Pragmatic Prefabs in Learners' Dictionaries

Sylvie De Cock
Centre for English Corpus Linguistics, Université catholique de Louvain
Collège Erasme, 1, Place Blaise Pascal, B-1348 Louvain-la-Neuve, Belgium
decock@lige.ucl.ac.be

Abstract

This paper sets out to investigate the coverage of one type of prefabricated expressions, namely pragmatic prefabs in learners' dictionaries. The first section seeks to find out whether or not a series of frequently recurring pragmatic prefabs in the Louvain Corpus of Native English Conversation (LOCNEC) are included in the latest editions of the five major advanced learners' dictionaries (CIDE 1995, COBUILD 2001, LDOCE 2001, OALD 2000, MED 2002) and raises some issues concerning the inclusion or non-inclusion of such expressions. The second section examines the description of *of course* in greater detail both from a quantitative (i.e. in terms of the number of uses distinguished and the number of examples given) and a qualitative point of view (i.e. in terms of the adequacy of the descriptions provided). The five dictionaries are compared and assessed and the contribution of learner corpus-based research to the compilation of learners' dictionaries is highlighted. The paper concludes by making concrete suggestions for more 'learner-aware' dictionary coverage of pragmatic prefabs.

1 Introduction

Until about two or three decades ago phraseology, taken in a wide sense to refer both to what Gvishiani et al. [2001] call idiomatic and non-idiomatic phraseology, was the poor relation of linguistic investigation in the West and ready-made or prefabricated multi-word expressions were regarded as marginal and rather 'problematic' phenomena [Chafe 1968]. Since then, the study of prefabs has moved to the forefront of linguistic research and has attracted growing interest from a variety of language-related fields such as language teaching and second language acquisition, NLP and corpus linguistics, pragmatics, psycholinguistics and lexicography, to list but a few. The sheer number and range of recent publications devoted to multi-word expressions of all kinds can be seen to bear witness to the present topical and even fashionable character of the study of these expressions.

This paper sets out to investigate the lexicographic treatment of prefabricated expressions in the latest editions of the five major advanced learners' dictionaries: the Cambridge International Dictionary of English 1995 (first edition, henceforth CIDE), the Collins Cobuild English Dictionary for Advanced Learners 2001 (third edition, henceforth COBUILD), the Longman Dictionary of Contemporary English 2001 (third edition, henceforth LDOCE), the MacMillan English Dictionary for Advanced Learners of English 2002 (first edition, henceforth MED) and the Oxford Advanced Learner's Dictionary of Current English 2000 (sixth edition, henceforth OALD). As becomes apparent from the introductory sections to the five learners' dictionaries under investigation a conscious effort has been made to include and describe multi-word expressions. A brief survey of the literature on the coverage of prefabs in learners' dictionaries reveals that idioms and restricted collocations with figurative meanings (e.g. *spill the beans* or *foot the bill*) seem to

have received most of the attention and, that, although the treatment of these expressions has benefited a great deal from the use of corpora in the dictionary-making process, there is still room for improvement [Herbst 1996; Moon 1996; Mittmann 1996; Alexander 1996; Svartvik 1996; Bogaarts 1996]. The focus of this study is on one type of prefabs which are not very salient physiologically and which seem to have been comparatively neglected in lexicographic studies: 'pragmatic prefabs' or 'formulae' used in spoken interaction.

2 Pragmatic Prefabs in Learners' Dictionaries

The starting point of this study is an analysis of recurrent word combinations [De Cock 2000] in the LOuvain Corpus of Native English Conversation (LOCNEC) and the French component of the Louvain International Database of Spoken English Interlanguage (LINDSEI). LOCNEC and the French component of LINDSEI[1] are two fully comparable corpora: they are made up of informal interviews with British university students and advanced French learners of English as a foreign language (university students of English in their third or fourth year) respectively and contain around 100,000 words of interviewee speech each. The analysis reveals that among the frequently recurring continuous sequences of words in the corpora there is a whole series of prefabs that can be labelled 'pragmatic prefabs'. Pragmatic prefabs or formulae (e.g. *you know*, *thank you*, *in a nutshell*) are not defined in terms of the segments of extra linguistic experience they designate but in terms of what language users 'do' with them. In other words, even if pragmatic prefabs contribute little, if anything, to the propositional content of speakers' utterances, they nevertheless play a major role in spoken discourse in that they are used to perform important functions on an interpersonal level (to signal turn management, speakers' attitudes towards their utterances and interlocutors, to perform speech-act functions), on a textual level (to signal information management: topic shifts, digressions, return to topics after digressions, exemplifiers, summarizers) and/or on a strategic level (to signal planning/encoding problems).

2.1 The inclusion of pragmatic prefabs in the five learners' dictionaries

As Table 1 shows, the pragmatic prefabs under investigation[2] are not given equal coverage in the five dictionaries. Coverage ranges from 8 prefabs included and described in a subentry (CIDE) to 19 (LDOCE) and reflects to a large extent editorial decisions and dictionary policies. LDOCE's extensive coverage of the formulae in this study appears to be in line with the claim that LDOCE, which is based on substantial corpora of not only written but also spoken American (The Longman Spoken American Corpus) and British English (the spoken component of the BNC), gives more prominence to spoken English than any other ELT dictionary (quite a few of the formulae in Table 1 can actually be labelled as more frequent in speech than in writing) and that is lays special emphasis on phrases and collocations. The apparent poor score of CIDE, which, its makers claim, is built around a large corpus of both written and spoken English, is largely due to the fact that, as was also noted by Herbst [1996], the examples in CIDE are not merely seen as illustrations of how the words are used but are designed to provide its users with information on collocations and phrases. Ten out of the 21 prefabs are actually presented in this rather low-key way. Only 4 of them are highlighted in bold type and 3 are followed by a synonym in brackets (*in a way*, *for instance*, *and then*). COBUILD, MED and OALD have similar scores but it should be noted that OALD also uses bold type in examples with extra explanations in brackets to highlight common phrases.

	Relative frequency (based on 100,000 words per variety)		CIDE 1995	COBUILD 2001	LODCE 2001	MED 2002	OALD 2000
	LOCNEC	LINDSEI					
You know	503	198[-***][2]	✔	✔	✔	✔	✔
Sort of	406	38[-***]	✔	✔	✔	✔	✔
Sort of like	41	0	✘	✘	✔	✘	✔
Kind of	79	48[-***]	✔	✔	✔	✔	✔
I mean	370	160[-***]	✔	✔	✔	✔	✔
I thought	100	53[-***]	Ex.	Ex.	✔	✔	Ex.
And then	293	196[-***]	Ex.	✔	Ex.	Ex.	✘
But then	25	11[-***]	**Ex.**	✔	✔	✔	✔
That's right	70	23[-***]	Ex.	✘	✔	✔	**Ex.**
I suppose	55	25[-***]	Ex.	✔	✔	✔	**Ex.**
I know	33	4	✔	✔	✔	✔	✔
In a way	16	3	**Ex.**	✔	✔	✔	✔
A bit of a	16	0	Ex.	Ex.	✔	✔	✔
And like	48	0	✘	Ex.	✘	✘	✘
And things (like that)	71	10[-***]	✘	✘	✔	✘	✔
I think	357	439[+***]	Ex.	✔	✔	Ex.	Ex.
In fact	5	225[+***]	✔	✔	✔	✔	✔
Of course	26	128[+***]	✔	✔	✔	✔	✔
And so on	2	66	✔	✔	✔	✔	✔
For example	7	61[+***]	**Ex.**	✔	✔	✔	✔
For instance	2	47	**Ex.**	✔	✔	✔	✔
Total			8 ✔ 3 ✘ 6 Ex. 4 **Ex.**	15 ✔ 3 ✘ 3 Ex.	19 ✔ 1 ✘ 1 Ex.	16 ✔ 3 ✘ 2 Ex.	15 ✔ 2 ✘ 2 Ex. 2 **Ex.**

Table 1: Pragmatic prefabs in LOCNEC and LINDSEI and the five learners' dictionaries

Legend: ✔ included and described in a subentry
✘ not included
Ex. In an example sentence (not discussed in a separate subentry)
Ex. In bold in an example sentence (sometimes followed by a word of explanation in brackets)

While there is widespread agreement on the inclusion of 8 pragmatic prefabs (*you know, sort of, kind of, I mean, I know, in fact, of course, and so on*) and relative agreement on the inclusion of 10 others (*for example, for instance, in a way, but then, I suppose, a bit of a, I think, I thought, that's right, and then*), there is no consensus on the treatment of *sort of like, and things (like that)* and especially *and like.*

It is noteworthy that, whereas the frequently recurring formulae listed are not all included in the dictionaries, 'classical' idioms such as *kick the bucket*, which have been reported to be rather rare even in large corpora (Moon 1998), appear to receive systematic treatment. This should not come as a surprise as these figurative and hence psychologically salient expressions have until fairly recently received the lion's share of attention in phraseology mainly because of their non-compositional meanings. Pragmatic prefabs are not on the whole semantically opaque. This does however not mean that they are easy to decode or that learners of English experience no problems when decoding and/or encoding them. Pragmatic prefabs are after all not defined in terms of the propositional meanings they convey but in terms of the pragmatic functions they are used to perform in discourse. Because of their essential interpersonal and discourse-organising functions, pragmatic prefabs are part and parcel of natural native-like English and as such require adequate treatment in advanced learners' dictionaries.

2.2 Issues regarding the inclusion of pragmatic prefabs in learners' dictionaries

Given the uneven treatment of some fairly frequent formulae in LOCNEC in most of the 5 dictionaries, I set out to check the frequencies of the 21 pragmatic prefabs under study in two larger native speaker (NS) corpora of spontaneous conversation, namely part of the demographic component of the British National Corpus (approximately 2,800,000 words, henceforth the BNC) and the private dialogue component of ICE-GB (approximately 205,000 words, henceforth ICE-GB). Table 2 displays the 21 pragmatic prefabs ranked in order of frequency in LOCNEC, the BNC, ICE-GB and in the French LINDSEI subcorpus. The table clearly shows that there are considerable variations in frequency between the various NS corpora. That said, the ranking is on the whole fairly similar with 8 of the top 10 and 3 of the bottom 5 formulae shared by the speakers in the three NS corpora. In spite of some differences in frequency between the NS corpora, a comparison of the frequencies of the formulae in LINDSEI with their frequencies in the BNC and ICE-GB nevertheless uncovers very similar patterns of over- and underuse to the ones exposed when using LOCNEC as the control corpus (cf. Table 1). The French learners in LINDSEI use the pragmatic prefabs *in fact*, *of course*, *I think*, *and so on*, *for example* and *for instance* significantly much more frequently and the prefabs *you know*, *I mean*, *sort of*, *that's right*, *I thought* and *I know* significantly much less frequently than the native speakers in the BNC, ICE-GB and LOCNEC.

Differences in the frequency of occurrence of pragmatic prefabs can be ascribed to a whole series of factors such as communication situation, age or individual usage variations to mention but a few. Unlike the BNC and ICE-GB, which are made up of spontaneous conversations, LOCNEC contains informal interviews. The significant overuse of *you know*, *and then*, *and things (like that)* by the native speakers in LOCNEC (compared with the native speakers in the other two NS corpora) may well stem from the fact that, as they are expected to carry on talking for longer than in normal conversations, they feel the need to use more of these expressions to keep them going as it were. As interviewees they are also less likely to use response items that are used to empathise with one's interlocutor. This may go some way towards explaining the underuse of *I know* in LOCNEC. The influence of the communication situation on the frequency of occurrence of pragmatic prefabs can further be

illustrated by the results Table 3. This table lists a series of frequently recurring formulae (based on a study of recurrent word combinations in the demographic component of the BNC sampler) that a study of recurring pragmatic prefabs in a corpus like LOCNEC fails to bring out because they are all used in specific communication situations to perform interpersonal functions that interviewees in an interview setting are unlikely to have to resort to.

Rank	LOCNEC		ICE-GB		BNC		LINDSEI	
1	You know	503	I mean	440	You know	334	I think	439
2	Sort of	406	You know	403	I mean	234	In fact	225
3	I mean	370	I think	323	I think	212	You know	198
4	I think	357	Sort of	253	And then	126	And then	196
5	And then	293	And then	123	I thought	96	I mean	160
6	I thought	100	I thought	71	I know	91	Of course	128
7	Kind of	79	I know	70	Sort of	88	And so on	66
8	And things (like that)	71	That's right	58	That's right	65	For example	61
9	That's right	70	Kind of	56	I suppose	29	I thought	53
10	I suppose	55	I suppose	52	Of course	24	Kind of	48
11	And like	48	Of course	44	But then	13	For instance	47
12	Sort of like	41	In fact	37	And things (like that)	9	Sort of	38
13	I know	33	And things (like that)	20	In fact	9	I suppose	25
14	Of course	26	But then	16	And like	7	That's right	23
15	But then	25	A bit of a	11	A bit of a	7	But then	11
16	In a way	16	In a way	10	Kind of	7	And things (like that)	10
17	A bit of a	16	And so on	9	Sort of like	5	I know	4
18	For example	7	Sort of like	7	In a way	2	In a way	3
19	In fact	5	And like	4	And so on	2	Sort of like	0
20	And so on	2	For example	4	For example	1	A bit of a	0
21	For instance	2	For instance	3	For instance	1	And like	0

Table 2: Relative frequencies of pragmatic prefabs (based on 100,000 words per variety)

The age of the informants in a corpus can also affect the frequency counts of some pragmatic prefabs. *And like* is a case in point. Its high frequency in LOCNEC (university students aged between 18 and 24) and the fact that the discourse marker *like* has been shown by Anderson [1997] to be a prevalent feature of teenage speech seem to point to an age-related phenomenon. What is more, an investigation of *and like* in the speech of speakers aged

between 15 and 24 in the BNC reveals that it occurs with a frequency of 22 per 100,000 words (vs. a frequency of 7 per 100,000 words in the speech of speakers aged 15 to 59) and that the speakers in that age group account for approximately half of the occurrences of the string in the whole corpus.

Formulae	BNC	LOCNEC
Come on	60	3
Thank you	35	1
I see	27	1
Mind you	21	1
See you	18	0
Hang on	14	0
Here you are	9	0
How are you?	6	0

Table 3: Recurring formulae in the BNC (relative frequencies based on 100,000 words per variety)

Another factor in the frequency variations between the three NS corpora is the fact that formulae are prone to significant individual usage variations. In other words, formulae can often be seen to act as 'lexical teddy bears' [Hasselgren 1994] for some speakers. For example, well over half of the occurrences of *sort of like* in LOCNEC are accounted for by one and the same speaker. It is worth noting that formulae like *you know* and *I mean*, which are otherwise fairly equally distributed throughout the corpora, can also be used with extremely high frequencies by some speakers.

In view of such variations, frequency can arguably not be regarded as a reliable criterion when selecting formulae for inclusion in learners' dictionaries. What is more, as was mentioned earlier, it is in terms of the functions they perform on an interpersonal, textual and/or strategic level and not in terms of their frequency of occurrence that pragmatic prefabs are defined. As a result, the primary criterion for inclusion should be whether or not a string is used by the members of the NS English speaking community to fulfil such functions. Corpus-based descriptions of NS English are essential here in order to establish the various functions formulae serve in discourse and the restrictions surrounding their use. Corpus-driven studies, such as studies of recurrent word combinations [De Cock 2000], are also called for. They can be considered to constitute a useful and powerful starting point for studies of pragmatic prefabs as they lead researchers to take into consideration a series of frequently recurring strings of words they may otherwise have overlooked because of their lack of psychological salience. In the same vein, it is essential for computer learner corpus based studies of pragmatic prefabs to move beyond merely exposing frequency patterns of over- and underuse, and to bring to light and closely scrutinise learners' misuses of formulae, i.e. those cases where learners use certain pragmatic prefabs to perform functions they are not normally used to serve in NS language. Information on learners' misuses can directly be used to inform learners' dictionary descriptions of pragmatic prefabs and provide them with genuinely helpful guidance about how to use them in English.

3 Quantitative and Qualitative Analysis of *Of Course* in Learners' Dictionaries

This section addresses the question of how formulae should be described in learners' dictionaries and whether current treatment of these expressions is suited to learners' needs. For lack of space, the coverage of only one pragmatic prefab, *of course,* is examined in greater detail both from a quantitative (i.e. in terms of ease of access, the number of uses distinguished and the number of examples given, labels) and a qualitative point of view (i.e. in terms of the adequacy of the descriptions provided). The five dictionaries are compared and assessed. *Of course* was chosen not only because, as we saw earlier, the French learners in LINDSEI tend to heavily overuse it (in comparison with the native speakers in LOCNEC, the BNC and ICE-GB), but also because a closer qualitative investigation of it reveals that the learners in the corpus show a tendency to misuse it.

3.1 Quantitative analysis of *of course*

Table 4 succinctly summarises the information the five dictionaries provide for *of course*:
(1) the place where it is recorded in each dictionary: whereas MED and COBUILD record it under *of course*, as a lexical item in its own right, LDOCE records it under both *course* and *of course*, and CIDE and OALD record it under *course*;
(2) the number of uses that are distinguished in each (sub)entry (ranging from 5 to 3)[4];
(3) the number of examples listed for each different use (ranging from 1 to 4);
(4) the style and pragmatic labels supplied: while CIDE and MED provide no labels whatsoever, LDOCE and OALD give style labels (spoken, informal) and COBUILD supplies both style (spoken) and framed pragmatic labels (formulae, emphasis);
(5) the usage notes recorded (in LDOCE and OALD);
(6) the frequency information presented: only in COBUILD (part of the most frequent 680 words in their corpus), MED (part of the most frequent 2,500 words in their corpus) and LDOCE (part of the most frequent 1,000 words in the spoken as well as in the written corpus).

The five dictionaries are not equally user-friendly in terms of ease of access and prominence (in decreasing order of user-friendliness): LDOCE and COBUILD can be regarded as the most user-friendly because they record *of course* both in a subentry under *course* (respectively in colour and in bold) and as a full entry under *of course*. MED's treatment of *of course* can be seen as fairly user-friendly (own entry under *of course* printed in red) provided the learners actually think of looking up the phrase under *of course* and not *course* (phrases and expressions are usually listed under the first lexical word they contain). OALD is relatively user-friendly as *of course* is listed fourth in bold type in the IDM (idiom) section (where the phrases are listed one after the other) at the end of the entry for *course*. CIDE is the least user-friendly as its users have to plod through 10 example sentences (with explanations in brackets) listed one after the other under the second main entry for *course* (meaning *development*) before reaching the five examples containing *of course* in bold type (each example sentence is followed by an explanation in brackets).

Of course	Headword	N° of uses (n° of ex.) [extra information, labels]
CIDE	@course	Use 1 (1), Use 2 (1), Use 3 (1), Use 4 (1), Use 5 (1)
COBUILD	-@course -@of course	-See 'of course' – cross-reference [=of course] - [frequ.: 5 black diamonds] Use 1 (3) [spoken, adv, adv with cl, =naturally]; Use 2 (2) [spoken, convention, formulae]; Use 3 (4) [spoken, adv, adv with cl, adv as reply, emphasis]
MED	@of course	[frequ.: 3 red stars, adv.] Use 1 general (1), Use 1a (1), Use 1b (2); Use 2 (2), Use 3 (1)
LDOCE	-@course -@of course	-Use 1 (1), Use 2 (1), Use 3 (1) [spoken, also **course** informal], Use 4 (1) [also **course** spoken]; [see OF COURSE USAGE], -[frequ.: S1, W1, adv.] Use 1 (3), Use 2 (2) + Usage note politeness and style
OALD	@course	Use 1 (2) [also informal **course;** spoken]; Use 2 (2) [also informal **course;** spoken]; Use 3 (1) [spoken]; Use 4 (2); + 'More about' box

Table 4: The treatment of *of course* in the five learners' dictionaries

3.2 Qualitative analysis of *of course*

Consider the following instances of *of course* taken from the French component of LINDSEI (speaker A = interviewer; speaker B = learner):

(1) B: it's a factor of motivation for the students
A: yes and I suppose also they are the ones that are in control on a computer
B: yes **of course**

(2) B: I'm working on er Robinson Crusoe's rewritings
A: oh yes
B: yeah it's fascinating
A: how many times has it been rewritten .. has it it's been rewritten?
B: er yeah yeah **of course**

(3) A: I've heard about this problem in Dublin as well that they can't study literature at all
B: mm
A: because all the courses are full it's a shame that isn't it?
B: yeah **of course**

Using *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. Two learners' dictionaries, namely LDOCE and OALD, actually address the inappropriate use of *of course* in such contexts. In the usage note provided for *of course*, LDOCE stresses the fact that:

> It is not usually polite to use of course or of course not as a reply to a request for information. If for example someone asked you: 'Is this the way to the station?' and you replied 'Of course (it is)',

this 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.

The 'more about *of course* box' in OALD even goes one step further. Not only does it attract learners' attention to problems of usage, but it also supplies them with appropriate alternative ways of reacting and responding[5]:

Of course is often used to show that what you are saying is not surprising or is generally not known or accepted. For this reason, and because it can be difficult to get the right intonation, you may not sound polite if you use of course or of course not when you answer a request for information or permission. It can be safer to use a different word or phrase.
'Is it the right room for the English class? 'Yes, it is' *'Of course' or *'Of course it is' (...)
If you say of course (...) it may sound as though you think the answer to the question is obvious and that the person should not ask. In the same way, of course should not be used as a reply to a statement of fact or when someone expresses an opinion. 'It's a lovely day' 'It certainly is'/'Yes it is.' - *'Of course it is' – I think you'll enjoy that play.' 'I'm sure I will.'/'Yes, it sounds really good' – *Of course.

The use of notes highlighting some of learners' possible misuses of pragmatic prefabs should certainly be encouraged in learners' dictionaries and learner corpus-based studies have a crucial part to play in the compilation of these notes. Pragmatic prefabs also ought to be explicitly and clearly marked as such in learners' dictionaries (by the use of colour and some sort of iconic symbol for example) so that learners can immediately identify them not as expressions that are defined in terms of their prepositional meaning but as expressions that the speakers and/or writers of English use to perform a whole series of interpersonal, textual and strategic functions in discourse. Efforts should ideally also be made to raise learners' awareness of the importance and the workings of pragmatics in their target language as they do not always realise that pragmatic conventions can differ considerably from one language to the next and that using a certain expression in an inappropriate context may well lead to some very awkward situations. MED's 'Language Awareness' section on pragmatics (written by Joanna Channel in very accessible style) and COBUILD's use of pragmatic labels, which are carefully explained and exemplified in a special pragmatics introductory section, can undoubtedly be regarded as a step in the right direction.

4 Conclusion

In the light of what was discussed above, learner-aware coverage of the pragmatic prefab *of course* in learners' dictionaries should be as follows:
(1) It should be recorded in a subentry under the headword *course* and/or as an entry in its own right under *of course* (possibly with a cross-reference under *course* as learners may be used to looking for multi-word units under the first lexical word they contain); (2) *Of course* should be clearly marked and highlighted as a pragmatic prefab; (3) the various uses should be identified on the basis of rigorous NS corpus-based studies; (4) each use should be illustrated by representative examples taken from NS corpora; (5) the (sub)entry for *of course* should be accompanied by a usage note warning learners against using it in inappropriate contexts and supplying them with examples of expressions they should be using instead; (6) pragmatic labels underlining the functions it serves in discourse and (6) frequency information in speech and in writing should also be included.

While thorough descriptions of pragmatic prefabs in NS corpora provide lexicographers with the information they need to distinguish the various uses of formulae, select representative examples, give them pragmatic labels depending on their functions and add any frequency information, the contribution of corpus-based studies using a corpus like LINDSEI can actually be seen as twofold. The multi learner mother tongue background composition of LINDSEI makes it possible for researchers to uncover which pragmatic prefabs tend to be misused by different groups of learners (i.e. cross-linguistic deficiencies [Granger 1998]) and therefore require special treatment in learners' dictionaries aimed at all learners regardless of their mother tongue backgrounds. In addition, the composition of LINDSEI also enables them to shed light on those formulae that are problematic for specific groups only (e.g. transfer-related deficiencies). As there is unfortunately no space in traditional learners' dictionaries to discuss such transfer-related misuses, information on transfer-related problems could in fact be exploited in bilingualised learners' dictionaries, i.e. those learners' dictionaries that are aimed at learners of a particular mother tongue. Although bilingualised learners' dictionaries are still few and far between, it may not be unreasonable to expect learner-corpus based research to be instrumental in the creation of more dictionaries of this kind.

Endnotes

1. The Louvain International Database of Spoken English Interlanguage (LINDSEI) is a corpus of informal interviews with advanced learners of EFL from different mother tongue backgrounds (French, Italian, Japanese, Spanish, Bulgarian, Swedish, etc). For more on LINDSEI: http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Lindsei/lindsei.htm

2 .The pragmatic prefabs in Table 1 have been selected because of their high frequency of recurrence in LOCNEC and/or because they are significantly overused or underused by the learners in the French component of LINDSEI.

3. The asterisked figures indicate statistically significant differences (* = chi-square with $p \leq 0.05$; ** = chi-square with $p \leq 0.01$; *** = chi-square with $p \leq 0.005$). Chi-square measures are not given for those formulae that do not recur at least five times in each corpus because this measure has been shown to be unreliable in such circumstances.

4. LDOCE's treatment of *of course* is somewhat puzzling in that it lists 4 different uses of the phrase under *course* but only 2 under *of course*.

5. 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 (the uses discussed in the note mainly concern spoken language).

Acknowledgement

I would like to thank Sylviane Granger for insightful comments on earlier versions of this paper.

References

[Anderson 1997] Anderson, G., 1997. They wanna see like how we talk and all that. The use of like as a discourse marker in London teenage speech, in: M. Ljung (ed.) *Corpus-based Studies in English.*, pp. 37-48. Rodopi, Amsterdam.

[Alexander 1998] Alexander, R. J., 1998. Really spoilt for choice? Fixed expressions in learners' dictionaries of English, in T. Fontenelle et al. (eds) *EURALEX '98 Proceedings, volume* II, pp. 535-543, Université de Liège, Département d'anglais et de néerlandais, Belgium.

[Bogaarts 1996] Bogaarts, P., 1996. Dictionaries for Learners of English, in: *International Journal of Lexicography, 9 (4),* pp. 277-319, Oxford University Press, Oxford.

[Chafe 1968] Chafe, W. L., 1968. Idiomaticity as an anomaly in the Chomskian paradigm, in: *Foundations of Language 4,* pp. 109-127, Reidel Publishing Company, Dordrecht.

[De Cock 2000] De Cock, S., 2000 Repetitive phrasal chunkiness and advanced EFL speech and writing, in: C. Mair et al. (eds) *Corpus Linguistics and Linguistic Theory, Papers from the 20th International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999,* pp. 51-68. Rodopi, Amsterdam and Atlanta.

[Granger 1998] Granger, S., 1998. The computer learner corpus: a versatile new source of data for SLA research., in: S. Granger (ed.) *Learner English on Computer*, pp. 3-18. Addison Wesley Longman, London and New York.

[Gvishiani et al 2001] Gvishiani, N. & O. Gerwe, 2001. From non-idiomatic to idiomatic phraseology: a contrastive analysis of learner and native speaker corpora., in S. De Cock et al. (eds) *Future Challenges for Corpus Linguistics. Proceedings of the 22nd ICAME Conference, Louvain-la-Neuve, Belgium,* pp. 35-37, Centre for English Corpus Linguistics, Université catholique de Louvain.

[Hasselgren 1994] Hasselgren, A., 1994. Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary, in: *International Journal of Applied Linguistics, 4(2),* pp. 237-260, Oxford University Press, Oxford.

[Herbst 1996] Herbst, T., 1996. On the way to the perfect learners' dictionary: a first comparison of OALD5, LDOCE3, COBUILD2 and CIDE, in: *International Journal of Lexicography, 9 (4),* pp. 321-357, Oxford University Press, Oxford.

[Mittman 1999] Mittmann, B., 1999. The treatment of collocations in OALD5, LDOCE3, COBUILD2 and CIDE, in: T. Herbst et al. (eds) *The Perfect Learners' Dictionary (?), pp.* 101-111. Max Niemeyer Verlag, Tübingen.

[Moon 1998] Moon, R., 1998. *Fixed Expressions and Idioms in English. A Corpus-Based Approach.* Clarendon Press, Oxford.

[Moon 1999] Moon, R., 1999. Needles and haystacks, idioms and corpora: Gaining insights into idioms, using corpus analysis, in: T. Herbst et al. (eds) *The Perfect Learners' Dictionary (?),* pp. 265-281. Max Niemeyer Verlag, Tübingen.

[Svartvik 1999] Svartvik, J., 1999. Corpora and dictionaries, in: T. Herbst, et al. (eds) *The Perfect Learners' Dictionary (?),* pp. 283-294. Max Niemeyer Verlag, Tübingen.

Cambridge International Dictionary of English. 1995. Cambridge University Press, Cambridge. P. Procter (ed.).

Collins COBUILD English Dictionary for Advanced Learners. 2001.: HarperCollins Publishers, Glasgow. J. Sinclair (ed.).

Longman Dictionary of Contemporary English. 2001. Pearson Education Limited, Harlow. D. Summers (ed.).

MacMillan English Dictionary for Advanced Learners. 2002. MacMillan Education, Oxford. M. Rundell (ed.)

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