yes I think . er .. **I don't er I don't come here** er .. for parties *or something like that* I prefer to go . to go and visit friends and I don't like to: to to stay .. to stay up all night or something like that <\B>

(20) <B> **I don't want to work** in a bank *or something* <X> I'm not interested in that <\B>

3.2.2 Sequences containing *sort of* and *kind of*

The sequences that contain *sort of* or *kind of* are ubiquitous in NS speech. All in all, there are over 460 instances of *sort of* and over 110 instances of *kind of* in 100,000 words in the corpus (instances of *sort of* and *kind of* occurring as part of VTs have been excluded from the frequency counts). The prevalence of sequences with *sort of* over sequences with *kind of* comes as no surprise in view of the fact that while the former is more typical in British English (the interviewees are all British), the latter is more frequently used in American English (see for example Biber *et al.* 1999). In the majority of cases *sort of* (in about 400 instances out of 460, i.e. 87%) and *kind of* (in about 80 instances out of 110, i.e. 72%) are used as a discourse item that introduces vagueness and fuzziness in discourse. Instances where *sort of* and *kind of* are not used as discourse items and where they can be paraphrased as 'type of' (if X is a sort of/kind of Y it means that X can be a hyponym of Y, Aijmer 2002: 176) typically occur as part of sequences such as *a sort of, that sort of* or *the sort of*:

(21) <B> erm so .. maybe *that sort of* area . but involving films so .. you know <XX> film magazines <\B>

(22) <B>I thought I think I'm *the sort of* person that can do this I was really confident [then I went in .. and <\B>

An interesting pattern that comes out of an analysis of sequences with *sort of* and *kind of* that can be seen to function as discourse items is the use of these sequences in front of verbs in over 35% of the cases (cf. examples below). Clausal sequences such as *I sort of* clearly reflect this tendency.

(24) probably speech therapist things like that and <X> none of <X> interest me both my parents are teachers they've *sort of* put me off <XX[X> <\B>

(25) <B> <laughs> <X> didn't really we just saw the pubs . and we didn't really get out into it the second time we *sort of* went round all the palaces <\B>

(26) I think it's the fact that I had a bossy German with me that *sort of* helped as well <\B>

(27) <B> yeah . well I was going to but you know a= as I say during the year I *kind of* changed my mind about what I wanted to do <\B>

Just like VTs, sequences with *sort of* and *kind of* signal an assumption of common ground and social closeness, which in turn contributes to the informality of the interaction and "creates a congenial atmosphere" (Aijmer 2002: 209). More specifically, speakers can be seen to use these discourse items to signal to their addressees that the word they are about to use may not be the perfect word for what they want to express or describe, either because they lack the vocabulary to talk about a particular topic or because the word may be too technical, formal or informal for example (e.g. *it was it was in a big sort of chateau villa thing so*). In addition *sort of* and *kind of* have softening and polite functions when they are used to tone down strong opinions or unpleasant or embarrassing topics or referents (*which wa= wasn't the biggest problem but I mean coping in the first I remember in the first week I just went around in a sort of daze because having to find out where everything was getting used to university life*).

A comparison of the sequences containing *sort of* and *kind of* (cf. Table 8) reveals that the learners in the spoken corpus tend to significantly underuse these sequences overall: *sort*

of NNS speech 85 vs. NS speech 460; *kind of* NNS speech 41 vs. NS speech 110 (relative frequencies)

	NNS	**NS**
kind of [4]	▲▲	▲▲▲
a kind of	▲	△
this kind of	▲	-
that kind of	△	-
some kind of	△	△
sort of [4]	▲	▲▲▲▲▲
a sort of	▲	▲
some sort of	▲	-
I sort of	-	▲
it's sort of	-	▲
it was sort of	-	△
you can sort of	-	△
I was sort of	-	△

Table 8. Examples of sequences with *sort of* and *kind of* in NNS vs. NS speech

What is even more significant is the way the learners use *sort of* and *kind of*. Unlike the native speakers in the spoken corpus, who can regularly be seen to use *sort of* and *kind of* in front of verbs, the learners tend to use them almost exclusively in front of nouns. There are a mere 5 cases of *sort of/kind of* followed by a verb in NNS speech, e.g.:

(28) but er the[i:] other was really bad and he was er <\B> <A> oh I see <\A> <B> *sort of* . destroying all the[i:] effect the good one was doing and em.

Most instances of *sort of* and *kind of* occur as part of patterns more typically associated with the literal non-pragmatic use of *sort of* (when *sort of/kind of = type of*, cf. Aijmer 2002):

(29) <B> it's *a sort of* drum [but er <\B>

(30) <B> yes *the kind of* English they have to deal with is er well scientific English [eh .. well <sighs> agri= agriculture and things around around this <\B>

(31) <B> mm I don't think so mm I mean no it's *the kind of* mentality which is developed in: the country so er and particularly in second= er at secondary school and this er *this kind of* mentality er we are really mm we

(32) <B> yes we do because there are loads of trolleys . and er every *kind of* [food going ar= around er well <\B>

Sort of and *kind of* can actually be seen to perform typical NS pragmatic functions in approximately 15-17% of the cases (vs. ca. 80% in NS speech!).

Sort of and *kind of* are sometimes used by the learners as a communication strategy to bridge gaps in their English vocabulary. This can, to some extent, be related to some of native speakers' uses of *sort of* and *kind of*, i.e. when using the sequences to signal to their interlocutors that the word they are going to use may not be the perfect word to denote what they have in mind because they lack the vocabulary to talk about a certain topic for example. Learners' use of *sort of* or *kind of* as a communication strategy involves

'approximation' and 'language switch' (cf. Tarone *et al.* 1983). In most cases the word following *sort of* or *kind of* is borrowed from French. Consider the following examples:

(33) <B> there was there were clowns everywhere there were erm .. *some sort of* ***braderie*** but .. <\B>

(34) <B> yes yes .. yeah because of the: *sort of* eh ***vapeur*** [I don't know how you say it <\B>

(35) <B> something really serious you know [er *a kind of* er .. ***tailleur*** <\B>

(36) and there there are also eh parks around er around there and a: . *a kind of* eh ***téléphérique*** I don't know how you say it in English <laughs>

(37) it was very beautiful it was *a sort of* .. erm .. yes erm ***promenade***

In the following examples *sort of* and *kind of*, which are followed by an English word, can be seen to be used in the same way. Note the presence of quite a few pauses and hesitation items, which clearly point to learners' encoding problems. The word *essay* or *assignment* appears to be problematic for several interviewees:

(38) <B> er yes we had .. erm to: to revise your courses or to: to make er .. *some sort of* er <?> little er <?> .. er <?> .. little .. erm <X> .. such a: .. [little work <X> work <\B>

(39) <B> but we have er . especially one course we have to make erm . *a kind of* work . for the[i:] end of the year <\B>

Learner's preferred use of *kind of* over *sort of* could be regarded as resulting from their essentially non-pragmatic use of these strings (*kind of* is more typically associated with 'type') and from possible influence from American English through films, sitcoms or songs (the discourse item *kind of* has been reported as more frequent in US English than in British English, cf. Biber *et al.* 1999).

Two other phrases which fulfil similar functions to *sort of* and *kind of* in the NS corpus and which are not part of learners' stocks of preferred sequences are *in a way* (▲ e.g. *not on the kibbutz because . it's very it was very small very isolated it was like a prison <u>in a way</u> you can't really leave*) and *a bit of a*. The sequence *a bit of a* (▲) is typically used to soften or tone down experiences or situations (cf. Channell 1994), either because they might be perceived as negative by the hearer (most examples are of this type) or because the speaker may come across as pretentious (cf. example 42). Consider the following examples of *a bit of a* in context:

(40) <B> and er Bangkok was *a bit of a* **disappointment** <\B>

(41) <B> [and I got on really well with them .. and er <X> a nice house <X> it's always *a bit of a* **tip** but . but you know I'm I'm really enjoying it <\B>

(42) you can control them I think it's more . fun to: to have something [to work with <\B>

<A> [uhu <\A>

<B> bit of *a bit of a* **challenge** anyway [<laughs> <\B>

It is noteworthy that incursions into the Chinese, Italian and Japanese LINDSEI subcorpora (the only other complete LINDSEI components when this study was carried out) show that these learners also tend to markedly underuse markers of vagueness in informal

speech. For example, *sort of,* which is recurrent in the Italian subcorpus (▲▲) only, is almost invariably used in the patterns *a sort of* and *this sort of* (in a way that is quite close to *type of*). Interestingly, *kind of* is fairly widely used in the three subcorpora (Japanese ▲▲▲▲; Chinese ▲▲▲▲, Italian ▲▲▲ - possibly due to influence of American English). However, here again, it predominantly occurs as part of the patterns *a kind of, this kind of* or *some kind of* and is closer in meaning to *type of* than its pragmatic use:

(43) <B> so I I I can only get used to *this kind of* dish <\B> (Chinese)

(44) to kiss him but Kevin Spacey say no er look mm I'm not *this kind of* person I understand but mm I'm not gay in the [i:] end (Italian)

(45) I want to . be a teacher of *some kind of* team sports club [ha-ha <\B> (Japanese)

Spot checks on *or something (like that)*, *or anything (like that)*, *and things (like that)* seem to indicate that vagueness tags are probably also generally underused by the learners in the other spoken learner corpora: while *or something (like that)* is used in all three corpora (Japanese: ▲▲, Chinese: ▲, Italian: ▲▲), *or anything (like that),* and *and things (like that)* appear to occur (admittedly with very low frequencies: 1 and 2 respectively!) in the Italian corpus only. Note that the more formal *and so on* seems to be particularly preferred in the Italian corpus (▲▲, cf. French learners). It is also interesting to note that *for example* (but not *for instance*) tends to be favoured by the learners in the three corpora (Chinese: ▲▲; Japanese ▲▲▲, Italian: ▲▲▲).

3.2.3 Of course!

Interestingly, beside significantly underusing markers of vagueness and sequences containing discourse items such as you know and I mean, the French-speaking learners in the NNS corpus seem to favour some rather forceful recurrent sequences. The sequences that contain of course are a case in point.

The sequence *of course,* whether or not occurring as part of longer recurrent sequences (e.g. *yes of course*, *well of course*), is significantly overused in NNS speech: NNS ▲▲▲▲▲ vs. NS speech ▲▲ (at $p \leq 0.005$). Learners' use of the response *yes/yeah of course (yes)* deserves special attention. Not only do the learners in the corpus overuse this sequences (NNS ▲▲ vs. NS -), but they also tend to misuse it. Examples (46) to (48) illustrate learners' typical misuse of the target language sequence.

(46) <B> it's a factor of motivation for the students <\B>

<A> yes and I suppose also they are the ones that are in control on a computer <\A>

<B> *yes of course* <\B>

(47) <B> I'm working on er Robinson Crusoe's rewritings <\B>

<A> oh yes <\A>

<B> yeah it's fascinating <\B>

<A> how many times has it been rewritten .. has it it's been rewritten? <\A>

<B> er yeah *yeah of course* <\B>

(48) <A> I've heard about this problem in Dublin as well that they can't study literature at all <\A>

<B> mm <\B>

<A> because all the courses are full it's a shame that isn't it? <\A>

<B> *yeah of course* <\B>

Using *yes/yeah of course* in this way to answer a request for information or to respond to an opinion expressed by another speaker may well make learners sound rather over-emphatic and even impolite.

It is interesting to note that two of the major learners' dictionaries, namely the *Longman Dictionary of Contemporary English* (LDOCE 2001: 980) and the *Oxford Advanced Learners' Dictionary* (OALD 2000: 287) actually address the inappropriate use of *(yes) of course* in such contexts. In the usage notes provided for *of course*, LDOCE and OALD respectively stress that using the sequences as a reply to a request for information "would sound as if you think the answer to the question is very clear and you think the person is stupid to need to ask you" and that "it may sound as though you think the answer to the question is obvious and that the person should not ask." OALD even supplies learners with appropriate alternative ways of reacting and responding (e.g. *yes it is*). Whether or not and the extent to which these usage notes were compiled with the help of studies based on learner corpora is unfortunately not clear (OALD makes no mention of the use of learner corpus data and the Longman Learners' Corpus appears to contain only written language while the uses discussed in the note mainly concern spoken language). Results from studies of recurrent sequences in learner language would certainly have a crucial part to play in the compilation of usage notes of this type.

An incursion into the Japanese, Chinese and Italian LINDSEI subcorpora shows that it is not just French learners who appear to favour the sequences *(yes/yeah) of course*: Chinese ▲▲▲, Japanese ▲▲▲▲, Italian ▲▲▲.

Learners' inappropriate use of *yes/yeah of course* and their underuse of the response *that's right* could be partly related. In examples (49) and (50) *(yes) that's right* could arguably be used as a preferred and more appropriate substitute for the awkward *yes of course*. Compare examples (49) and (50) with examples of native speakers' use of *(yes) that's right* in (51), (52) and (53).

(49) <B> no today er I've got er . nothing special er neither on on Friday er so m= Monday and Friday are . day off [days off <\B>

<A> [easy . easy days <\A>

<B> easy day <\B>

<A> so you just come and see me instead <\A>

<B> [*yes of course* <\B>

(50) <A> you presumably came here with other people from Namur <\A>

<B> er *yes of course* and er .. and I I've got a: . well er not a flat I I don't know <\B>

(51) <A> so you were you were studying in Copenhagen <\A>

<B> *yes that's right yeah* <\B>

(52) <A> <laughs> yes but as you say it's the people [who make the place mhm <\A>

<B> [*yeah that's right* <XX> yeah definitely <\B>

(53) <A> mhm but erm so it was a holiday for you there <\A>

<B> *yeah that's right yeah* <\B>

4 Conclusion

Kjellmer's assumption that learners' building material is individual bricks rather than prefabricated sections appears to be simplistic. This hypothesis is confirmed provided that only those sequences that do not contain repeats and/or hesitation items are taken into consideration (see also the results for NS and NNS writing in De Cock 2003). What emerges from this study is that advanced learners' use of frequently recurring sequences of words displays a complex picture of overuse, underuse, misuse of target language NS sequences and use of learner idiosyncratic sequences. A number of recurrent sequences of words also point to learners' acute encoding problems. The findings also suggest that the learners are lacking in routinized ways of interacting and building rapport with their interlocutors and of toning down and weaving the right amount of imprecision and vagueness (a typical feature of NS informal interactions).

Studies of recurrent sequences in NS and NNS speech and writing undoubtedly have a very valuable contribution to make to pedagogical lexicography (enhancing the information included in learners' dictionaries) and to English Language Teaching (ELT). Although there is now widespread agreement that prefabs of all kinds should be taught (through awareness-raising and/or explicit teaching activities), recent studies (e.g. Nesselauf 2003 and forthcoming, Schmitt et al. 2004, Jones and Haywood 2004) have highlighted the difficulty of actually choosing the prefabs that should be included in ELT material. The results of investigations such as the one reported on here have a major part to play in this selection because, not only do they provide us with real NS usage, but they also bring to light the sequences learners appear to find problematic.

This paper has mainly focused on the spoken productions of French-speaking advanced learners and there have been only extremely limited incursions into learner corpora of other mother tongue backgrounds. As Granger (1998b) puts it, learners are not "phraseologically virgin territory". Large-scale cross-linguistic investigations are therefore called for to assess the role played by transfer from the mother tongue, transfer of training and developmental processes in the development of learners' preferred ways of saying things. The multi learner mother tongue background composition of learner corpora such as LINDSEI and ICLE makes it possible for researchers to uncover which prefabs tend to be problematic for different groups of learners regardless of their mother tongue backgrounds but also which prefabs are problematic for specific groups only (e.g. transfer-related deficiencies), which will in turn influence the design of ELT material (textbooks/methods aimed at all learners or at learners from the same mother tongue background). Such studies are crucial because, as Nesselhauf's study of verb-noun collocations suggests, the influence of learners' mother tongue "in the area of word combinations (…) seems to be considerably stronger than even those researchers who have suspected its importance have assumed" (Nesselhauf 2003: 237; see also Spöttl and McCarthy 2004).

An important issue connected with studies of recurrent sequences of words will need to be dealt with in greater detail in the near future. This issue, which was raised by a limited though exploratory and pioneering study by Schmitt, Grandage and Adolphs (2004), concerns the psycholinguistic validity of automatically extracted recurrent sequences of words and the relationship between recurrence and the storage of sequences of all kinds as wholes in the brain (see also Wray 2002). In other words, are recurrent sequences of words

actually stored in the mind as wholes and does recurrence actually cause a sequence to be stored as a unit or is recurrence a result of a sequence being stored whole and therefore easily accessible? This is still an open question.

Notes

1. The other three components of communicative competence include 'Whether (and to what degree) something is formally possible', 'Whether (and to what degree) something is feasible', 'Whether (and to what degree) something is appropriate' (Hymes 1972: 284-285).

2. Legend: >>> means statistically highly significant overuse (at $p \leq 0.005$); >> means statistically significant overuse (at $p \leq 0.01$); > means statistically significant overuse (at $p \leq 0.05$); <<< means statistically highly significant underuse (at $p \leq 0.005$); << means statistically significant underuse (at $p \leq 0.01$); < means statistically significant underuse (at $p \leq 0.05$). = is used when the difference is not statistically significant.

3. The asterisked figure indicates a statistically highly significant difference (chi-square with $p <$ or $= 0.005$).

4. The frequency count given for this sequence does not include instances of the string when it occurs as part of longer recurrent sequences in the corpus.

References

Adolphs, S. and V. Durow (2004) Social-cultural integration and the development of formulaic sequences. In N. Schmitt (ed.) *Formulaic sequences*. Amsterdam: Benjamins. 107-126.

Aijmer, K. (2002) *English discourse particles: Evidence from a corpus*. Amsterdam: Benjamins.

Altenberg, B. (1990) Speech as linear composition. In G. Caie, K. Haastrup, A.L. Jakobsen, J.E. Nielsen, J. Sevaldsen, H. Specht and A. Zettersten (eds.) *Proceedings from the fourth Nordic conference for English studies*, Helsingor, May 11-13 1989. Copenhagen University, Department of English. 133-143.

Altenberg, B. (1998) On the phraseology of spoken English: The evidence of recurrent word combinations. In A. P. Cowie (ed.) *Phraseology: Theory, analysis and applications*. Oxford: Oxford University Press. 101-122.

Altenberg, B. and M. Eeg-Olofsson (1990) Phraseology in spoken English: Presentation of a project. In J. Aarts and W. Meijs (eds.) *Theory and practice in corpus linguistics*. Amsterdam: Rodopi. 1-26.

Béjoint, H. (2000) *Modern lexicograph: . An introduction*. Oxford: Oxford University Press.

Biber, D. (1988) *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (2003) 'Take a look at…': Lexical bundles in university teaching and textbooks. Paper presented at PALC 2003 (Practical Applications in Language and Computers), Łódź University, 4-6 April 2003.

Biber, D. and S. Conrad (1999) Lexical Bundles in Conversation and Academic Prose. In H. Hasselgård and S. Oksefjell (eds.) *Out of Corpora: Studies in Honour of Stig Johansson.* Amsterdam and Atlanta: Rodopi. 181-190.

Biber, D., Conrad, S. and V. Cortes (2003) Towards a taxonomy of lexical bundles in speech and writing. In A. Wilson, P. Rayson, and T. McEnery (eds.) *Corpus linguistics by the Lune: A festschrift for Geoffrey Leech.* Frankfurt: Peter Lang. 71-92.

Biber, D., Johansson, S., Leech, G., Conrad, S. and E. Finegan. (1999) *Longman grammar of spoken and written English*. London: Longman.

Chafe, W. L. (1982) Integration and involvement in speaking, writing, and oral literature. In D. Tannen (ed.) *Spoken and written language: Exploring orality and literacy*. Norwood: Ablex. 35-54.

Chafe, W. L. (1987) Cognitive constraints on information flow. In R S. Tomlin (ed.) *Coherence and grounding discourse*. Outcome of a symposium, June 1984, Eugene, Oregon. John Benjamins: Amsterdam Philadelphia, pp. 21-51.

Chafe,W. L. and J. Danielewicz (1987) Properties of spoken and written language. In R. Horowitz and S. J. Samuels (eds.) *Comprehending oral and written language*. New York: Academic press. 82-113.

Channell, J. (1994) *Vague Language*. Oxford: Oxford University Press.

Cortes, V. (2002a) *Lexical bundles in published and student academic writing*. Paper presented at AAACL 2002. Fourth North American Symposium on Corpus Linguistics and Language Teaching, Indianapolis, 2-4 November 2002.

Cortes, V. (2002b) Lexical bundles in Freshman composition. In R. Reppen, S. M. Fitzmaurice and D. Biber (eds.) *Using Corpora to Explore Linguistic Variation*. Amsterdam: Benjamins. 131-146.

Cowie, A. P. (1999) Phraseology and corpora: Some implications for dictionary-making. *International Journal of lexicography* 12.4: 307-323.

Crystal, D. and D. Davy (1975) *Advanced conversational English*. London: Longman.

De Cock, S. (2003) *Recurrent sequences of words in native speaker and advanced learner spoken and written English*. Unpublished PhD dissertation. Louvain-la-Neuve: Centre for English Corpus Linguistics: Université catholique de Louvain.

Drave, N. (2000) Vaguely speaking: A corpus approach to vague language in intercultural conversations. Paper presented at Icame 2000, Sydney, 21-25 April 2000.

Fernando, C. (1996) *Idioms and idiomaticity*. Oxford: Oxford University Press.

Granger, S. (1998a) The computerized learner corpus: A versatile new source of data for SLA research. In S. Granger (ed.) *Learner English on computer*. London: Longman. 3-18.

Granger, S. (1998b) Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (ed.) *Phraseology: Theory, analysis and applications*. Oxford: Oxford University Press. 145-160.

Hymes, D. (1972) On communicative competence. In J. Pride and J. Holmes (eds.) *Sociolinguistics*. Harmondsworth: Penguin. 269-293.

Jones, M. and S. Haywood (2004) Facilitating the acquisition of formulaic sequences: An exploratory study in an EAP context. In N. Schmitt (ed.) *Formulaic sequences*. Amsterdam: Benjamins. 269-300.

Kjellmer, G. (1991) A mint of phrases. In K. Aijmer and B. Altenberg (eds.) *English corpus linguistics*. London: Longman. 111-127.

Kjellmer, G. (1994) *A dictionary of English collocations*. 3 Vols. Oxford: Clarendon Press.

Milton, J. and R. Freeman (1996) Lexical variation in the writing of Chinese learners of English. In C. E. Percy, C. F. Meyer and I. Lancashire (eds.) *Synchronic corpus linguistics. Papers from the sixteenth international conference on English language research on computerized corpora (ICAME 16)*. Amsterdam: Rodopi. 121-131.

Moon, R. (1998) *Fixed expressions and idioms in English*. Oxford: Clarendon Press.

Nesselhauf, N. (forthcoming) *Collocations in a learner corpus*. Amsterdam: Benjamins.

Nesselhauf, N. (2003) The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics* 24.2: 223-242.

Overstreet, M. and G. Yule (1997) On being explicit and stuff in contemporary American English. *Journal of English Linguistics* 25.3: 250-258.

Pawley, A. and F. H. Syder (1983) Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards and R. Schmidt (eds.) *Language and communication.* London: Longman. 191-226.

Raupach, M. (1984) Formulae in second language speech production. In H. Dechert, D. Mohle and M. Raupach (eds.) *Second language productions*. Tuebingen: Gunter Narr. 115-137.

Schmitt, N. and R. Carter (2004) Formulaic sequences in action: An introduction. In N. Schmitt (ed.) *Formulaic sequences*. Amsterdam: Benjamins. 1-22.

Schmitt, N., Dörnyei, Z., Adolphs, S. and V. Durow (2004) Knowledge and acquisition of formulaic sequences: A longitudinal study. In N. Schmitt (ed.) *Formulaic sequences*. Amsterdam: Benjamins. 55-86.

Schmitt, N., Grandage, S. and S. Adolphs (2004) Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (ed.) *Formulaic sequences*. Amsterdam: Benjamins. 127-151.

Sinclair, J. M. (1991) *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Spöttl, C. and M. McCarthy (2004) Comparing knowledge of formulaic sequences across L1, L2, L3 and L4. In N. Schmitt (ed.) *Formulaic sequences*. Amsterdam: Benjamins. 191-225.

Stubbs, M. (2002) *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

Sugiura, M. (2002) Collocational Knowledge of L2 Learners of English: A case study of Japanese Learners. In T. Saito, J. Nakasnura and S. Yamazaki (eds.) *English Corpus Linguistics in Japan*. Amsterdam: Rodopi. 303-323.

Tarone, E., Cohen, A. and G. Dumas (1983) A Closer look at some interlanguage terminology: A framework for communication strategies. In C. Faerch and G. Kasper (eds.) *Strategies in interlanguage communication.* London: Longman. 4-14.

Wray, A. (2002) *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Oxford Advanced Learner's Dictionary (2000) Oxford: Oxford University Press. X. S. Wehmeier (ed.).

Longman Dictionary of Contemporary English (2001) Harlow: Pearson. D. Summers (ed.).