English for Specific Purposes World
Online Journal for Teachers

ESP WORLD

ISSN 1682-3257
Indexed in
Genamics JournalSeek

Home Contents Resources Links Editors History

Small corpora as assisting tools in the teaching of English news language:
A preliminary tokens-based examination of Michael Swan’s Practical English Usage news language wordlist.

Pascual Pérez-Paredes

Universidad de Murcia, Spain

Departamento de Filología Inglesa

Plaza de la Universidad s/n, Campus de La Merced

30071 Murcia

Spain

TEL: +34 968364378

FAX: +34 968363185

E-mail: pascualf@um.es

Abstract

One of the methodological challenges for any EFL teacher is to decide on the range of vocabulary to be taught and used in classroom work. This has an immediate impact on the lexical content of the learning activities and, if applicable, the scope of language achievement evaluation. Teaching the language of news has become a major feature of English as a Foreign Language (EFL) courses, even if these have just a General English (GE) syllabus or a pre-intermediate audience. EFL teachers find that textbooks and reference books present their very own choices. This is of great aid and a valuable teaching tool. However, very frequently teachers find this solution far from satisfactory. In this article, we will discuss the teaching implications which derive from the use of two small 250.000-word corpora of news language in English and how these relate to the lexical items of news language found in a GE reference manual: Michael Swan’s *Practical English Usage* (OUP). A corpus-based analysis was performed to examine to what extent Swan’s particular choice of words matched the actual use of lexical items in natural news texts. The paper concludes with some critical comments on how small-corpora analyses can, based on evidence, challenge different sorts of certain *mythologies* and help teachers focus their efforts in terms of selection of language contents.

Keywords: ELT, Corpus Linguistics, small corpora, news-language

People in Western countries probably hear more language from the media than they do directly from the lips of their fellow humans in conversation. (Bell 1991: 1)

I. News-Language and English Language Teaching

Although syllabus design is not usually under their direct influence, language teachers certainly have a say on a wide range of methodological decisions that affect students' learning experiences. Choice of vocabulary is one of them. Other important vocabulary areas comprise sequencing, format, presentation, monitoring and assessment (Nation 2001a: 384). In this article, we will concentrate on issues related to lexical content selection of the language used in the news narrative for advanced learners of EFL.

Textbooks usually present vocabulary orderly and in a way that is relevant to the topic under consideration. Once the study unit is completed, students feel they have a core vocabulary that can meet their communicative needs within a particular topic. However, when teaching the language of news[1] most textbooks simply list out a small group of words or, very frequently, a newspaper article is exploited. In Leo Jones' *Progress to Proficiency* (1993), for example, we find both combined. In Gude and Duckworth's *Proficiency Masterclass* (1994), after discussing differences between popular and serious newspapers (p. 199), students are asked to complete a twenty-gap cloze activity on headline news. Note that the two textbooks mentioned above have a worldwide distribution as they are published by leading and highly respected English Language Teaching (ELT) publishing companies. Any of these two approaches might, presumably, leave teachers and students, especially very advanced ones, dissatisfied. We believe that the reason is straightforward. Media, according to Bell (1991), generate a lot of the language that is heard in society and it is only natural that students feel they need supplementary training in that important area: "The language of news media is prominent and pervasive in society, and it is worth understanding how that language works… ”(Bell 1991: xxiii). In a similar way, it is sensible to think that material developers should make sure that their choice of language is representative enough and as comprehensive as possible so as to provide students with ample opportunities to deal with the language of the news.

In terms of diffusion, the news narrative has been said to be the *primary* language *genre* and the one which exerts a more pervasive influence on a given audience (Jensen and Pauly 1997). We should note here that the term *genre* has a complex rendering in ELT contexts as it may subsume different roles. Besides offering a comprehensive account of the issue, Henry and Roseberry (2001) establish a distinction between genre and register, where the latter is said to include the language patterns associated with the former. This implies that word choice is a feature of the language register which, in turn, makes up a genre. Sanderson (1999:2) presents a wealth of arguments that stresses the convenience of using this genre in the ELT classroom. According to this author, newspapers in general have an educational value, carry cultural information which is important for building a good communicative competence in students, reflect changes in the language rapidly as they are "linguistically topical" and up-to-date, expose varieties of English, present students with opportunities to read something of interest to them, offer teachers authentic material which can be easily converted into English for Specific Purposes (ESP) content and, finally, can be used "effectively with a wide range of levels". On top of this group of advantages, Rademann (1998: 49) claims that newspaper articles are "a much more representative sample of a given language than most others", which makes this genre an attractive source of interest to both language researchers and teachers. The LSWE (Longman Spoken and Written English) corpus, which was used for the Longman Grammar of Spoken and Written English, owes 26.68% of its word stock to news language[2]. Together with conversation, fiction and academic prose, the authors state that "these four registers are major categories that span much of the range of situational and linguistic variation in English" (Biber et al. 1999: 25). Hence, it is sensible to think that a detailed account of the language used in this context is required if we are to meet the needs of advanced students of English as a Foreign Language (EFL). That account will necessarily include a selection of lexis that is representative of the news genre as part of what Nation (2001b) calls a *special purposes approach*. Within this framework, learner language needs are matched up by research in specialised vocabulary pertaining to specific language domains or genres. This may eventually lead to better informed choices in vocabulary selection, particularly in the staging and sequencing processes as frequency stands out as one of the most obvious criteria for selection and grading of language (White 1998).

Our first step into a general type of special purposes approach will be to consider already published EFL material that describes the specific vocabulary of newspapers.

II. Swan's Newspaper-Language Wordlist

One of the resources frequently used when introducing the language of news is Swan's 1995 *English Basic Usage Second Edition*. The author himself claims that the

explanations [in the book] deal mainly with standard modern British English, and the examples are as realistic as I can make them (…) A good deal of information is given about American usage, but the book is not intended as a systematic guide to American English (p. xii).

In the blurb on the back of the book, the following can be read:

It answers the learner's questions, "Is this right or wrong, and why?" and the teacher's question, "How can I explain this to my classes?" It gives information and advice that is practical, clear, reliable and easy to find. Most of the book is about grammar, but it also covers selected points of vocabulary, idiom, style, pronunciation and spelling.

Interestingly, and for the purposes of our research, this edition offers a comprehensive group of words that are normally found in newspaper headlines.One may argue that headlines and the rest of the news text do not necessarily share the same lexical properties, but it is a fact that headlines are as much a part of the structure of news stories as the lead or the body of the news story itself. In this sense Bell (1991: 169) allocates the headline within the abstract section of the news text model, together with the lead, the other two sections being the attribution, which is not always explicit, and the story, which can be composed of one or several episodes. Besides, the language of headlines encapsulates much of the contents developed in the news body as well as the newspaper position towards those events.

For many English language teachers, especially in Europe, Swan's book is a reference and an undisputable authority when teaching, explaining and accounting for regularities and irregularities of the English language[3]. As a token of this appreciation, a recent reviewer of the volume said that it "is the bible for the TEFL educator", and continued "this is the ultimate reference guide"[4]. In this sense it is important to reckon the impact of authorised vocabulary choices made by native speakers on the EFL professional community. Sinclair (1997:2) states that language teachers are dependant on those who describe the L2 for their "basic information and its stability and reliability". A good example of this is West's (1953) *A General Service List of English Words* ***[5]*** which, according to Bauer and Nation (1993), is basically a collection of 2000 thousand headwords representing word families. West's wordlist influenced ELT authors for decades as *his* particular selection of common words was behind different learning materials (Nation 1990; Sánchez 2000). Precisely, one of the strengths of this selection was that the words therein were chosen on frequency claims to meet the communicative needs of a particular group of people[6]. Bauman and Culligan went further on and, using Bauer and Nation´s (1995) multilevel standard for determining headwords, extended the list to 2284 words and used the Brown [7]Corpus to determine frequency. Nation's (2001b) latest works someway still rely on this word listing. As we will see now, machine-readable collection of texts have opened up new ways into vocabulary handling and teaching applications.

III. The Use of Small Corpora in ELT Research and Pedagogy

The sub-corpus, or mini-corpus, approach is gaining recognition and users. The very strength of small corpora is that they do not need to be as large as a general, big corpus if, considered other design factors, they are representative of the language domain and instrumental in meeting the needs of designers and corpora users alike. In this respect, Tognini-Bonelli (2001: 57) put it very straightforwardly when she stated that "it would be difficult to see the reasons for choosing an unrepresentative corpus". Certainly this is the major asset of small corpora: they are not small because *small* is better, but because *small* can be more authentic and representative in a specific language domain and genre than reference corpora.

It should be said that using approaches like ours, that is, top-down research where the corpus is assisting teachers and researchers in checking existing statement against evidence, some of the difficulties which derive from polysemy are not so much of a problem given the restricted contexts of communication which small corpora represent. Of course, this is always subject to analysis and revision but it is not likely that words which are characteristic of a language domain will present 21 meanings on average such as those in West's list[8]. Nation (2001:34) admits that even recent software like *Range* or *Vocabulary Profile* are not capable "of distinguishing different meanings of the same forms".

Different researchers have made use of small corpora[9] in their investigations. For example, Jackson (1997) reports students' use of mini-corpora to implement computer-based stylistic analysis of texts. Cheng and Warren (1999) collected spontaneous naturally occurring Hong Kong spoken English and built the Hong Kong Corpus of Conversational English (HKCCE), which comprises around 500.000 words. Gavioli (2000) reports the use of an English medical corpus of 258.622 words which comprise five specialist sub-domains of similar size. Simpson (2000) documented a corpus of spoken academic English which consisted of four mini corpora of around 250.000 words each. Other researchers have used portions of existing corpora to narrow the scope of their investigations. One of these is Westergren (1996: 5) who placed her "focus of interest on the place and function of contractions in the Lancaster-Oslo-Bergen (LOB) Corpus press text categories A–C, each representing its own genre". Tribble (2000) used a similar approach to contrast written genres. In our work the genre is the same but the corpora are different.

Ma (1993) reports using concordancing based on small corpora in English as a Second Language (ESL) settings in three different domains: as an aid for syllabus design, for classroom teaching and for test construction purposes. This approach places itself well within the pedagogical approach advocated by Granger and Tribble[10] (1998: 200) and which can be summarized in the following two basic assumptions: (a) form-based students' instruction is conducive to the success of Second Language Acquisition (SLA) and (b) corrective feedback is positive and necessary for successful SLA. A recent work in this field has been carried out by Kenning (2000) who preferred to lay emphasis on the use of Key Word in Context (KWIC) software to identify comprehension pitfalls rather than exploiting concordancing lines for teaching purposes. The author states that

These various forms of remedial action might usefully employ concordance-based materials. Indeed, can there be, for instance, a more effective means of developing student awareness of the broad range of contexts in which (a particular node or word) acts (…) and of the varied ways in which this (syntactic) function is conveyed in English, than through sets of parallel sentences drawn from concordance data? (p. 168).

These works share an interest in (a) an empiricist approach to language analysis, (b) language performance and (c) the description of language, together with researchers' evaluation, based on quantitative methods. In this article, we aim at attempting a reconciliation between statement and evidence (Tognini-Bonelli 2001) by exploring how the words listed by Swan manifest themselves in two different mini-corpora of news-language. By doing so, we seek to present teachers and students with the opportunity to reflect on how standard EFL material shapes the way in which the teaching and learning experience is confronted.

IV. Corpus-aided Analysis

IV.1. Preliminary insights into the mini-corpora used

In our analysis, 157 items in Swan's wordlist were scrutinized. These items can be found in Appendix 1[11]. Considering the fact that we had applied approach in mind, we decided to accomplish our examination of actual, authentic and attested data by focusing our attention on two mini-corpora which intentionally differed in different ways. See Figures 1 and 2 for details.

CORPUS 1

Variety: British English

News content: British home news

Number of tokens: 209.610

Types: 15.706

Corpus origin: Oxford University Press MicroConcord Corpus Collection A: Home News.

Publishing Date: 1993

Texts used: articles from the Home News Page from The Independent and The Independent on Sunday

Figure 1: OUP MicroConcord Corpus

CORPUS 2

Variety: American English

News content: US main stories, European news, Asian stories, entertainment, the environment and human interest stories. Each scope accounted for a 16.6% of the total

Number of tokens: 254.485

Types: 20.151

Corpus Origin: InfoBeat. URL at http://www.infobeat.com

Compilation date: November1 1999 – February 21 2000

Texts used: summaries of news stories, usually 6-lines long

Figure 2: American Corpus

Corpus 1 is one of the published sources which include the news genre. Other available, commercial corpora were not considered because their publication date (BROWN or LOB). Both corpora- Stubbs (1996:81) would call them collection of large texts- or mini-corpora (Tribble 1997), are similar, but not identical, as for length and content are concerned. Corpus 2 is more varied and one could expect a wider range of types than in Corpus 1. But certainly it is language variety and language production date that make these two corpora worth contrasting. While the first is British, the second is American; while the first was published in 1993, and accordingly the language in the corpus can presumably be dated back to 1991 or 1992, the second is very recent, which implies that both contents and vocabulary are closer to speakers nowadays. Whether this poses an advantage or not, only researchers and teachers are to say. However, we believe that when it comes down to teaching a language, *proximity* is always a plus, not an extra; a turn-on, never a turn-off for students.

In our case, we thought that using these two mini-corpora would expose the status of the prescriptive wordlist in Swan (1995) in terms of the different parameters which are usually present in language teaching contexts. Language variety and date of compilation appear as most outstanding but others such as news content and publication source also play a major role in the way information and language data are conveyed. Classroom-oriented corpus-based research, in this sense, is highly beneficial as may become a tool to assess the relevance of native speakers' choices of vocabulary which, for different reasons, may or may not be adequate to the needs of specific learners.

In a situation of rapid linguistic change (Crystal 2001), teachers must be sensitive to apparently diverging trends in the English language. Accommodation of native speakers to international standards, on the one hand, and more flexible attitudes towards variety and language norm, on the other, do coincide. This "dynamic linguistic relativism" (Crystal 2001: 63) calls for a more active role of teachers in the handling of the contents which are delivered to their students. This is where small corpora come in. Let us go back to our corpora. Some interesting data are really food for thought: in Corpus 2, *Pinochet* accounts for 0.01% of all tokens, *Kosovo* accounts for 0.04% and *Timor* accounts for 0,01%. Which sounds very familiar to our ears, or at least it did at the time Corpus 2 was compiled. In terms of Data Driven Learning, students can relate to these stories and, no doubt, they will be interested in them as they shape the perception of the society in which they are living. Now compare the following. In Corpus 1, Margaret accounts for 0.02% of all tokens, Thatcher accounts for 0.03% and Heseltine accounts for 0.01%. This is just an example of the important contribution of corpora to the teaching of languages, which, following McEnery and Wilson (1996: 104) can be of two types:

(First) corpus examples are important in language learning as they expose students at an early stage in the learning process to the kinds of sentences and vocabulary which they will encounter in reading genuine texts in the language or in using the language in real communicative situations (…) However, corpora, much more that other sources of empirical data, have another important role in language pedagogy which goes beyond simply providing more realistic examples of language usage. A number of scholars have used corpus data to look critically at existing language teaching materials.

Although not the main focus of this research, serendipity[12] and DDL are at the heart of corpus based approaches to FLT. It seems that politicians come and go and this fact must be part and parcel with our experience of life. But one thing is politicians’ names and quite another is actual language use. Swan (1995) outlines *EC*, European Community, as an important word in English news language. After a word frequency analysis, we found out that Corpus 1 returned 18 tokens of such item; Corpus 2 none. However, Corpus 2 returned 42 occurrences of EU, European Union; Corpus 1, none. It is in these easily-outdated contexts that small corpora can be highly beneficial to teachers and learners of a FL. In this sense, Rademann (1998: 66) sees important advantages in using electronic news corpora as "linguistic features across English-language quality dailies from different nations and international editions of a given paper" can be easily obtained and researched.

Both corpora are different in two other crucial aspects: readership and distribution. Corpus 1 addresses a cultivated audience through, say, traditional distribution devices. Corpus 2, as a way of contrast, has a “hidden” audience of Internet users who receive the news for free on a daily basis. The information is stored in their e-mail accounts and, even when e-mail is downloaded, this can be left in the inbox folder. We were interested in comparing two opposing collection of texts in the hope that this could bring about a richer discussion and a more diversified perspective on the topic under consideration. In fact, our two corpora are statistically different. We ran a Coefficient Interval Test for Two Proportions and found out that, considering the token and type magnitudes of both corpora, these differed significantly: z= -5.41 (p=0.000)[13]. Different sources imply different selections made by language users as, using Maletzke’s 1963 classic model of the mass media, communicators’ self-image, personality, working team, organization role, social environment and media constraints all interact even before the content structures and vocabulary selections are approached. This shows how complex vocabulary selection can get and how necessary tools like small corpora can be in a context where the teaching of a genre or a domain are implied.

The limitations of analysing language using small corpora are essentially related to corpus size. Biber, Conrad and Reppen (1998: 30) state that a very large corpus is needed if we are to study the meaning and use of words. Bearing the LOB corpus in mind, we may consider that one million words are not much to explore all the potential meanings of a given lexical item. However, it is nonetheless true that the occurrence of moderately uncommon vocabulary is largely dependant on the topics and text typology represented in the corpus. Engwall (1994:51) thinks that

> no scientific criteria exist for determining the size of any corpus. It has to be decided simply with reference to a balance of depth and breath, but the lack of resources sometimes restricts the desired design.

However, the question of size is of great importance in domain-specific corpora. In fact, one may ask how decisive is corpus size when specific texts are scanty. There are empirical ways to explore the potential limitations and virtues of any n-word corpus. The use of the Type-Token-Formula (Sánchez and Cantos 1997) allows us to calculate estimations of the quantity of types for any given corpus based on n-tokens. Yang et al. (2000) and Cantos (2000) report that the projections of different multi-million token samples are greatly accurate. In our case, using the formula mentioned above, the types projections for both corpora would be as follows[14]:

Projection over	Corpus 1	Corpus 2
Actual number of types	15706	20151
2-million words	48514.835	56491.161
3-million words	59418.299	69187.260
5-million words	76708.695	89320.369

Table 1: Projections of types for Corpus 1 and 2.

This gives us a precise idea of the sort of corpora size needed to encounter a more varied range of *types* of news language vocabulary. Based on these data, we may assess how adequate Swan’s news-language wordlist can be for an exploratory EFL-oriented analysis. We should remember here that, to the best of our knowledge, there exists no commercially-available specific corpus of English news language which reaches up figures of standard general corpora.

With an educational use in mind, corpora 1 and 2 do clearly serve the purpose of presenting us with opportunities to test the actual occurrence of *the* specific lexical items listed by Swan. In this sense, we believe that specialized corpora compiled by teachers or researchers can play an important role in a narrowly-defined contexts such as the study of how standard and authorized news-language vocabulary (as proposed by Swan) actually can be tracked down in specific news-language use. This is the scope of the next section.

IV.2. An empirical lexical taxonomy of news-language vocabulary

After running a frequency analysis on both corpora, the original list of lexical items was classified according to occurrence criteria. Note that tokens are used as primary data in this analysis as Swan´s choice is conceptually tokens-oriented.

One in 150 tokens[15] was selected as a cut-off point for frequency significance. In Corpus 2, 26 occurrences stand for a minimum 0.01%, while the figure decreases to 22 for the same percentage in Corpus 1. This is so because of the different size of both corpora. As an example, we could take *deal* which in Corpus 1 appeared in 55 instances reaching a 0.03% token percentage. The same item appeared 101 times in Corpus 2 reaching a 0.04% figure. Despite the word amount and proportional differences, both corpora present very similar Type/ Token figures, 7.49 (C1) and 7.92 (C2), and similar standardised Type/ Token measures, 46.83 (C1) and 53.43 (C2).

In order to examine the behaviour of Swan's wordlist we decided to classify the items taking into account their appearance on (a) both or only one corpus and (b) their frequency (> 0.01% or <0.01%). Five different groups emerged:

Group A. Swan’s words with frequency >0.01% found in both corpora: 21 words

act, aid, allege, appear, ban, block, call, campaign, charge, claim, clear, cut, deal, hit, IRA, link, move, plant, press, threat and US.

Table 2: Swan's wordlist Group A

Group B. Swan’s words with frequency >0.01% found in one corpus only: 14 words

bar, cash, Commons, firm, jail, key, mission, MP, poll, storm, Tory, troops, UK and Ulster.

Table 3: Swan's wordlist Group B

Group C. Swan’s words with frequency <0.01% found in both corpora: 65 words

alert, bid, blast, blaze, blow, bolster, bond, boom, boost, brink, clash, con, curb, deadlock, dole, drama, edge, foil, fraud, go-ahead, grab, grip, hail, halt, haul, hike, leak, loom, Lords, mercy, net, odds, oust, pact, peak, peer, peril, pit, plea, pledge, pull out, raid, rampage, rift, row, rule out, saga, scare, scrap, seize, set to, shed, slate, slump, spark, split, stake, swap, sway, switch, toll, trio, UN, vow and wed.

Table 4: Swan's wordlist Group C.

Group D. Swan’s words with frequency <0.01% found in one corpus only: 43 words

axe, BR, chop, crackdown, cutback, envoy, feud, flak, flare, freeze, gag, go for, head for, jobless, landslide, lashed, leap, marred, mob, nailed, opt for, PC, peg, PM, pools, Premier, probe, push for, quake, quit, quiz, rap, riddle, sack, slam, slash, snatch, soar, spree, stun, surge, swoop, and walk out.

Table 5: Swan's wordlist Group D

Group E. Swan’s words not found in any corpus at all: 14 words

BA, clamp down on, dask, demo, gaol, gems, gun down, hit out at, hitch, in the red, mar slay, storm out of and VAT

Table 6: Swan's wordlist Group E

The percentage distribution of all 5 groups is found on Figure 3.

GROUPA: 13.4%	GROUP B: 8.9%	GROUP C: 41.4%
GROUP D: 27.4%	GROUP E: 8.9%	

Figure 3: percentage distribution

Group A and B comprise 35 lexical items which can be tracked down in the language evidence presented here. These groups stand for 22.3% of Swan’s words under scrutiny. Group C is composed of 65 lexical items present in Swan’s list, that is 41.4% of the items under scrutiny. This is, by large, the word collection which most significantly contribute to the *newsness* of our data. The fact that they fail to come over the 0.01% significance cut-off tells us that, despite their presence, we were only able to track down somewhat sporadic use of the items. Group D is particularly interesting as 43 lexical items present in Swan’s list are found in exclusively one corpus, irrespective of which this is, which implies that their use in highly context-dependent and their usefulness potentially restrictive. These items stand for 27.4% of Swan’s items studied here. Groups C and D, those items listed by Swan with <0.01% representativeness, account for 68.8% of the total wordlist. In Group E we find 14 lexical items which are not found in either corpus.

The word frequency information we obtained could have very well served rationalist approaches to pedagogical decisions about news-language representativeness in the time scope and diffusion contexts that our two corpora were originated. A word of caution is required here, however. Obviously, when listing out items not selected by Swan, one had to decide on words which, based on intuition, belonged to this particular news-language field. So one had to approach the issue exactly in the same way as Swan did when he compiled his wordlist, this time using corpora as an auxiliary tool. For instance, we could hypothesize about Group H, composed of *forensic, lured, mocked, slick* and *snapped*, words not listed out by Swan with frequency <0.01% found in both corpora. It is arguable whether these should be included in a news-language vocabulary repertoire; however, it is very sensible to acknowledge that, at least for a period of time, those words were represented in the two collection of news texts. In this respect, it is interesting to see how wordlists become easily dated. Take “pit”, one of Swan’s choices. Coal mines made quite a splash in British papers during the eighties as the sector problems were a major home issue at the time. While pits are still there, they are no longer a priority in news stories these days, and so the word fails to appear significantly in either corpus (11 and 5 occurrences in Corpus 1 and 2, respectively). Group F could be composed of words, not listed out by Swan, with frequency >0.01% found in both corpora: like NATO, spokesman, survey, trial, and Union. We could have Group G, composed of words, not listed out by Swan, with frequency >0.01% found in one corpus only: 45 words like abuse, acid, courts or constituency. The percentage distribution of these three groups is found on Figure 4.

GROUP F: 9%	GROUP G: 82%	GROUP H: 9%

Figure 4: percentage distribution

Groups F and G comprise 50 lexical items which, potentially, might contribute significantly to the news-language vocabulary data used for EFL work. Group H is composed of five lexical items which are not found in Swan's repertoire which, as stated above, we must admit rather intuitively, belong to the range of news vocabulary the EFL community may come up with.

V. Pedagogical considerations

It follows from the results above that Groups A and B, those composed of items listed by Swan with frequency > 0.01 %, reach together 22.3% of the representativeness of the total word stock presented by him, while almost 9% of *his* words are not even found once in the 464.095-token joint corpus used in this research. Group C alone makes a 41.4% figure while 27.4 % of the word stock examined is only found in one of the corpora. This shows that low-frequency words, 68.8% in the joint-corpus, help demarcate topics and carry essential meaning in the context of news-English. All things considered, several points which stem from these insights are relevant to corpus-based studies in the field of FLT.

First, surely enough, different corpora will cast different conclusions on the same range of vocabulary under study. Only *unfeasible* corpora of all the possible linguistic performances would exactly return the same data, but according to current standards this is just a chimera. With available machine readable texts, topics of interest to particular FL students can be narrowed down to fully adjust to learner's needs, interests or classroom learning schedule. In view of this, it is necessary to reckon that when using small corpora, contrasting and combining different collections of texts, such as the ones used here, will increase the scope of the lexical analysis to be performed. The use of combined small corpora will result in a more diversified picture of language behaviour in specific domains where somehow classroom application is in the agenda:

What sort of corpus do foreign language learners and teachers need then? I would say first of all that they almost certainly need many corpora rather than one. Students of English want to be able to write different kinds of text and become effective language users in many different contexts. They need a rich set of potential models for their own language behaviour (Tribble 1997).

The combination of small corpora of different nature helps teachers make informed choices as to the adequacy of vocabulary lists such as Swan's. Obviously, the basic, underlying principle advocated here is that frequency, ranging from absence to statistical significance, has a value in its own and that can be used as a tool by instructors or materials developers in order to explore and classify the specialised vocabulary to be introduced to learners[16]. Going back to McEnery and Wilson (1996), we can see that a communion of authentic data, empirical findings about the language and corpora as a source of materials will play up the sense of purposefulness and language verification which is so positively valued by students. Besides, the type of work we present here

Second, EFL teachers must be aware of the importance of using materials that effectively meet their students' needs, whether these are more form, communicative, cultural-oriented or even all three combined. Besides, teachers know that students come to classrooms with their own particular baggage in terms of background knowledge, language, expectations, learning styles, confidence, motivation and personal circumstances (McKay and Tom 1999). It is reasonable to think that some groups of students will be pleased to engage in work with one type of vocabulary and will find another unmotivating, too dated, or simply will reject the idea of working with raw language items. This circumstance is to be carefully weighed down by teachers when implementing corpus-based selection of vocabulary and learning activities and, visibly, this is where small corpora outweigh their bigger relatives as teachers can collect data that is fit to their praxis. In this way we can note here that commercial specialized, non-general, corpora are not abundant, and, given the availability of machine readable texts, it is reasonable to put forward that FL teachers adopt an active role in pursuing corpora-gathering efforts which can be pedagogically beneficial to their students.

Third, the type and range of classroom vocabulary work proposed by Hoey (1997) and Tribble (2000) is more readily accomplished with small corpora as they tend to be more sensitive to specific domains than larger corpora. The BNC offers, according to Lee's BNC Index[17], fifteen texts where news stories are collected. They total 282.500 words. One of the texts is composed of 158.242 words, which actually implies that it alone covers 56% of the representativeness of the BNC news stories. One wonders how representative[18] this subcorpus can get even if it is a part of a well-established and reputed larger corpus. Again, we find in this illustration another reason to advocate the non-commercial compilation and use of specialised corpora for the teaching and learning of well-narrowed communicative language domains.

As for the corpus-aided selection of news-language vocabulary, we can see how Group G constitutes a highly context-dependent selection of items as they are found in one corpus only, while group F should rationally be included in any account of news language vocabulary as the frequency and presence of those words in both corpora make it a candidate for students' learning and teachers' description. However, we should be careful here. Corpus-aided selection and speakers' intuition go hand in hand. Terminological extraction would be fairer in terms of down-to-top processes -from evidence to statement-, that is, the opposite to the one advocated here which examines statement with the aid of evidence. In this respect, we recommend that such possibility be explored and results contrasted against the data we present here.

As already stated, the corpora we used can allow teachers assess the usefulness of the wordlist in terms of frequency on the two corpora as well as evaluate potential uses of Swan's listing. One may think that the fact that only 14 of Swan's lexical items did not make it into any corpus wordlist group can tell us about how appropriate his selection was. Some of the items in this group are too specific and presumably in other news-language contexts they would appear on the scene (take VAT, for instance). Or, all the way round, others may argue that if 41.4% of "his" vocabulary did not even reach the 0.01% significance threshold, Swan's very own selection is far from depicting actual news-language use. An old debate starts again whether corpus-based analyses should exclusively rely on data-driven procedures and conclusions or whether language data is always to be complemented by teachers and researchers' own insight into the concerning question. Fillmore (1992) addressed the issue a decade ago and recently (Fillmore 2001) reaffirmed his idea that both resources, that is, raw language data and speakers/ researchers' introspections are necessary to "succeed in the language business" (2001:1).

Sánchez (2000) holds that digitalized corpora, whether compiled by teachers/linguists or commercial software, are useful in, at least, three areas (a) as a source of materials for investigation; (b) as a source of classroom work and (c) as a source of information to establish the adequacy of texts being used in the classroom. The work here presented shows how all three aspects can be interconnected and provide feedback that helps teachers make informed choices as to what vocabulary is to be introduced in the classroom. By comparing two small, although different, news language corpora we can present students with contextualized instances of data-driven news language that is updated, with vocabulary that is based on evidence and which is faithful to native speakers' usage, with an account of vocabulary which is relevant in different varieties (i.e. BrE and AmE). Similarly, the use of different, diverging corpora may be convenient when the context of communication is also intrinsically interesting as producer, readership, distribution and genre may be scrutinized as potential factors for language diversity.

Although it is true that corpora still remain unknown territory to most FL teacher, one of the ideas underlying this work is that teachers can understand that they can build *their own* corpora for *their own* specific purposes, and that their students can benefit from that in numerous ways:

> A corpus is not a magic tool that might solve all the problems teachers have to face. But it helps in solving some of them and might bring into the classroom real language usage in connection to what modern technology can offer in assisting teaching. (Sánchez 2000:29)

As for the case of news-language, after our analysis we can say that a teacher, especially a non-native one, is better equipped to present his or her students with actual examples of such vocabulary through concordancing lines or any other form he or she may think fit. This contextualized stock will soundly reflect usage at two different points in time within the same decade and, presumably, will bring about cultural discussion on socio- cultural issues which may enrich the learning process.

References

Atkins, B.T.S. and Zampolli, A. (Eds.) (1994). *Computational Approaches to the Lexicon*. Oxford: Oxford University Press.

Bauer, L. and Nation, I.S.P. (1993).Word families. *International Journal of Lexicography*, 6 : 253-79.

Bauman and Culligan. URL http://plaza3.mbn.or.jp/~bauman/gsl.html (as of July 15th, 2001).

Bell, A. (1991). *The Language of News Media*. Oxford: Blackwell.

Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: CUP.

Biber, D. et al. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Burnard, L. and McEnery, T. (eds.) (2000). *Rethinking language pedagogy from a corpus perspective: papers from the third international conference on teaching and language corpora*. Hamburg: Peter Lang.

Burns, A. and Coffin, C. (Eds.). *Analysing English in a Global Context*. London: Routledge.

Cantos, P. (2000). A Multidimensional Corpus-Based Analysis of English Spoken and Written-To-Be-Spoken Discourse. *Cuadernos de Filología Inglesa*, Vol 9 (1), 39- 70. Corpus-based Research in English Language and Linguistics. Universidad de Murcia.

Cheng, W. and Warren, M. (1999). Facilitating a description of intercultural conversations: the Hong Kong Corpus of Conversational English. *ICAME Journal* 23, 5-20.

Crystal, D. (2001). The future of Englishes. In Burns, A. and Coffin, C. (Eds.).

Engwell, G. (1994). Criteria in corpus creation. In Atkins, B.T.S. and Zampolli, A. (Eds.)

Fillmore, C. J. (1992). “Corpus linguistics” or “Computer-aided armchair linguistics”. In

Fillmore, C. J. (2001). “Armchair linguistics vs. corpus linguistics revisited”. *ICAME 2001 Future Challenges for Corpus Linguistics. Proceedings of the 22nd International Computer Archive of Modern and Medieval English Conference.* De Cock, S., Gilquin, G., Granger, S. and Petch-Tyson, S. (Eds). 2001, pp. 1-2. Louvain: Universite Catholique de Louvain.

Ferguson, M. & Golding, P. (Eds.) (1997). *Cultural Studies in Question.* London: Sage

Gavioli, L. (2000). Some thoughts on the problem of representing ESP through small corpora. Presentation delivered at Teaching and Language Corpora (TALC) 2000. Graz, Austria.

Ghadessy, M., Henry, A. and Roseberry, R. (Eds.). (2001). *Small Corpus Studies and ELT. Theory and Practice.* Amsterdam: John Benjamins.

Granger, S. (ed.) (1998), *Learner English on Computer*. Harlow: Longman.

Granger, S. and Tribble, C. (1998), Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning. In: Granger, S. (Ed.).

Gude, K. and Duckworth, M. (1994). *Proficiency Masterclass*. Oxford: Oxford University Press.

Henry, A. and Roseberry, R. (2001). Using small corpus to obtain data for teaching a genre. In Ghadessy, Henry and Roseberry (Eds.).

Hoey M (1997). *From concordance to text structure: new uses for computer corpora*. Paper presented at the First international conference: Practical Applications in Language Corpora (1997), University of Lodz.

Howatt, A.P.R. (1984). *A History of English Language Teaching*. Oxford: Oxford University Press.

Jackson, H. (1997).Corpus and Concordance: Finding out about Style. In Wichman, A., Fligelstone, S., McEnery, T. and Knowles, G. (Eds). *Teaching and Language Corpora*. Harlow: Longman.

Jensen, J. & Pauly, J.J. (1997). Imagining the Audience: Losses and Gains in Cultural Studies. In Ferguson, M. & Golding, P. (Eds.).

Johns, T. (1991). Should you be persuaded: two examples of data-driven learning. In Johns, T. and King, P. (eds.), *Classroom Concordancing.* Birmingham: University of Birmingham, pp. 1-16.

Johns, T. (1992). From printout to handout: grammar and vocabulary teaching in context of data-driven learning. In Johns, T. and King, P. (eds.), op. cit., pp. 27-45.

Johnson, K. and Johnson, H. (Eds.). (1998). *Encyclopedic Dictionary of Applied Linguistics.* Oxford: Blackwell.

Jones, L. (1993). *Progress to Proficiency. Student’s book.* CUP.

Kenning, M. (2000). Concordancing and comprehension: preliminary observations on using concordance output to predict pitfalls. *RECALL* 12 (2): 157-169.

Ma, B. (1993). Small-Corpora Concordancing in ESL Teaching and Learning. *Hong-Kong-Papers-in-Linguistics-and-Language-Teaching*; v16 p11-30.

Maletzke G (1963) *Psychologie der Massenkommunikation* Hamburg: Verlag Hans Bredow-Institut.

McEnery, T. and Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Microconcord Corpus Collection A. (1993). Oxford: Oxford University Press.

Nation, I.S.P. (1990). *Teaching and Learning Vocabulary*. New York: Newbury House.

Nation, I.S.P. (2001a). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, I.S.P. (2001b). Using small corpora to investigate learner needs. In Ghadessy, Henry and Roseberry (Eds.).

Rademann, T. (1998). Using online electronic newspapers in modern English-Language press corpora: Benefits and pitfalls. *ICAME Journal* 22, 49-72.

Sánchez, A. and Cantos, P. (1997). Predictability of Word Forms (Types) and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus: An 8-Million-Word Corpus of Contemporary Spanish. *IJCL*, 2(2): 259-280.

Sánchez, A. (2000). Language Teaching before and after "digitilized corpora". Three main issues. *Cuadernos de Filología Inglesa*, Vol 9 (1), 5-37. Corpus-based Research in English Language and Linguistics. Universidad de Murcia.

Sanderson, Paul (1999). *Using Newspapers in the Classroom*. Cambridge: Cambridge University Press.

Simpson, R. (2000). Cross-disciplinary Comparisons in a Corpus of Spoken Academic English. Presentation delivered at Teaching and Language Corpora (TALC) 2000. Graz, Austria.

Sinclair, J. (1997). Corpus evidence in language description. In Wichman, A. et al. (Eds.)

Sinclair, J. (2001). Preface to *Small Corpus Studies and ELT*. In Ghadessy, Henry and Roseberry (Eds.).

Stubbs (1996). *Text and Corpus Analysis.* Oxford: Blackwell.

Svartik, J. (Ed.) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August*. 1991, pp. 35-60. Berlin/ New York: Mouton de Gruyter.

Swan, M. (1995). *English Basic Usage. Second Edition.* Oxford: Oxford University Press.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work.* Amsterdam: John Benjamins.

Tribble (1997) *Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching.* Paper presented at the First international conference: Practical Applications in Language Corpora (1997), University of Lodz, Poland. Available at URL http://ourworld.compuserve.com/homepages/Christopher_Tribble/Palc.htm#Which Corpus (As Of July 15th 2001).

Tribble (2000). *Genres, Keywords, Teaching: towards a Pedagogic Account of the Language of Project Proposals.* In Burnard and McEnery (Eds.).

West, M. (1953). *A General List of English Words*. London: Longman.

Westergren, M. (1996). Contracted forms in newspaper language: Inter-and intra-textual variation. *ICAME Journal*, 20, 5-22.

White, R. (1988). *The ELT Curriculum. Design, Innovation and Management*. Oxford: Blackwell.

Wichman, A. et al. (eds.) (1997), *Teaching and language corpora.* Harlow: Longman.

Yang, D H., Cantos, P. and, Song, M. (2000). An algorithm for predicting the relationship between lemmas and corpus size. *ETRI Journal*, 22, 2, 6.

---

[1] We understand *news language* as the English language used when conveying news by a publishing organization, irrespective of the subgenre.

[2] 10.679.300 words out of 40.025.700.

[3] The book sold around 400 copies in Spain in 2000, according to Oxford University Press representatives in Spain. They also pointed out that the figure is more than acceptable as the book is not especially promoted in the OUP catalogue as teachers and students have traditionally bought it for a decade as it is recognized within the ELT community as an important and very useful reference book.

[4] Review by Kenneth G. Johnson, at amazon.com (10/02/2002)

[5] This list comprises 80 % of the words in any written text according to the Encyclopedic Dictionary of Applied Linguistics.

[6] See Howatt (1984: 245-50) for a discussion.

[7] The Brown Corpus is a collection of 1,014,312 words of running text of edited American English prose printed in the United States during 1961. Francis, W. N. and Kucera, H. (1979) *BROWN Corpus Manual. MANUAL of Information to accompany a Standard Corpus of Present-Day Edited American English, for use with Digital Computers.Revised and amplified.* Providence, Rhode Island: Brown University.

[8] Johnson and Johnson (1998)

[9] Sinclair (2001: ix) states that corpus sizing is always relative: "the dimensions of a small corpus vary with the date it is compiled".

[10] For a ground-breaking discussion on Data-Driven Learning see Johns (1991) and Johns (1992).

[11] 23 words were not considered as they conform a group of lexical items very frequently found in non news-language contexts. They appear italicised in Appendix 1.

[12] See Bernardini, S. 2000a. “Systematising serendipity: Proposals for concordancing large

corpora with language learners." In Burnard, L. and T. McEnery (eds). *Rethinking language pedagogy from a corpus perspective: Papers from the third international conference on teaching and language corpora*. (Lodz Studies in Language). Hamburg: Peter Lang. 183-190.

[13] Minitab 13.2 was used to perform this analysis.

[14] If types growth was linear, which is not, and we had 5-million-word corpora, one could expect 374648 and 395917 types in corpus 1 and 2, respectively.

[15] 0.01%

[16] Other options include terminological extraction. By doing so, the focus will be placed on *terms* rather than on traditional FL vocabulary renderings. This represents a restrictive view of what the language of news might be like in terms of learnability and familiarity teaching principles.

[17] It can be reached at http://www.davidlee00.freeserve.co.uk/corpus_resources.htm

[18] In terms of purely linguistic research purposes. However, large corpora do not necessarily behave in an identical way. For instance, in the LSWE corpus, domestic/local/ city news subcorpora are composed of 1.233.900 and 995.000 words in the British and American English varieties, respectively

Appendix I

ACT
AID
ALERT
ALLEGE
APPEAR
AXE
BA
BACK
BAN
BAR
BID
BLAST
BLAZE
BLOCK

BLOW
BOLSTER
BOND
BOOM
BOOST
BR
BRINK
CALL
CAMPAIGN
CASH
CHARGE
CHOP
CITY
CLAIM
CLAMP DOWN ON
CLASH
CLEAR
COMMONS
CON
CRACKDOWN
CRASH
CURB
CUT
CUTBACK
DASK
DEADLOCK

DEAL
DEMO
DOLE
DRAMA
DRIVE
DROP
DUE
EC
EDGE
ENVOY
FACE
FEUD
FIND
FIRM
FLAK
FLARE
FOIL
FRAUD
FREEZE
GAG
GAOL
GEMS
GO
GO FOR
GO-AHEAD

GRAB
GRIP
GUN DOWN
HAIL
HALT
HAUL
HEAD
HEAD FOR
HIKE
HIT
HIT OUT AT
HITCH
HOLD
IN THE RED
IRA
JAIL
JOBLESS
KEY
LANDSLIDE
LASH
LAUNCH
LEAD
LEAK
LEAP
LIFE
LINK

LOOM
LORDS
MAR
MERCY
MISSION
MOB
MOVE
MP
NAIL
NET
ODDS
ON
OPT FOR
OUST
OUT TO
OVER
PACT
PAY
PC
PEAK
PEER
PEG
PERIL
PIT
PLANT

PLEA
PLEDGE
PM
POLL
POOLS
PREMIER
PRESS
PROBE
PULL OUT
PUSH FOR
QUAKE
QUIT
QUIZ
RAID
RAMPAGE
RAP
RECORD
RIDDLE
RIFT
ROCK
ROW
RULE OUT
SACK
SAGA
SCARE
SCRAP

SEEK
SEIZE
SET TO
SHED
SLAM
SLASH
SLATE
SLAY
SLUMP
SNATCH
SOAR
SPARK
SPLIT
SPREE
STAKE
STORM
STORM OUT OF
STUN
SURGE
SWAP
SWAY
SWITCH
SWOOP
THREAT
TOLL

TOP
TORY
TRIO
TROOPS
UK
ULSTER
UN
URGE
US
VAT
VOW
WALK OUT
WED

Top Home Contents Resources Links Editors History