Corpus Tools and Methods, Today and Tomorrow: Incorporating Linguists' Manual Annotations

Nicholas Smith
School of English, Sociology, Politics & Contemporary History,
University of Salford, Manchester, UK

Sebastian Hoffmann
Department of Linguistics & English Language, Lancaster University, Lancaster, UK

Paul Rayson
Department of Computing, Lancaster University, Lancaster, UK

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011

Correspondence:
Nicholas Smith, School of English, Sociology, Politics & Contemporary History, University of Salford, Manchester M5 4WT, UK.
E-mail:
n.smith@salford.ac.uk

Abstract

Today's corpus tools offer the user a wide range of features that greatly facilitate the linguistic analysis of large amounts of authentic language data (e.g. frequency distributions, collocations, keywords, etc.). However, these tools typically fail to address the fundamental need of the linguist to add interpretive information to a concordance or query result, by coding individual concordance lines for structural, functional, discoursal, and other features in a flexible way. The ability to add such qualitative data is indispensable to a fuller understanding of the phenomenon under investigation as it allows the linguist to produce more rigorous descriptions—and theories—about language in use.

Our article has two aims: first, to assess the merits and drawbacks of existing solutions, by surveying what can be achieved using state-of-the-art corpus tools and generic database software; second, we draw up a set of desiderata and recommendations for the incorporation of flexible encoding features into future corpus tools. We describe an initial step in this direction, with a recent enhancement to the *BNCweb* corpus analysis software. More generally, we hope our suggestions will lead to linguists and software developers working together more closely to ensure that the needs of the former are provided for by the available technology.

1 Introduction

Standard corpus-processing tools currently offer a wide range of features for the automatic analysis of corpus data (for example, advanced sorting, collocations, *n*-grams, and distributions across metatextual categories). However, the value of these types of analysis varies considerably as a function of the accuracy and specificity of the query run over the corpus, and the quality and extent of text data and their annotations. As most users of corpora have discovered, corpus searches typically

© The Authors 2008. Published by Oxford University Press on behalf of ALLC and ACH. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

doi:10.1093/llc/fqn004 Advance Access Published on 17 March 2008

retrieve irrelevant instances—sometimes a large proportion of the data set—and unless such instances are discarded, they have the potential to distort the linguistic findings. Perhaps even more importantly, for many research questions the level of annotation provided in existing corpora is inadequate, and linguists therefore often need to categorize query results according to their own classification criteria.

Considering this pervasive need for manually manipulating the output of corpus searches, one would expect modern corpus tools to offer abundant support to linguists in this endeavour. However, as we will demonstrate in some detail, this support is patchy, and nowhere is an entirely satisfactory solution provided. Furthermore, the technical and methodological aspects of categorizing corpus data receive very little attention in manuals and introductory textbooks on corpus linguistics, and novices in the field of corpus-based studies may therefore fail to appreciate its fundamental importance.

Our article pursues several aims: first, we offer some background information concerning the manual annotation of linguistic searches and briefly review the existing literature; second, we present a survey of a number of state-of-the-art programs, widely used by linguists, to undertake manual linguistic annotation of concordance data. This part of our discussion is particularly geared towards linguists who may until now have preferred the traditional approach of carrying out their subsequent analysis of concordance listings on paper. While we are aware that such an approach can have its benefits, we would like to present the computer-based approach as an equally valid, and certainly much more powerful and flexible method of analysing data. However, our overview will also show that the currently available tools have some serious limitations and drawbacks, and these will be discussed and evaluated in turn. In the final part of our article—which is particularly aimed at software programmers—we proceed to outline some of the issues that need addressing in the development of future corpus tools. As a first step in this direction, we present a newly implemented feature in *BNCweb*, a web-based interface to the British National Corpus (BNC), which allows users to re-import a set of results that have previously been exported from the tool, and manually categorized in a database application.

2 Background

2.1 Top-down, concordance-based annotation

One useful way of looking at corpus annotation is in terms of a bottom-up versus top-down distinction. In the bottom-up approach, the linguist (or, more usually, a team of linguists) carries out a close manual inspection of all the texts in a corpus, coding instances of linguistic features as they are encountered (possibly drawing on a particular theoretical framework). In the top-down approach, the linguist conducts a more focused enquiry of one or more linguistic items, retrieving them by means of a concordance query, and subsequently marking, for each attested example, anything deemed to be significant (or potentially significant) for characterizing the behavior of that item.[1] It is this top-down, concordance-based approach that is the focus of our article.

2.2 Why manually annotate a concordance?

The benefits of manual annotation of concordance lines can very quickly be demonstrated with the help of a practical example. Let us imagine that someone wishes to explore the use of tag questions in British English. In an ideal situation, the researcher would have access to a corpus where all tag questions have been pre-annotated by the compilers of the corpus.[2] If this is not the case, a search strategy has to be developed that retrieves all—but ideally only—instances of tag questions on the basis of a set of formal criteria. At first, this would seem like a fairly straightforward exercise since the form of tag questions can be well defined: they consist of an auxiliary (a form of *be*, *do*, *have* or a modal), followed by an optional negative element and a personal pronoun, *there*, or *one* (e.g. *didn't she*, *isn't it?* or *is there?*). Furthermore, it would be a reasonable assumption that tag questions

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011

predominantly occur in utterance-final position. However, it is in fact very easy to find instances of utterance-internal tag questions. As a case in point, consider example (1):

(1) And right on the almost on the final whistle just before United scored in injury time, I think mid-fielder Martin Cool got in a very good volley *didn't he* from some distance, but it really was whistling toward goal? (BNC:KS7:744)

A retrieval strategy that relies on a sentence- or utterance-final position—e.g. by searching for auxiliary/modal–personal pronoun sequences that are immediately followed by a question mark—will thus miss at least some relevant tag questions and may therefore give a skewed picture of actual tag question usage. However, if the immediately adjacent question mark is omitted from the search pattern, a large number of irrelevant instances will also be retrieved. Two typical examples are shown in (2) and (3):

(2) I mean *are you* talking about a hundred and fifty? (BNC:F7J:360).

(3) The first thing he *did he* made friends amongst the young men in the college. (BNC:HE3:91).

In example (2), *are you* is the beginning of a regular yes–no question and in example (3), there is in fact a clause boundary between *did* and *he*.

It is of course possible to optimize the precision of this type of search by specifying additional criteria for retrieval. For example, it would make sense to automatically exclude instances of potential tag questions if they are immediately preceded by a *wh*-word (e.g. *why don't you do this?*), and the same action is warranted if a verb immediately follows the pronoun (e.g. *doesn't he like cheese?*). However, such restrictions are often problematic because they are liable to reduce recall, i.e. they discard instances that should have been included in the analysis. At some stage, it will therefore normally be necessary for the linguist to distinguish between relevant and irrelevant instances by way of a manual analysis.

Besides removing false positives, manual annotation is crucial in identifying relevant parameters of linguistic variation. Although modern corpora like the BNC are richly annotated and may provide the researcher with various types of metatextual information, these pre-existing categories are often not fully adequate to meet the demands of the research question at hand and they may therefore have to be complemented with additional levels of (manual) annotation. For example, in the case of tag questions, the researcher may wish to look at the issue of polarity. Thus, a tag question can have reversed polarity—positive anchor/negative tag or negative anchor/positive tag—as in *He loves that, doesn't he?* or *You don't like cheese, do you?* or constant polarity—usually positive/positive—as in *This is dangerous, is it?* As in the case of the retrieval of tag questions mentioned above, this type of distinction can to some extent be determined by automated measures—e.g. by trying to determine the polarity of the anchor clause by checking for the existence or absence of negation. However, unless the corpus is grammatically parsed, any such automated detection is likely to result in partially erroneous classifications; a manual annotation of individual instances will produce much more reliable results. Furthermore, manual annotation is likely to be the only solution if a researcher wishes to investigate the different pragmatic functions of tag questions (cf. Tottie and Hoffmann, 2006, who distinguish between informational, confirmatory, facilitating, attitudinal, peremptory, and aggressive tag questions), as the fluid relationship between form and function will render any attempt at determining pragmatic functions via formal criteria highly ineffective.

The example of tag questions is by no means unusual. If one selects at random from the topics that have occupied corpus linguists' minds over the years—topics in grammar (e.g. passives, nominalizations), discourse/pragmatics (e.g. hedges, interjections), or lexical semantics (e.g. polysemy, synonymy)—it is difficult to find one that does not require (1) a degree of manual filtering to obtain a clean and full set of corpus instances, and (2) further annotation to characterize its uses adequately in different linguistic and social contexts.

Naturally, not even manual annotation will always be trouble free. A number of issues need to

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011

be borne in mind, such as inter-rater consistency (the degree to which different linguists agree about a given analysis) or intra-rater consistency (the degree to which the same linguist is consistent in his or her own analysis). One further consideration in this regard is that it is usually desirable not only to record linguistic features but also to document metadata about the annotation *process* itself (cf. Leech, 2005). This aids replicability and consistency of coding, and also assists researchers when they return to difficult examples at a later time. For example, such meta-annotation may include the following information:

- explicit documentation of the criteria which were used to define each category that has been applied, especially if they are complex or controversial;
- information about who coded the data and when; and
- the level of confidence of the researchers about their judgement.

In sum, there is a strong case for manual annotation of concordance lines to be treated as a central component of most linguistic research projects that use corpora. It is often an essential step in cleaning up data that has been retrieved by way of automated search strategies and it forms the basis for more complete linguistic descriptions that are in turn key to subsequent interpretations and theoretical explanations.

2.3 A brief overview of the relevant literature

The fundamental importance that we have just attributed to the manual annotation of concordance lines is not fully reflected in the existing literature. It is interesting to note that in a whole volume of papers entitled *Corpus Annotation* (Garside *et al.*, 1997), extensive coverage is given to automated and manual techniques of bottom-up linguistic annotation, but top-down approaches do not feature.

The concept of manually annotating concordance lines is of course not new, and several publications have alluded to its relevance, to varying degrees. For example, de Haan (1984) introduces the term 'problem-oriented tagging' and exemplifies its application with the study of post-modifying clauses in the English noun phrase. He argues that the effort of adding such annotations is warranted because they can provide a much greater level of detail on the 'problem' in question than any pre-existing annotations in the corpus. He also shows how manually added classification codes can form the basis for statistical analyses. Although de Haan describes problem-oriented tagging as *ad hoc*, in that it will only be used for the project for which it was devised, he adds optimistically that the annotation scheme 'can be modified easily to serve the needs of other investigators in different research projects' (p. 123).

Ten years later, Kirk (1994) revisits many of these issues in a detailed discussion of the advantages and drawbacks of concordances. In his view, concordances 'do little more than rearrange the selected data as a special kind of list' (p. 261). However, when the raw and non-interpreted data are subjected to classification, they can come to be used in an 'intelligent' way (p. 261). For this purpose, Kirk suggests that the concordance be imported into a database application. In this way, manually added classifications can be used to filter or sort the concordance in very flexible ways. In other words, the concordance is no longer a mere list: 'although the database is merely reordering the analysis keyed in, it is reorganizing the analysis in its own terms and is thus able to reflect back to the analyst the actual distribution of the data' (p. 263).

A similar approach to de Haan's and Kirk's is described in Tottie *et al.* (1984), who report on their method of investigating negative sentences in which each instance is annotated with values belonging to 35 distinct categories. But in contrast to the aforementioned annotation processes, which appear to have been entirely manual, Tottie *et al.* highlight the fact that the manual annotation of linguistic data can be simplified and accelerated considerably with the help of a specialized computer tool that interacts with the linguist and which automatically provides the most probable value for each category. This is an important reminder that manual and

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011

automated approaches to annotation need not be discrete activities, but may go hand in hand.

To the best of our knowledge, there are no more recent papers devoted to the topic of manual annotation of concordance lines. Possible reasons for this are that the topic has been exhausted, or that it is perceived as too trivial to address in a research article. However, even if either of these were true, one would expect the practicalities of manual annotation to be covered in introductory manuals and handbooks in corpus linguistics. It is to these that we now turn.

The topic of manual post-editing of automatically retrieved corpus data does not feature at all in Kennedy's (1998) introduction to corpus linguistics, even though many of the case studies he summarizes will no doubt have relied on such a manual post-editing process.[3] And while McEnery and Wilson (2001, p. 69) briefly mention problem-oriented tagging as a 'very important' (non-exhaustive) type of corpus annotation, they state that the need for manual annotation arises only 'occasionally' (p. 69). Furthermore, they do not offer any explicit guidance on the individual steps involved. The reason given for this choice is that they consider the process to be too specific to an individual research question to offer any further generalizations.

Biber *et al.* (1998, pp. 71, 73) also discuss the fact that data retrieved by automatic techniques may require hand editing, and they point out that this process is more conveniently carried out with the help of a specialized annotation tool. In their case, they refer to 'an interactive text analysis program' (p. 112) which is unfortunately not further specified. Their illustrative example concerns the classification of noun phrases with respect to their informational characteristics (e.g. whether their information status is 'given' or 'new', or whether their reference is 'anaphoric', 'exophoric', or 'inferrable'). The interactive tool first automatically detects possible noun phrases—essentially by looking for nouns and pronouns in a POS-tagged corpus—and then presents an initial analysis to the user. This analysis can then be confirmed or, if necessary, corrected on the basis of a list of possible choices.

Meyer (2002, pp. 97–8, 111), too, mentions 'problem-oriented tagging' and he exemplifies its application with a sample analysis of pseudo-titles (e.g. ***fugitive financier*** *Robert Vesco*, ***linguist*** *Noam Chomsky*). He highlights the problem that non-mnemonic codes (e.g. numerical values) 'increase the likelihood that errors will occur during the encoding process' (p. 111), and like Biber *et al.* (1998), he recommends the use of a specialized annotation tool. The program of his choice is again a privately developed one (called *Tagger*, cf. Meyer and Tenney, 1993), which displays 'both the actual variables being studied in a particular corpus analysis, and the values associated with each variable' (p. 113).

Finally, McEnery *et al.* (2006) devote rather more attention to problem-oriented annotation, and their textbook is the only introduction to corpus-based language studies that explicitly guides its readers through the practical steps necessary for manually annotating data. The authors focus both on the time-consuming, exhaustive annotation of all texts in a corpus—i.e. the bottom-up approach—as well as on the tagging of individual concordance lines—i.e. the top-down approach. They exemplify the practical application of both approaches in the context of analysing errors in learner corpora. Interestingly, an evaluative distinction is drawn between the two approaches: whereas the former is presented as a 'proper way', the latter is dubbed 'a dirty way' of annotating a corpus (p. 253). This negative evaluation is presumably based on an assumption that the annotations which are manually added to concordance lines cannot be reused at a later stage.

In sum, the coverage of this type of annotation to date is rather patchy. While the advantages of manual annotation of concordance lines have been clearly documented in a number of publications, its treatment in most introductory textbooks is not as extensive as might be expected. As a result, linguists who are new to corpus-based studies may fail to be sensitized to its fundamental value for many types of linguistic analysis. We therefore believe that the time is now ripe for a reassessment of the existing processes and tools.

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011

3 Currently Available Tools and Strategies for Manual Analysis of a Corpus Query Result

This section will present a survey of some currently available tools for the manual annotation of concordance lines. In a first step, we will focus on 'in-corpus-tool categorization' by reviewing the functionality offered by three state-of-the-art packages. This will then be followed by a discussion of the merits and drawbacks of employing external software for this purpose. More specifically, we will explore the viability of two widely used packages, namely the spreadsheet application Microsoft *Excel* and the relational database *FileMaker Pro*.

We should emphasize again that we are exclusively concerned with the top-down approach to manual annotation. As a result, a number of tools that are specifically designed to handle the task of qualitative text analysis by social scientists (Alexa and Zuell, 2000) will not receive any attention. Although tools such as *Dexter* and *ATLAS.ti* offer considerable flexibility and functionality in incorporating manual annotations, they offer no facilities (or extremely limited facilities) for concordancing.[4]

3.1 Annotating within the corpus tool itself

In this section, the following three tools will be reviewed:

- *MonoConc* Version 2.2. This is a Windows-based PC concordance tool with a range of powerful features (Advanced Search: Full Regular Expression search, Part-of-Speech Tag Search, Collocations), published by Athelstan.[5]
- *WordSmith Tools* Version 5. This recently updated Windows-based package is described as an advanced set of tools providing 'an integrated suite of programs for looking at how words behave in texts' (Scott, 2007).[6]
- *BNCweb (CQP-edition)* (Hoffmann and Evert, 2006; Lehmann *et al.*, 2000).[7] BNCweb is a web-based client program for searching and retrieving lexical, grammatical, and textual data from the BNC. Originally, it relied on the BNC server program SARA, whose functionality was extended by the integration of *MySQL*, a very fast and powerful SQL database server. A new version integrated with *CQP* has recently been completed. *CQP* is the corpus query processor originally integrated into *IMS Corpus WorkBench*[8] (Christ, 1994) and its graphical user interface (*Xkwic*).

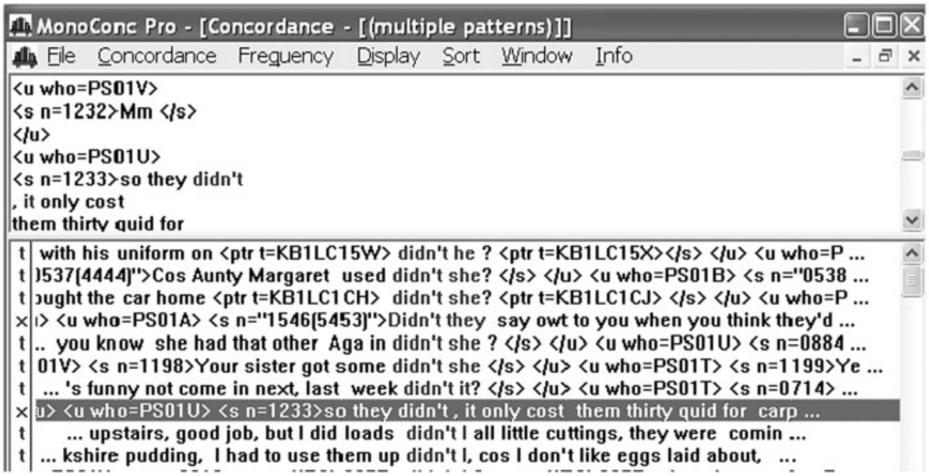
For general reviews of these packages, either individually or in comparison to others, see Lee and Rayson (2000), Hockey (2001), Reppen (2001), Rayson (2003, pp. 74–84), Roberts *et al.* (2006), Ari (2006) and Wiechmann and Fuhs (2006).[9] Given the aims of the present article, our focus is solely on their capacity to support manual corpus annotation.

3.1.1 *MonoConc 2.2*

Figure 1 displays the result of a search of the BNC Sampler corpus (spoken demographic component) for *didn't* + personal pronouns (*I*, *you*, *he, she, it*, etc.) as presented by *MonoConc 2.2*. Manual annotations can be entered via a pop-up menu that appears when the user right-clicks an example in the concordance listing. This function (which *MonoConc* refers to as 'Assign letter') offers the user just one field for inserting annotations. Within that field a single letter code can be assigned to each concordance hit: here 't' represents a valid case of a tag question, and 'x' represents 'not a tag question'. In other words, the classification is restricted to a maximum of 26 annotation values. Although in principle it is possible to devise a compact coding system that would register several levels of categorization—e.g. 'x' for 'not a tag question', 'p' for 'positive polarity', 'n' for 'negative polarity', etc., in practice, it is sometimes difficult to choose a code with mnemonic value.

The annotation field can be sorted, so that concordance lines sharing a common feature (e.g. positive polarity uses of tag questions) can be grouped together. This makes it easier to assess patterns of behaviour, and to count relevant instances of the feature of interest. However, *MonoConc* does not readily allow the user to examine the annotation field in other ways, e.g. to find distribution characteristics, or collocations. This can only be achieved by a workaround method, i.e. if the user explicitly deletes the cases that are not wanted. This workaround can

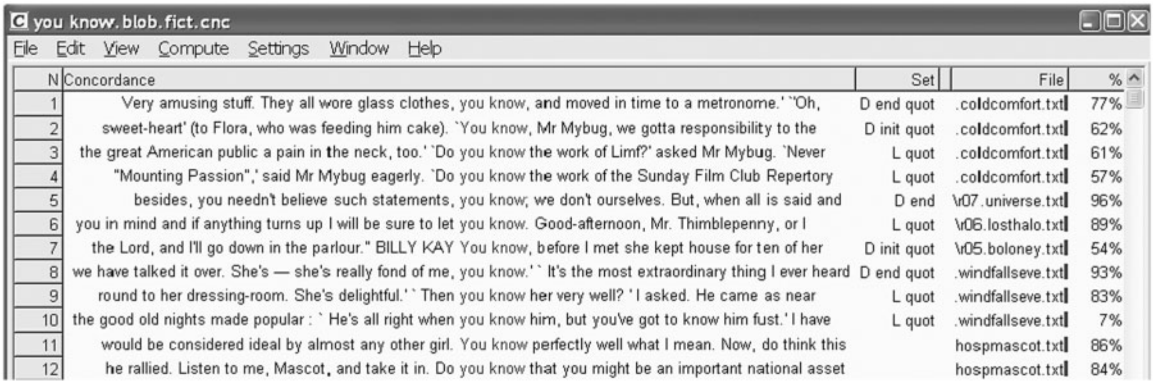
Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011



```
<u who=PS01V>
<s n=1232>Mm </s>
</u>
<u who=PS01U>
<s n=1233>so they didn't
, it only cost
them thirty quid for
```

```
t  with his uniform on <ptr t=KB1LC15W> didn't he ? <ptr t=KB1LC15X></s> </u> <u who=P ...
t )537(4444)">Cos Aunty Margaret  used didn't she? </s> </u> <u who=PS01B> <s n="0538 ...
t ɔught the car home <ptr t=KB1LC1CH>  didn't she? <ptr t=KB1LC1CJ> </s> </u> <u who=P ...
x ı> <u who=PS01A> <s n="1546(5453)">Didn't they  say owt to you when you think they'd ...
t ..  you know  she had that other  Aga in didn't she ? </s> </u> <u who=PS01U> <s n=0884 ...
t  01V> <s n=1198>Your sister got some didn't she </s> </u> <u who=PS01T> <s n=1199>Ye ...
t   ... 's funny not come in next, last  week didn't it? </s> </u> <u who=PS01T> <s n=0714> ...
x u> <u who=PS01U> <s n=1233>so they didn't , it only cost  them thirty quid for  carp ...
t       ... upstairs, good job, but I did loads  didn't I all little cuttings, they were  comin ...
t  ... kshire pudding,  I had to use them up didn't I, cos I don't like eggs laid about,   ...
```

Fig. 1 User-defined categorizations in *MonoConc* Version 2.2



N	Concordance	Set	File	%
1	Very amusing stuff. They all wore glass clothes, you know, and moved in time to a metronome.' `'Oh,	D end quot	.coldcomfort.txt	77%
2	sweet-heart' (to Flora, who was feeding him cake). `You know, Mr Mybug, we gotta responsibility to the	D init quot	.coldcomfort.txt	62%
3	the great American public a pain in the neck, too.' `Do you know the work of Limf?' asked Mr Mybug. `Never	L quot	.coldcomfort.txt	61%
4	"Mounting Passion",' said Mr Mybug eagerly. `Do you know the work of the Sunday Film Club Repertory	L quot	.coldcomfort.txt	57%
5	besides, you needn't believe such statements, you know; we don't ourselves. But, when all is said and	D end	\r07.universe.txt	96%
6	you in mind and if anything turns up I will be sure to let you know. Good-afternoon, Mr. Thimblepenny, or I	L quot	\r06.losthalo.txt	89%
7	the Lord, and I'll go down in the parlour." BILLY KAY You know, before I met she kept house for ten of her	D init quot	\r05.boloney.txt	54%
8	we have talked it over. She's — she's really fond of me, you know.' ` It's the most extraordinary thing I ever heard	D end quot	.windfallseve.txt	93%
9	round to her dressing-room. She's delightful.' ` Then you know her very well? ' I asked. He came as near	L quot	.windfallseve.txt	83%
10	the good old nights made popular : ` He's all right when you know him, but you've got to know him fust.' I have	L quot	.windfallseve.txt	7%
11	would be considered ideal by almost any other girl. You know perfectly well what I mean. Now, do think this		hospmascot.txt	86%
12	he rallied. Listen to me, Mascot, and take it in. Do you know that you might be an important national asset		hospmascot.txt	84%

Fig. 2 User-defined categorization in *WordSmith*

be problematic, e.g. if there are some cases which the user is uncertain how to categorize.

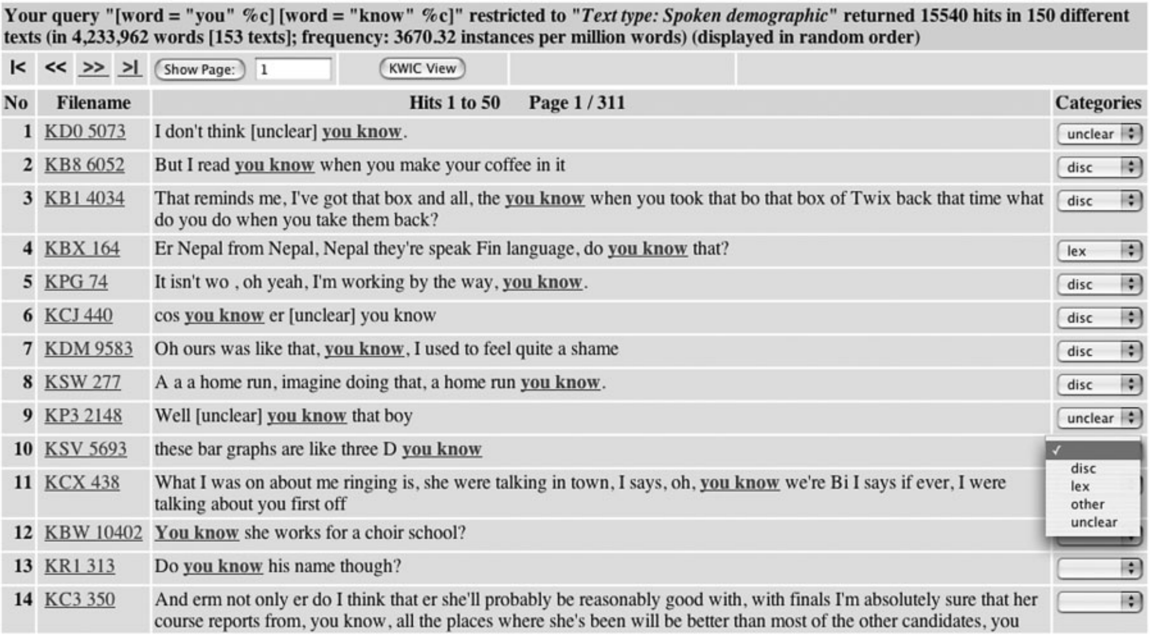
3.1.2 *WordSmith Tools Version 5*

WordSmith was one of the first programs, if not the first, to offer the user the opportunity to attach qualitative codes to concordance lines. Not surprisingly then, Version 5 of the package offers a little more flexibility than *MonoConc* Version 2.2 in this respect.

Like the 'Assign letter' function of *MonoConc*, the 'Set' function of *WordSmith* inserts a single letter code in a field adjacent to the concordance text. (This code is entered simply by typing the letter while in concordance view.) Moreover, by double-clicking in the 'Set' field, longer—and therefore more mnemonic—labels can be entered, containing spaces if desired. This opens up the possibility of entering multiple levels of annotation. Figure 2 illustrates this with an annotated concordance of *you know*: the codes here represent lexical ('L') and discourse marker ('D') uses, as well as whether the example is from quoted (direct) speech ('quot'), and its position in the utterance as initial ('init') or final ('end').

In practice, however, the fact that these annotations are stored in a single field means that it is difficult to exploit and analyse them individually.

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011



Your query "[word = "you" %c] [word = "know" %c]" restricted to "*Text type: Spoken demographic*" returned 15540 hits in 150 different texts (in 4,233,962 words [153 texts]; frequency: 3670.32 instances per million words) (displayed in random order)

No	Filename	Hits 1 to 50 Page 1 / 311	Categories
1	KD0 5073	I don't think [unclear] **you know**.	unclear
2	KB8 6052	But I read **you know** when you make your coffee in it	disc
3	KB1 4034	That reminds me, I've got that box and all, the **you know** when you took that bo that box of Twix back that time what do you do when you take them back?	disc
4	KBX 164	Er Nepal from Nepal, Nepal they're speak Fin language, do **you know** that?	lex
5	KPG 74	It isn't wo , oh yeah, I'm working by the way, **you know**.	disc
6	KCJ 440	cos **you know** er [unclear] you know	disc
7	KDM 9583	Oh ours was like that, **you know**, I used to feel quite a shame	disc
8	KSW 277	A a a home run, imagine doing that, a home run **you know**.	disc
9	KP3 2148	Well [unclear] **you know** that boy	unclear
10	KSV 5693	these bar graphs are like three D **you know**	
11	KCX 438	What I was on about me ringing is, she were talking in town, I says, oh, **you know** we're Bi I says if ever, I were talking about you first off	
12	KBW 10402	**You know** she works for a choir school?	
13	KR1 313	Do **you know** his name though?	
14	KC3 350	And erm not only er do I think that er she'll probably be reasonably good with, with finals I'm absolutely sure that her course reports from, you know, all the places where she's been will be better than most of the other candidates, you	

Fig. 3 The categorization feature of *BNCweb (CQP-edition)* at work

Manually annotated concordances can be saved for later re-use. In addition, *WordSmith* offers the option of inserting user-defined categories as tags into the original source texts.[10] An advantage of this feature is that manual annotations can easily be made available to other users of the data. On the basis of these tags, it possible to perform flexible searches of the corpus which combine pre-existing and user-defined levels of annotation (e.g. to retrieve all instances of *you know* that are tagged as discourse markers and which are uttered by speakers aged 25–35 years). However, at the time of writing, no simple solution appears to exist to combine several of these manually added sets of categories in a single search. As in *MonoConc*, users are therefore still effectively restricted to a single manual annotation field.[11]

3.1.3 *BNCweb (CQP-edition)*

With *BNCweb (CQP-edition)*, again the user is effectively limited to a single type of manual annotation, since only one field is provided for this purpose. By default, the values 'unclear' and 'other' are added to the values defined by the user. Labels for values can be up to 50 characters long, but they may only contain letters of the alphabet, numbers, and the underscore character ('_'). Manually annotated data are automatically saved, and remain available for corrections or further annotation at a later stage. While it is possible to add new values to a categorization scheme, existing values cannot be deleted. Figure 3 illustrates a user's classification of the string *you know* as a discourse marker, as a lexical use of *know*, or as 'unclear'.

Having 'unclear' available as a standard option is invaluable for dealing with indeterminate cases of various sorts—vague cases, ambiguous cases, or cases where the user lacks confidence about their own decision making. However, there is currently no convenient way provided to indicate which of these types of indeterminacy is involved, should the user wish to distinguish them. Furthermore, no facilities are offered as yet for sorting the concordance on the basis of user-defined category values. In other words, the user cannot arrange the KWIC list with the discourse marker uses of *you know* appearing all together, followed by the lexical

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011

verb uses, then the 'other' uses (if there are any), and the unclear uses. It is, however, possible to divide the manually annotated data into separate saved query outputs. The user can then, for example, view all discourse marker uses in a single concordance and also apply the full set of post-processing features offered by *BNCweb* to this data (collocations, distributional statistics, etc.).

3.2 Evaluation of 'in-corpus-tool' annotation

The method described in Section 3.1 makes it possible for linguists to effectively 'clean up' their data—i.e. remove unwanted instances—as well as to perform basic classifications according to a limited set of values within one category. The key advantage of annotating concordances within the corpus tool is that the user always has access to the larger context of a concordance line; this is often essential for a correct classification of individual instances. Another important benefit is that once the data have been manually post-processed, the whole range of advanced functions of the corpus tool—e.g. collocations, keywords, *n*-grams, etc.—remains available for further linguistic analysis, although two of the programs (*MonoConc* and *BNCweb*) reviewed require a workaround—i.e. deletion of unwanted examples or division into separate saved query outputs—before these additional functions can be accessed. Finally, some researchers may also consider it an advantage that the use of only one piece of software needs to be learnt.

In contrast, the main limitation of annotating concordances within corpus tools is that they generally lack the advanced features for querying and manipulating data that are to be found in standard spreadsheet and database packages. Only with the latter can the user, for example, filter data by multiple criteria; validate his/her annotations (e.g. preventing typos such as 'pasive' from appearing); incorporate annotations consisting of free-flowing text (including annotations about the process of editing, e.g. 'need to check annotation of feature *X* for consistency'; 'example most likely given, but possibly new'; or 'good example for paper'); and perform advanced arithmetic functions.

3.3 Exporting the concordance to external software

Given the limitations just mentioned, it is often more convenient to save the set of concordance lines to the user's hard disk, and then import the data to a database or spreadsheet application such as *FileMaker*, *Access*, *Excel*, etc. These applications support a number of data formats, the simplest one being a tab-delimited file.

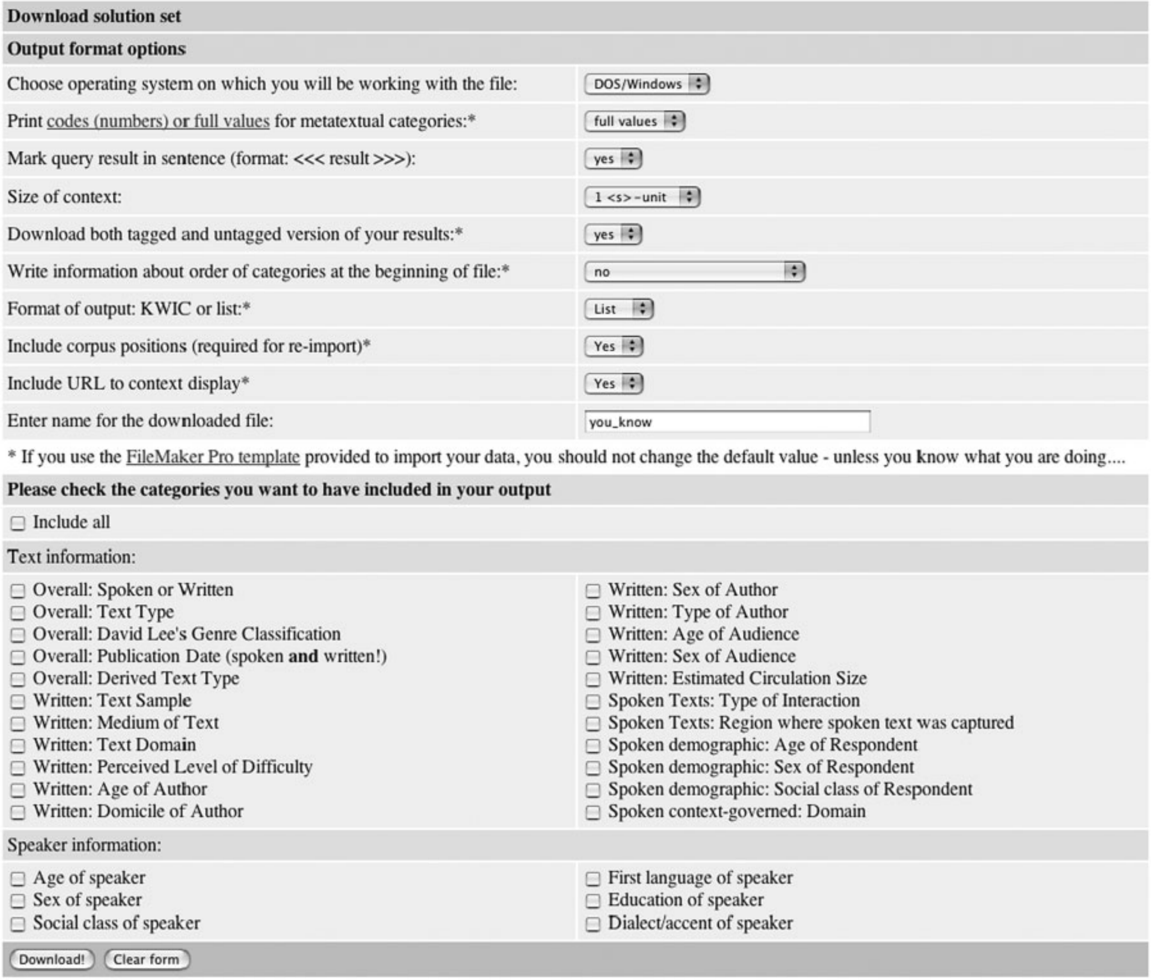
In addition to the actual concordance lines, it is of course essential for referencing purposes (e.g. when reporting the results of one's analysis) to export information about the position of the data in the corpus, for example, a text-identifier and sentence number. This information may also be required to link to any of the meta-textual information—such as the age or educational level of the speaker—that has been pre-annotated in the corpus (see below). Alternatively, such meta-textual information can be exported together with the concordance lines and imported into the database tool. The choices available to users of *BNCweb* are shown in Fig. 4.[12] Among the various options available, the user can also export a hyperlink that can be used to access the larger context of the concordance line in the corpus via any web browser.

A simple but very effective environment for annotating the concordance data is the flat database format provided in spreadsheet packages such as *Microsoft Excel*.[13] Figure 5 shows a number of types of annotation that have been entered manually alongside a concordance. The corpus in question was already in POS-tagged form, and the query was for any form of BE, followed by *being*, followed by a past participle. This found instances of the progressive passive construction with very high recall.

Sequences that were not examples of the progressive passive, such as the sequence *is being trimmed* in this example (4), were marked as spurious:

> (4) WHAT is the objection to utilising the old burial ground in Beaumont-fee, Lincoln, as a car park? It is untidy and, apart from what grass there *is being trimmed* now and then, it is not particularly well looked after. (LOB: B22).

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011



Download solution set

Output format options

Choose operating system on which you will be working with the file:	DOS/Windows
Print codes (numbers) or full values for metatextual categories:*	full values
Mark query result in sentence (format: <<< result >>>):	yes
Size of context:	1 <s>-unit
Download both tagged and untagged version of your results:*	yes
Write information about order of categories at the beginning of file:*	no
Format of output: KWIC or list:*	List
Include corpus positions (required for re-import)*	Yes
Include URL to context display*	Yes
Enter name for the downloaded file:	you_know

* If you use the FileMaker Pro template provided to import your data, you should not change the default value - unless you know what you are doing....

Please check the categories you want to have included in your output

□ Include all

Text information:

□ Overall: Spoken or Written
□ Overall: Text Type
□ Overall: David Lee's Genre Classification
□ Overall: Publication Date (spoken **and** written!)
□ Overall: Derived Text Type
□ Written: Text Sample
□ Written: Medium of Text
□ Written: Text Domain
□ Written: Perceived Level of Difficulty
□ Written: Age of Author
□ Written: Domicile of Author
□ Written: Sex of Author
□ Written: Type of Author
□ Written: Age of Audience
□ Written: Sex of Audience
□ Written: Estimated Circulation Size
□ Spoken Texts: Type of Interaction
□ Spoken Texts: Region where spoken text was captured
□ Spoken demographic: Age of Respondent
□ Spoken demographic: Sex of Respondent
□ Spoken demographic: Social class of Respondent
□ Spoken context-governed: Domain

Speaker information:

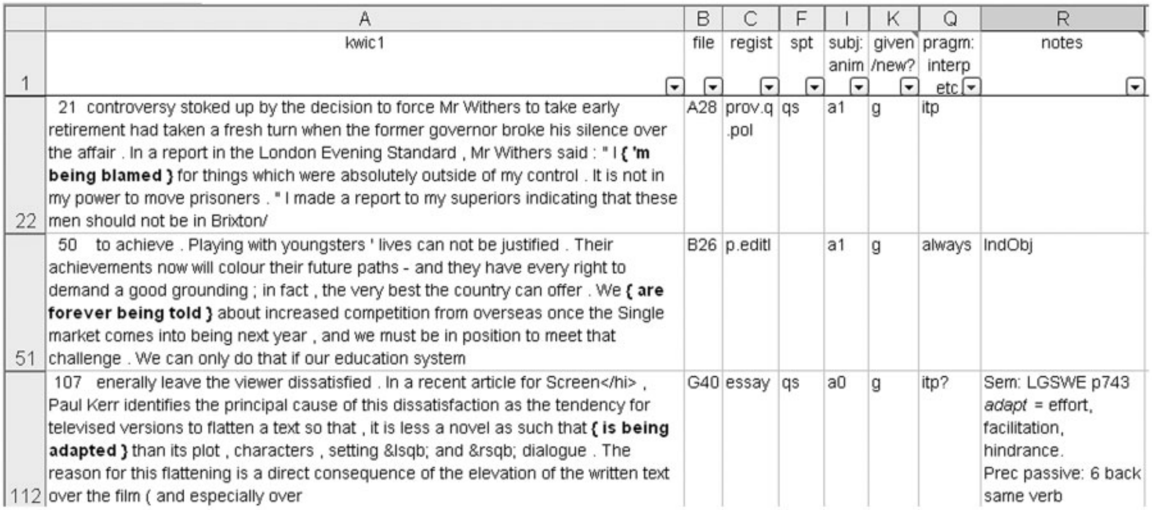
□ Age of speaker
□ Sex of speaker
□ Social class of speaker
□ First language of speaker
□ Education of speaker
□ Dialect/accent of speaker

Fig. 4 Options for exporting a concordance from *BNCweb (CQP-edition)*, including choices for metatextual categories

The spreadsheet shown in Fig. 5 has already been filtered to show only genuine cases of the progressive passive.[14] The column 'spt' indicates whether the example occurs in a marked category of speech and thought representation [e.g. 'qs' = 'quoted (i.e. direct) speech']. The next three columns indicate different linguistic characteristics of the retrieved instances: the (in)animacy of the subject ('a1' = animate, 'a0' = inanimate); the information status of the subject (i.e. whether it represents given or new information); and pragmatic aspects of usage (e.g. a metalinguistic or interpretative use of *I'm being blamed* in the first citation).

It is notable that in this particular data set, many examples are not amenable to discrete classification. Information status, for example, is notoriously difficult to assign (see e.g. Prince, 1981 for discussion). The user can reflect his/her uncertainty about the value to assign simply by using a question mark, e.g. '?' = 'completely unclear whether it is given or new', 'g?' = 'unclear, but more likely to be given'. Since a high degree of subjectivity is often involved in such decisions, it is a good idea—as an aide-memoire and a guide to others studying one's work—to document canonical examples of each type, including the unclear cases. (In *Excel*, such

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011



	A	B	C	F	I	K	Q	R
1	kwic1	file	regist	spt	subj: anim	given /new?	pragm: interp etc	notes
22	21 controversy stoked up by the decision to force Mr Withers to take early retirement had taken a fresh turn when the former governor broke his silence over the affair . In a report in the London Evening Standard , Mr Withers said : " I **{ 'm being blamed }** for things which were absolutely outside of my control . It is not in my power to move prisoners . " I made a report to my superiors indicating that these men should not be in Brixton/	A28	prov.q .pol	qs	a1	g	itp	
51	50 to achieve . Playing with youngsters ' lives can not be justified . Their achievements now will colour their future paths - and they have every right to demand a good grounding ; in fact , the very best the country can offer . We **{ are forever being told }** about increased competition from overseas once the Single market comes into being next year , and we must be in position to meet that challenge . We can only do that if our education system	B26	p.editl		a1	g	always	IndObj
112	107 enerally leave the viewer dissatisfied . In a recent article for Screen</hi> , Paul Kerr identifies the principal cause of this dissatisfaction as the tendency for televised versions to flatten a text so that , it is less a novel as such that **{ is being adapted }** than its plot , characters , setting &lsqb; and &rsqb; dialogue . The reason for this flattening is a direct consequence of the elevation of the written text over the film (and especially over	G40	essay	qs	a0	g	itp?	Sem: LGSWE p743 *adapt* = effort, facilitation, hindrance. Prec passive: 6 back same verb

Fig. 5 An annotated database of the progressive passive, using *Excel*

information can be added to another worksheet incorporated in the same file). The database can be easily filtered to retrieve cases with the respective probabilities (e.g. 'g*' will find both clear and probably given cases, while 'g?' will only display the latter of the two categories).

The last column contains various notes added by the linguist while coding the data; e.g. a note to check the *Longman Grammar of Spoken and Written English*, page 743, for discussion of the semantic categorization of the verb *adapt* (which is the main verb of the progressive verb phrase in this example). Such cases highlight the fact that annotation is not only a product, but also a *process*, often undergoing continuous revision and refinement through repeated readings of the data.[15]

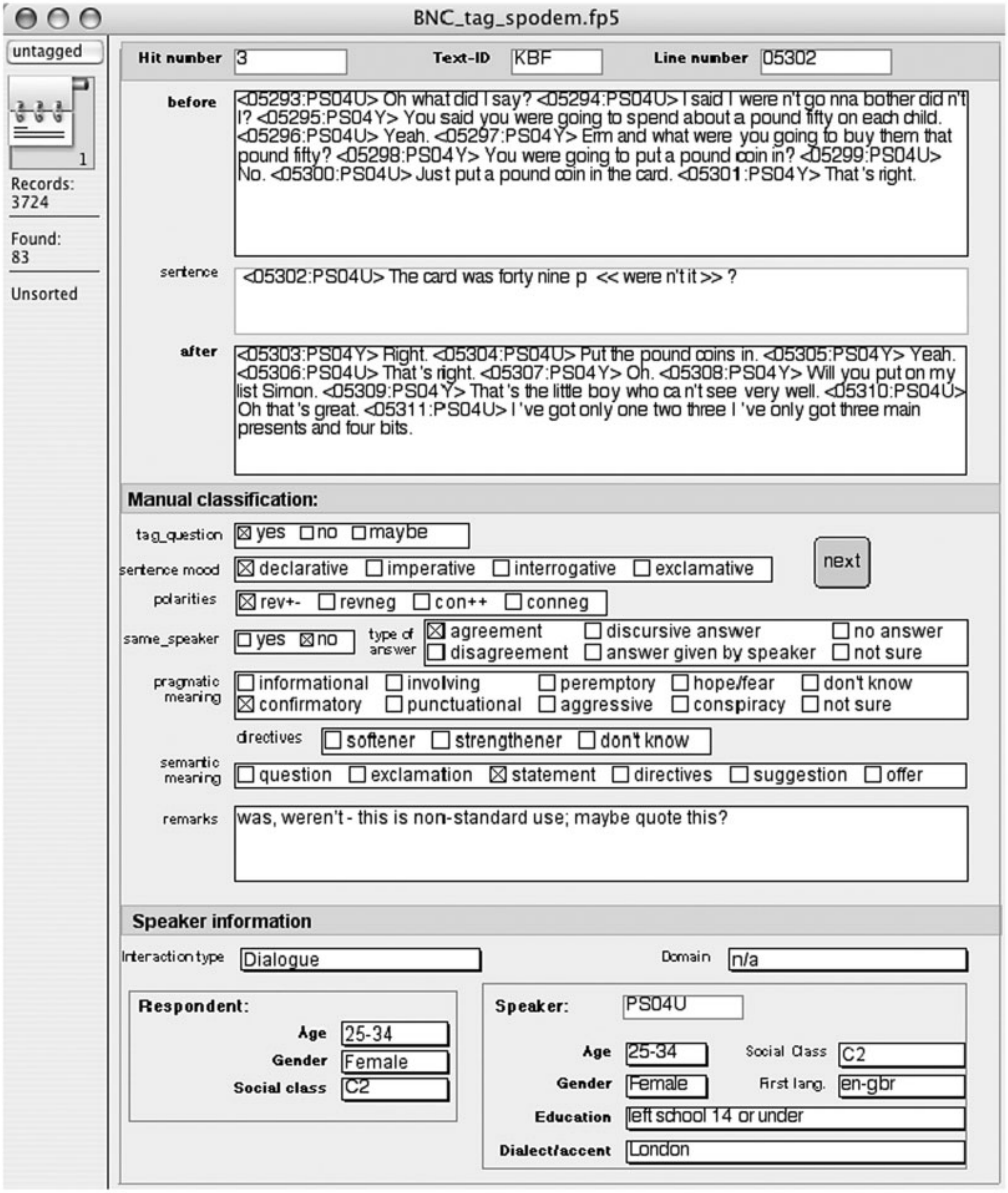
Still more extensive analytic capabilities are offered by relational database programs such as *FileMaker*.[16] Figure 6 displays a fairly elaborate database created in *FileMaker* to categorize sentences with tag questions in the spoken-demographic component of the BNC. At the top of the screen, three text fields contain the sentence in question and the preceding and following context. Below this, the researcher has set up a wide range of categories for manual analysis, including, for example, whether the sentence indeed contains a tag question, what kind of syntactic structure the sentence has, and what type of pragmatic function the tag question performs. Finally, the database mask also displays information about the speaker who has uttered the tag question. This information is part of the annotation provided in the BNC. Since *FileMaker* is a relational database, this type of (fixed) information can be stored in a separate database and linked into any newly created database by way of a common field—or set of fields—e.g. a speaker identification code or a text-ID and sentence number.

Once all entries in the database have been categorized, the user can filter the set of data in very flexible ways and use this information for the compilation of descriptive statistics. For example, it only takes a few clicks to find all declarative sentences uttered by female speakers up to the age of 34 years, which were classified as informational tag questions.

3.4 Evaluation of 'export-to-database' method

The method described in Section 3.3 clearly offers linguists much more flexibility in their manual annotations than what is currently offered by the in-corpus-tool approach described in Section 3.1. At the same time, exporting to a database also entails

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011



BNC_tag_spodem.fp5

untagged

Records: 3724

Found: 83

Unsorted

Hit number 3 | Text-ID KBF | Line number 05302

before: <05293:PS04U> Oh what did I say? <05294:PS04U> I said I were n't go nna bother did n't I? <05295:PS04Y> You said you were going to spend about a pound fifty on each child. <05296:PS04U> Yeah. <05297:PS04Y> Erm and what were you going to buy them that pound fifty? <05298:PS04Y> You were going to put a pound coin in? <05299:PS04U> No. <05300:PS04U> Just put a pound coin in the card. <05301:PS04Y> That 's right.

sentence: <05302:PS04U> The card was forty nine p << were n't it >> ?

after: <05303:PS04Y> Right. <05304:PS04U> Put the pound coins in. <05305:PS04Y> Yeah. <05306:PS04U> That 's right. <05307:PS04Y> Oh. <05308:PS04Y> Will you put on my list Simon. <05309:PS04Y> That 's the little boy who ca n't see very well. <05310:PS04U> Oh that 's great. <05311:PS04U> I 've got only one two three I 've only got three main presents and four bits.

Manual classification:

tag_question: ☒ yes ☐ no ☐ maybe

sentence mood: ☒ declarative ☐ imperative ☐ interrogative ☐ exclamative

polarities: ☒ rev+- ☐ revneg ☐ con++ ☐ conneg

same_speaker: ☐ yes ☒ no

type of answer: ☒ agreement ☐ discursive answer ☐ no answer ☐ disagreement ☐ answer given by speaker ☐ not sure

pragmatic meaning: ☐ informational ☐ involving ☐ peremptory ☐ hope/fear ☐ don't know ☒ confirmatory ☐ punctuational ☐ aggressive ☐ conspiracy ☐ not sure

directives: ☐ softener ☐ strengthener ☐ don't know

semantic meaning: ☐ question ☐ exclamation ☒ statement ☐ directives ☐ suggestion ☐ offer

remarks: was, weren't - this is non-standard use; maybe quote this?

Speaker information

Interaction type: Dialogue

Domain: n/a

Respondent:

Age: 25-34

Gender: Female

Social class: C2

Speaker: PS04U

Age: 25-34

Social Class: C2

Gender: Female

First lang.: en-gbr

Education: left school 14 or under

Dialect/accent: London

Fig. 6 Screen shot of a *FileMaker* database used to categorize sentences containing tag questions in the spoken-demographic component of the BNC

potentially serious drawbacks. First of all, the link between concordance and the original source text is normally severed. This means that the only way the user can consult the larger context of any concordance lines of interest is to return to the corpus tool and search for them manually. In the case of *BNCweb*, this problem has been circumvented by exporting, for each corpus citation,

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011

a hyperlink containing the URL address of the example. Naturally, however, such a solution can only work for web-based corpus tools.

A far more serious drawback of this method is that the advanced functions of the corpus tool—collocations, keywords, *n*-grams, etc.—can no longer be applied to the manually processed set of results. As a result, although the linguist may now be in possession of a wealth of relevant and accurate information, it may prove difficult to perform some of the standard quantitative analyses on the data.[17]

4 Requirements for Future Tools—and a First Step in this Direction

As Section 3 has shown, none of the existing tools offers linguists a fully satisfactory option for manually post-processing a corpus query result. There are at least two ways of working towards a solution to these issues. First of all, it would clearly be beneficial if the functionality of corpus tools were extended to give users greater flexibility in entering and manipulating manual annotations on concordance lines. However, this effectively means that developers of corpus tools would have to integrate the full feature-set of existing (relational) database applications. It seems to us that such an attempt at re-inventing the wheel is not the most efficient way to go forward.

Instead, we would like to see more enterprise in finding ways of exchanging the relevant data between existing applications. In this context, we need to stress that this exchange would have to work both ways. While it is no problem to export data to a database application, the solutions described in Section 3.3 do not make it possible to re-import any data back into the corpus tool—a necessary step if the linguist is to take advantage of the full feature-set of the corpus tool, using the manually cleaned and annotated set of concordance lines.

In order to achieve this type of interoperability, formats need to be established that are compatible with both types of applications, the corpus tools and the database programs. We can envisage three different ways of implementing this:

(1) The corpus tool and the database application use the same data source. Any changes made in one of the tools will be immediately available in the other tool.
(2) Manual annotations added in a database application are re-imported in a format that can be fully interpreted by the corpus tool. Manually added categories and their values become available to the user in the corpus tool in the same way as pre-existing corpus annotations.
(3) A common referencing system between the two types of tools is established, making it possible to compile a set of concordance lines in the corpus tool that corresponds to a manually cleaned and filtered set of instances in the database application.

Let us elaborate a little on each of these three suggestions. In principle, the first method is not very difficult to implement. For example, the corpus tool could use an SQL database such as *MySQL* to store query results. To some extent, this already is the case for *BNCweb* that heavily relies on *MySQL* for its automated post-processing features, such as the calculation of distribution statistics and collocation analyses. Although it would reduce the overall speed of queries made with *BNCweb*, it would require relatively few changes in the software to store all data relating to query results in SQL tables. The most recent version of *FileMaker Pro* (Version 9), in turn, can now directly access—both read and modify—data in SQL tables of various formats, including those used by *MySQL*. A researcher could thus perform a query with *BNCweb* and manually modify the resulting SQL table with *FileMaker Pro* by creating additional fields for classification. As long as *BNCweb* is adapted to recognize these newly added fields automatically, any manually added values could then be used in conjunction with the full existing feature set.[18] Unfortunately, the initial configuration steps to set up such a system on the client computer are far from trivial. They involve the installation of operating system-specific ODBC drivers and setting up a Data

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011

Source Name (DSN) to enable mutual communication between the *MySQL* server and the *FileMaker* client. Clearly, for such a system to acquire wide acceptance among corpus linguists, it would have to be implemented in a user-friendly fashion that does not require specialist knowledge and a major effort to be set up. Future developments in the area of shared data resources between corpus tools and database applications should thus particularly focus on the seamless integration of the two components.

The second of the three suggested solutions does not require a common data source for the corpus tool and the database, but instead relies on an exchange of the relevant information by way of an export/import procedure. Users would follow the method described in Section 3.3 to export a concordance to a database application and thereby make manual additions as required. When the modified data are re-imported into the corpus tool, all additional categories and their values would be recognized automatically in order for them to become available for further 'in-corpus-tool' use. Manually added categories and pre-existing annotation could then be combined in automated quantitative analyses. For this exchange of data, standoff annotation in the form of XML and a corresponding XML schema would seem to be the most suitable approach (Carletta *et al.*, 2005; Ide and Romary, 2004). A great advantage of this type of solution would be that manually added categories could easily be shared among different corpus users. However, the existing standards have been developed with the natural language processing or language engineering communities in mind and there is little awareness of these within the corpus linguistics community. Hence, such standards are currently not implemented within tools used by corpus linguists, nor in fact do their current implementations directly address the problems of manual annotation as described in our article. In any case, usability criteria would again have to feature high on the agenda of a software developer in the implementation of a solution that relies on standoff annotation for the exchange of information between different tools.

The third of our solutions is the easiest to implement. In fact, work on the current article has motivated changes to *BNCweb*, which represent a first successful step in the direction of user-friendly, bi-directional communication between corpus tools and databases. The implementation builds on the knowledge of how CQP creates an index of the BNC texts: each token in the corpus—i.e. a lexical item or a punctuation mark—is given a number. The count starts with the first token in the first text of the corpus and then proceeds through all texts in ascending order. Each token thus has a unique numerical identifier that is indexed by *CQP* and that represents the position of the token with respect to all other tokens in the corpus. The query result of a CQP search over the BNC internally returns a list of corpus positions that are then used to refer to the actual tokens.

When a query result is displayed, CQP optionally also outputs the corresponding corpus positions. If a user chooses to download a *BNCweb* concordance to the hard-disk (cf. Fig. 4 above), these numerical values can be exported alongside all other information. This data can then be imported into a database application of the user's choice as described in Section 3.3—currently, a *FileMaker Pro* template similar to the one displayed in Fig. 6 is provided for this purpose. Once the user has finished manually annotating the database, he or she may choose to filter the data, e.g. by restricting the display to instances that correspond to as many of the pre-existing or manually added categories as desired. The result of this filtering process will thus be a subset of the full query result.

The *FileMaker* template also contains a button named 'Export'. If the user clicks this button, *FileMaker* does not export all of the information that would be available for the selected entries. Instead, it only saves a list of the corresponding corpus positions to a file. The newly implemented feature in *BNCweb* subsequently allows the user to upload this list to the server. Using the 'undump' command in *CQP* (cf. Evert, 2005, pp. 38–9), a new concordance is then created on the basis of the corpus positions. This concordance will look exactly like the output of the initial query that was previously downloaded and imported into *FileMaker*. For example, the display still highlights *didn't he* as if the sequence had been matched by a

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011

normal search. The only difference is that the concordance now only contains those entries that were the result of the manual filtering in *FileMaker*. And as with any normal query result in *BNCweb*, the full set of features for automated analysis can be applied.

In an ideal situation, all manually added categories of annotation would be seamlessly integrated for use in the corpus tool. This would be the case if the first two of our three suggestions were to be implemented. The solution we have just described is clearly a compromise: for each filtered set of database entries, the process of re-importing into *BNCweb* has to be repeated. Although this is somewhat cumbersome, it still represents a significant step forward: at least for the BNC, linguists now have a simple way of applying automated quantitative analyses to manually cleaned-up and annotated data.

One drawback of the third solution outlined here is that the corpus and the newly added manual annotation levels remain fully separate. As a result, it is difficult to share the annotation scheme with other researchers in a convenient way (cf. the evaluation by McEnery *et al.*, 2006 of the 'top-down approach' as a 'dirty way'). For this reason, it is clearly worth moving beyond a solution that merely exchanges corpus positions—or similar methods of referencing—to one that fully integrates automated analysis with manual annotation.

As long as the internal mechanisms of saving query results are known, a procedure that exploits corpus positions as a way of referencing could easily be implemented for other corpus tools. Instead of corpus positions, these tools may rely on other ways of establishing unique references to elements in a corpus. We would therefore recommend that developers of corpus tools document their internal referencing mechanisms and implement the functionality to both export and (re-)import query results in the relevant format.

However, given that such diversity in referencing methods will require specific solutions for different software packages, we would also like to propose that developers of corpus tools strive towards the adoption of a common referencing system in order to support interoperability. This could be achieved by a system of unique corpus position identifiers (CPIs), which enables unique references to be generated for any word in a corpus text using pre-existing encoding. Hence, a corpus position identifier for the BNC could consist of the file identifier (file name), followed by sentence number and then word position, e.g. *KS9.45.6* for file KS9, sentence 45 and word position 6. Such a system of CPIs would enable a bridge between corpus software and the text itself and allow corpus users to share annotation on a word at position *KS9.45.6* without sharing the underlying text.[19] However, since such a system still relies on built-in markup, it appears to us that a potential candidate for a truly flexible and generic solution would be a system that relies on character offsets. Such an implementation would, among other advantages, be fully compatible with different writing systems or with layers of annotation that refer to token-internal segmentations. The only functionality that would have to be added to existing tools would thus be a mechanism to import and then translate character offsets to the already existing, tool-internal referencing system. One important disadvantage of a referencing system that relies on character offsets—or, for that matter, any type of CPI—is of course that any change in the original data will invalidate the existing character offsets and will therefore render any previously established standoff annotation obsolete. Clearly, further conceptual work is still required to optimize the interoperability between corpus tools and database applications.

5 Summary and Conclusion

The past decades have seen an enormous development in natural language processing and, as a result, automated annotation methods are nowadays a viable option to encode many linguistically relevant features of a corpus in an accurate way. However, for the foreseeable future, linguists and others wishing to maximize the benefits of searching a corpus will continue to need to devise and apply their own varieties of manual annotation in order to be able to clean up and enrich the data they retrieve. While it is of course possible to mark such

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011

annotations on paper, we would argue that inserting annotations in electronic form is nearly always to be preferred.

Two basic approaches can be taken: (1) annotating the categories within the corpus tool itself; and (2) exporting the query result to a general-purpose database or spreadsheet program. As our overview of available tools has shown, neither approach is as yet implemented in a way that maximizes the linguist's control over the data. In our discussion of future directions in the development of computer tools, we have advocated enhancing the latter approach, since it combines the already existing strengths of currently available tools. Most importantly, though, additional means have to be developed to allow the manually annotated data to be seamlessly re-imported into the corpus tool. We have briefly presented a newly implemented feature in *BNCweb* as a first step in this direction. We have also highlighted a number of issues that will require further attention from software developers.

Finally, we would exhort linguists and software developers to liaise more closely with each other in order to ensure that facilities for incorporating (inserting, manipulating, etc.) their manual annotations are optimized. This in turn should lead to improved descriptions, and interpretations, of empirical language data.

Acknowledgements

We would like to thank Stefan Evert, Geoffrey Leech, and Gunnel Tottie for their valuable comments on a draft version of this article. Michael Barlow and Mike Scott kindly provided test versions of their software, *Monoconc 2.2* and *Wordsmith 5*, respectively. We also acknowledge the support of The Leverhulme Trust (grant number F/00 185/J).

References

Alexa, M. and Zuell, C. (2000). Text analysis software: commonalities, differences and limitations: the results of a review. *Quantity and Quality*, **34**(3): 299–321.

Ari, O. (2006). Review of three software programs designed to identify lexical bundles. *Language Learning and Technology*, **10**(1): 30–7.

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

Carletta, J., McKelvie, D., Isard, A., Mengel, A., Klein, M., and Moller, M. B. (2005). A Generic Approach to Software Support for Linguistic Annotation Using XML. In Sampson, G. and McCarthy, D. (eds), *Corpus Linguistics: Readings in a Widening Discipline*. London/New York: Continuum International, pp. 449–59.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research (Budapest, 7–10 July 1994)*. Budapest, Hungary, pp. 23–32.

Evert, S. (2005). *The CQP Query Language Tutorial*. Unpublished manuscript. Available online from <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPTutorial/cqp-tutorial.pdf> (accessed 14 September 2007).

Garside, R., Leech, G., and McEnery, A. (eds) (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

de Haan, P. (1984). Problem-Oriented Tagging of English Corpus Data. In Aarts, J. and Meijs, W. (eds), *Corpus Linguistics: Recent Developments in the Use of Computer Corpora*. Amsterdam: Rodopi, pp. 123–39.

Hockey, S. (2001). Concordance Programs for Corpus Linguistics. In Simpson, R. C. and Swales, J. M. (eds), *Corpus Linguistics in North America: Selections from the 1999 Symposium*. Ann Arbor: University of Michigan Press, pp. 76–97.

Hoffmann, S. and Evert, S. (2006). BNCweb (CQP-Edition) – The Marriage of Two Corpus Tools. In Braun, S., Kohn, K, and Mukherjee, J. (eds), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt am Main: Peter Lang, pp. 177–95.

Hofland, K. (1991). Concordance Programs for Personal Computers. In Johansson, S. and Stenström, A.-B. (eds), *English Computer Corpora: Selected Papers and Research Guide*. Berlin: Mouton de Gruyter, pp. 283–306.

Hughes, L., & Lee, S. (eds) (1994). *CTI Centre for Textual Studies Resources Guide (1994)*. Oxford: CTI Centre for Textual Studies.

Ide, N. and Romary, L. (2004). International standard for a linguistic annotation framework. *Natural Language Engineering*, **10**(3–4): 211–25.

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London/New York: Longman.

Kirk, J. M. (1994). Corpus–Concordance–Database–VARBRUL. *Literary and Linguistic Computing*, **9**(4): 259–66.

Lancashire, I. (ed.) (1991). *The Humanities Computing Yearbook (1989–90)*. Oxford: Oxford University Press.

Lee, D. Y. W. and Rayson, P. (2000). Xkwic: a powerful concordancer for research. In *Workshop at Teaching and Language Corpora conference (TALC2000)*, 19–23 July 2000, Graz, Austria.

Leech, G. (2005). Adding Linguistic Annotation. In Wynne, M. (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, pp. 17–29. Available online from http://ahds.ac.uk/linguistic-corpora/ (accessed 20 June 2007).

Lehmann, H. M., Schneider, P., and Hoffmann, S. (2000). BNCweb. In Kirk, J. M. (ed.), *Corpora Galore*. Amsterdam: Rodopi, pp. 259–66.

McEnery, T. and Wilson, A. (2001). *Corpus Linguistics*. 2nd ed. Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R., and Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

Meyer, Ch. F. (2002). *English Corpus Linguistics. An Introduction*. Cambridge: Cambridge University Press.

Meyer, Ch. F. and Tenney, R. (1993). Tagger: An Interactive Tagging Program. In Souter, C. and Atwell, E. (eds), *Corpus-based Computational Linguistics*. Amsterdam: Rodopi, pp. 302–12.

Prince, E. (1981). Toward a Taxonomy of Given/New Information. In Cole, P. (ed.), *Radical Pragmatics*. New York: Academic Press, pp. 223–55.

Rayson, P. (2003). *Matrix: A Statistical Method and Software Tool for Linguistic Analysis through Corpus Comparison*. Ph.D. thesis, Lancaster University.

Reppen, R. (2001). Review of MonoConc Pro and WordSmith Tools. *Language Learning & Technology*, **5**(3): 32–6. Available online at http://llt.msu.edu/vol5num3/pdf/review4.pdf (accessed 1 March 2008).

Roberts, A., Al-Sulaiti, L., and Atwell, E. (2006). aConCorde: towards an open-source, extendable concordancer for Arabic. *Corpora*, **1**(1): 39–60.

Scott, M. (2007). *Introduction to WordSmith Tools*. Available online at <http://www.lexically.net/downloads/version5/HTML/index.html> (accessed 12 February 2008).

Smith, N. and Rayson, P. (2007). Recent change and variation in British English use of the progressive passive. *ICAME Journal*, **31**: 107–37.

Tottie, G. and Hoffmann, S. (2006). Tag questions in British and American English. *Journal of English Linguistics*, **34**(4): 283–311.

Tottie, G., Eeg-Olofsson, M., and Thavenius, C. (1984). Tagging Negative Sentences in LOB and LLC. In Aarts, J. and Meijs, W. (eds), *Corpus Linguistics: Recent Developments in the Use of Computer Corpora*. Amsterdam: Rodopi, pp. 173–84.

Wiechmann, D. and Fuhs, S. (2006). Concordancing software. *Corpus Linguistics and Linguistics Theory*, **2**(1): 107–27.

Notes

[1] Continuing the spatial analogy further, one may say that the concordance offers the user a bird's eye view of all the occurrences of the linguistic feature of interest.

[2] Two rare instances of corpora that have tag questions already annotated are ICE-GB and ICE-Ireland.

[3] He does, however, mention that automated annotation (e.g. part-of-speech tagging or syntactic parsing) is not 100% accurate and would therefore—if feasible, given the size of modern corpora—profit from manual post-editing.

[4] For example, *Dexter*'s website states that 'Dexter is also not a corpus concordancer' (http://www.dextercoder.org/about.html (accessed 4 September 2007).

[5] http://www.athel.com (accessed 1 March 2008).

[6] http://www.lexically.net/wordsmith/ (accessed 1 March 2008).

[7] http://www.bncweb.info (accessed 1 March 2008).

[8] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/ (accessed 1 March 2008).

[9] Earlier reviews appear in Hofland (1991), Lancashire (1991), Hughes and Lee (1994) and Kirk (1994).

[10] Wordsmith automatically creates a backup of the original corpus file(s). Furthermore, it keeps a log of these changes and offers an option to undo them if the annotation is no longer required.

[11] Mike Scott (personal communication) has indicated that this feature of *WordSmith* is still in its early stages of development and that future versions are likely to remove some of the existing restrictions.

[12] So far the only file format supported is tab-delimited text.

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011

13 We are, of course, aware that *Excel* is a spreadsheet program, designed primarily for numerical analysis. The program is, however, in widespread use in some areas of linguistics, such as field linguistics and language typology. The fact that Version 4 of *WordSmith* has introduced an export filter specifically for *Excel* is presumably in recognition of the latter program's convenience for manual annotation, coupled with its widespread availability.

14 This database of progressive passive instances is derived from the F-LOB corpus of 1990s British English. It has been used extensively in combination with databases extracted from several comparable corpora of British and American English (LOB, F-LOB, Brown, Frown) to examine variation and recent change in use of the progressive passive (Smith and Rayson, 2007).

15 This echoes a point that has already been made (e.g. by Leech, 2005) with regard to bottom-up annotation.

16 http://www.filemaker.com (accessed 1 March 2008).

17 Note, however, that quantitative analysis of the variable rule kind (e.g. using *VARBRUL*) can still be performed (cf. Kirk, 1994).

18 It must, however, be noted that the *MySQL* tables currently created by *BNCweb* are in a format that is far from human-readable. As a result, it would be quite a challenge to represent their contents to users of *FileMaker*—or any other SQL-based database application—in such a way that the data is amenable to manual work. The first of our three solutions clearly requires a careful design of internal data formats and it may therefore best be implemented in newly developed concordancers and corpus tools, for which this kind of interoperability is envisaged right from the outset.

19 In fact, such unique CPIs could also be provided as part of the pre-existing corpus annotation.

Downloaded from http://llc.oxfordjournals.org/ at UB Leipzig on December 9, 2011