End of Award Report

Creating a Parsed and Searchable Diachronic Corpus of Present-Day Spoken English (DCPSE)

R000239643

Survey of English Usage
University College London

1. Background

1.1 Diachronic and synchronic linguistics

Traditionally a distinction is made between diachronic and synchronic approaches to linguistics. The first considers language as it develops through time, whereas the latter takes a 'snapshot' look at languages viewed from the present. This old Saussurean dichotomy has recently been called into question, and some linguists have argued that the distinction is an artificial one. These linguists would argue that languages change all the time, even within the synchronic phases. As a result of these new attitudes to language development there is a new research impetus in linguistics which concerns itself with **recent change** (see Mair 1995, 1997; Mair and Hundt 1995, 1997, Denison 1998, Leech 2000, Smith and Leech 2001).

At the core of this project are two corpora of Modern British English, both founded at the Survey of English Usage (SEU) at University College London: the *London-Lund Corpus* (LLC), compiled in the 1960s, and the British Component of the *International Corpus of English* (ICE-GB, part-funded by ESRC R000232077), compiled in the 1990s.

The aim of this project was the construction of a diachronic corpus of spontaneous spoken English by carefully selecting directly comparable texts from the LLC and ICE-GB corpora. This corpus is a unique resource for linguists studying the spoken English of a period spanning 25-30 years. There is currently no comparable resource available, and the corpus will be the first of its kind enabling research into current change in spoken language.

For linguists who are interested in recent change corpora are especially valuable for data-gathering. Prior to this project they would have needed two separate corpora from two different periods to study a particular grammatical construction. Naturally, these corpora would have to be comparable as regards their internal composition (i.e. sampling criteria). An example of work done in this area is Aarts and Aarts (2002) which investigates the use of the English relative pronoun *whom*. In order to compare data from two periods of Present-Day English, the authors looked at material from the LLC and ICE-GB. They found that the overall use of *whom* as a Direct Object has become 90% less frequent over thirty years. Although ICE-GB is grammatically annotated and fully searchable, manual counts had to be carried out to find data in the older corpus. Thus, while the corpora were indispensable tools for this study, the research phase still required the careful pre-selection of comparable texts and manual searching of the LLC.

A parsed LLC sample is essential to permit the systematic exploration of grammatical variation over time, and will greatly facilitate research of this type, especially if it involves complex grammatical patterns.

In order to support research into current change Professor Christian Mair at the University of Freiburg has constructed two corpora of 1990s English: FLOB (Freiburg-Lancaster-Oslo-Bergen) and FROWN (Freiburg-Brown). These corpora are intended to match the LOB (Lancaster-Oslo-Bergen) and Brown corpora containing written English from the 1960s. These are excellent resources enabling linguists to research changes in written English over 30 years. Manual searches are still unavoidable, however, as these corpora have not been parsed.

This project has taken Mair's initiative further by constructing a corpus of British English comprising selections of spontaneous spoken English from the LLC and from ICE-GB. The new corpus will provide linguists interested in recent changes in English with a new, innovative and searchable database containing spoken English covering a period of 25-30 years. The focus on spoken English is justified because it is generally recognised that spoken language is primary and the first locus of changes in lexis and grammar.

1.2 The annotation of ICE-GB

In ICE-GB (Nelson, Wallis and Aarts 2002)[1] each sentence is *tagged* which means that a word class label is assigned to every word. In addition ICE-GB is *parsed*, i.e. it is given a full grammatical analysis. This is done in the form of tree diagrams. An example of a tree from ICE-GB is shown in Figure 1 below.

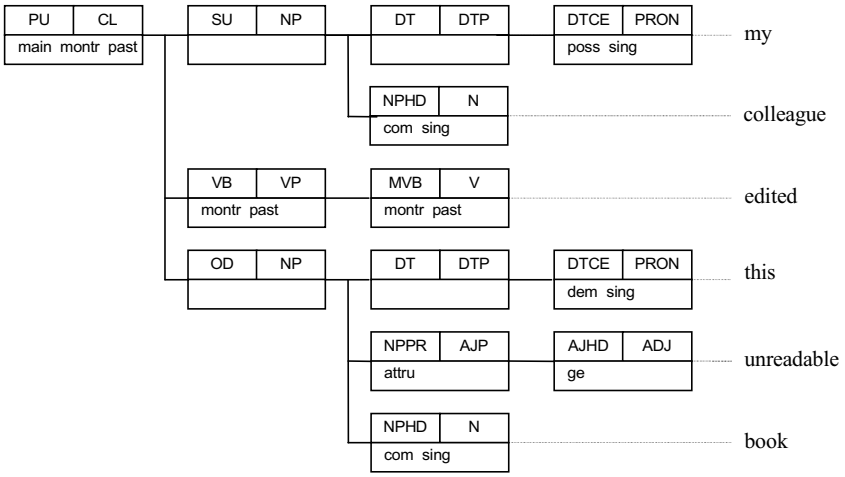

Figure 1: An ICE-GB tree for the sentence *My colleague edited this unreadable book.*[2]

[1] See www.ucl.ac.uk/english-usage/ice-gb, where information about the corpus and software downloads are available for review.

[2] Gloss (features are in italics): PU=parse unit, CL=clause, *main=main, montr=monotransitive, past=past tense,* SU=subject, NP=noun phrase, DT=determiner, DTP=determiner phrase, DTCE=central determiner, PRON=pronoun, *poss=possessive*, *sing=singular*, NPHD=NP head, N=noun, *com=common*, VB=verbal, VP=verb phrase, MVB=main verb, V=verb, OD=direct object, *dem=demonstrative*, NPPR=NP premodifier, AJP=adjective phrase, AJHD=adjective head, ADJ=adjective, *attru=attributive, ge=general.*

In this representation each node (box) is assigned a *function* label (top left), a *category* label (top right), as well as *features* (lower part of the box) which may percolate upwards (e.g. the features 'montr' and 'past' have percolated up from the verb *edited* to the highest level, PU). ICE-GB contains 600,000 words of transcribed speech and 400,000 words of written English.

The LLC component of DCPSE has been annotated in the same way.

1.3 The exploration of ICE-GB

Once corpora have been contructed tools must be developed to exploit them. At the SEU we developed ICECUP (the International Corpus of English Corpus Utility Programme), which allows user-friendly, fast and effective grammatical searches. Central to the exploitation of any dataset, including a parsed corpus, is the *query*. A corpus query is best understood as an abstract pattern which represents a particular type of linguistic structure. In a corpus consisting only of plain text, a query might simply be a string of characters or words, or a wildcard. If the corpus contains part of speech (POS) tags, then one can express a query exploiting these tags, e.g. one can search for a pronoun followed by a verb. Much more interestingly, queries performed on a parsed corpus are able to exploit tree configurations and extract particular kinds of structures, e.g. a verb followed by a direct object and an adjunct, or a noun phrase containing an adjective phrase.

In a previous project (ESRC R000222598) we concentrated on the retrieval of grammatical structures in a parsed corpus, developing a system of structural queries which we call *Fuzzy Tree Fragments* (FTFs, Wallis and Nelson, 2000a).[3] An FTF is a kind of syntactic search wildcard which can be constructed by users, expressed as a partial tree (Figure 2). Just like full trees, FTFs contain nodes and lexical items, both of which can be more or less completely specified (hence *fuzzy*), according to the user's needs. Links between nodes may be specified exactly or inexactly. For example, in Figure 2 the user may specify that the subject NP must immediately precede the verb phrase, or allow for there to be intervening material, e.g. a sentence adverb. Similarly, a user may insist that a particular node is the last one in a sequence of nodes. The result is an intuitive search system for grammatical queries.

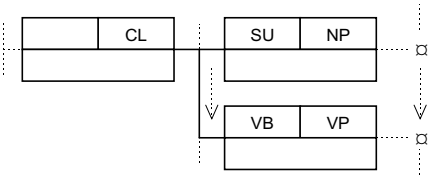

Figure 2: a simple FTF matching a clause with a subject NP and a verb phrase

FTFs form the centrepiece of a corpus exploration, annotation and maintenance platform. ICECUP supports the construction of FTFs, the efficient retrieval of matching structures and the browsing of results. It also allows linguists to search for logical combinations of FTFs. In addition, a 'wizard' tool in ICECUP lets users extract a more general FTF from an existing tree in the corpus, allowing them to find many more cases of the same structure. As a result, users can learn the underlying grammar through a process of corpus exploration, a benefit for experienced researchers and novices alike. Fuzzy Tree Fragments have been used for theoretical

[3] See www.ucl.ac.uk/english-usage/ftfs for a detailed description of FTFs, links and edges. ICECUP is available, with sample corpus and documentation, from www.ucl.ac.uk/english-usage/ice-gb/sampler.

syntactic research, teaching, dictionary construction, knowledge discovery (Wallis and Nelson, 2001), and corpus experimentation (Nelson, Wallis and Aarts, 2002; Wallis, 2003b).

2. Objectives

As mentioned above the principal objective of this project has been the construction of an 800,000-word text corpus of spoken English to provide a unique platform for research in recent change in grammar and lexis.

This resource, which we call a *Diachronic Corpus of Present-Day Spoken English* (DCPSE), will allow researchers to investigate changes in the grammar and usage of Present-Day English over a period of 30 years. DCPSE differs from FLOB and FROWN (see section 1.1) in a number of important ways. Firstly, the corpus is unique in containing exclusively **spontaneous spoken English**.[4] In due course we hope to provide a playback facility enabling linguists to listen to the original recordings. Secondly, the corpus is **parsed** which will permit research into synchronic and diachronic grammatical variation. Thirdly, the corpus is **fully searchable** using the ICECUP software that we developed for ICE-GB. This software has been modified to operate on the new data.

DCPSE will be a major new resource complementing the Freiburg corpora, allowing access for the first time to recordings that could hitherto only be listened to at the SEU premises.

Aims and objectives, as listed in the Research Proposal, and how they have been achieved.

1. **Select a total of 800,000 words of spoken English from comparable categories in the LLC and in ICE-GB (400,000 words from each corpus). The design of these corpora is similar, and it will thus be possible to select identical categories of spontaneous spoken English. In each case we will select a matching pair of texts, and cross-check the structural markup and tagging in the LLC.**

   Completed. See Appendix 1 for a list of the texts selected from the LLC and ICE-GB.

2. **Integrate the LLC and ICE-GB material.**

   Completed. LLC texts were labelled 'DL-*xnn*' and ICE-GB texts labelled 'DI-*xnn*' where *x* stands for a category label and *nn* for a two-digit number. These replace previous index labels. A number of processes were carried out to normalise the annotation (see below). Some very long monologue utterances (>1,000 words) were initially broken into feasible segments. These could then be read into ICECUP and indexed in an integrated fashion. We also decided to allow corpus correctors to manually alter the segmentation during the parse correction process

[4] According to our estimates, there are only under a million words of publicly available, parsed and checked, orthographically transcribed spoken material, of which 600,000 are in ICE-GB, 144,000 in the Penn Treebank and 130,000 in Geoffrey Sampson's CHRISTINE corpus. CHRISTINE includes about one tenth of the LLC with the prosodic annotation removed.

(see 5(c) below), to follow the same general semantic criteria as were used with ICE-GB.

3. **Modify ICECUP to handle the combined data. ICECUP was originally developed to operate on ICE-GB. It will need to be modified to handle the proposed 'two-in-one' corpus.**

   Completed. There are a number of issues here, including the fact that LLC texts are a standard 5,000 word length and may consist of as many as a thousand individual 'sentences'. A *corpus definition file* now describes the summary structure of the corpus. Sociolinguistic information was extracted from a number of database sources and integrated both to construct a unified corpus map and further elaborate the speakers in the corpus.

4. **Parse the LLC material. Given the 'messy' nature of spoken English this will be a major task.**

   This was carried out automatically to phrasal level using a **partial parser** developed for the project (see Methods, below). The parsing process consists of a number of subtasks, which are described in Section 3. (Note that the corpus had been tagged with POS labels, but not checked, prior to the start of the project.)

5. **Ensure 'analytic consistency' across the two subcorpora, i.e. make sure that analytical decisions for the LLC material are consistent with those made for ICE-GB. This will necessitate the writing of additional software.**

   We have ensured that analytic consistency was applied to every type of annotation from segmentation, the consistent use of extra-corpus annotation and the employment of identical POS-tagging criteria to the detail of the grammatical analysis. Additional tools were written to (a) translate tags and **re-tag the corpus** and (b) **parse the corpus** using ICE-GB as a knowledge base. The ICECUP software was extended to include additional correction and proofing tools, including (c) **new editing commands** for segmentation, self-correction and macro-insertion, (d) a **lexicon** and (e) a **grammaticon** (See Section 3, Methods).

6. **Manually check the parse results.**

   Research assistants were fully trained to check the parse results using the ICECUP software (see Section 3). Given the complexities of spoken language (see below) this has proved to be a major task, which we have now virtually completed. A small amount of checking remains to be done, which we envisage to complete early in the new year. We organised a workshop involving all the members of the research team, as well as Dr Gerry Nelson, to discuss the analysis of particularly difficult parsing units.

7. **Prepare and enhance the digitised LLC sound recordings, so that these can be used by researchers. The LLC recordings have not been disseminated until now. We will 'bleep out' the names and subdivide them into 'sentences' or groups of overlapping 'sentences'.**

   This task was proposed as a parallel task to the checking of the parsing. As a result of the checking phase taking longer than anticipated, the preparation of the sound

material has been delayed. *Assessors, please note that we did not request funding from the ESRC for this task.* (See Technical Appendix of the original proposal.) However, we hope to carry out this task by seeking additional funding. We also intend to upgrade DCSPE to include the original prosodic markup in the LLC.

8. **Write documentation and disseminate the new fully searchable diachronic corpus with the ICECUP software.**

   Completed. We have updated and expanded the *Getting* Started manual originally designed for the ICE-GB corpus. We will disseminate a free sample corpus with ICECUP over the web. The full corpus can be purchased from the Survey of English Usage.

3. Methods

As noted above, the corpus had been tagged, but not checked, before the project was started.

One of the principal tasks in this project has been the correction and proofing of the automatic parsing procedure. As is well known, the grammatical analysis of spoken utterances is not a straightforward process. Whereas the written word is subject to varying degrees of editorial correction by authors, spontaneous speech is not subject to such explicit correction. Instead, speakers tend to repeat themselves, 'self-correct' their utterances. They produce ungrammatical utterances, make false starts and slips of tongue, and mispronounce words, Many conversations are subject to overlap (people speaking at the same time). The difficulty of completing the parse analysis of spoken material is a challenge that has faced a number of research projects, such as the *HCRC Maptask Corpus* (ESRC R000236800). The corollary of this is that in the scholarly community a relatively small amount of such material has been analysed.

A related issue concerns segmentation. Parsing a corpus requires that texts are split into segments ('text units') which are then parsed. However, the LLC corpus is subdivided into *speaker turns*. As a result, extempore monologues can run for a thousand words or more (interrupted only by audience laughter or coughing), while in conversational texts, speaker interruptions and overlap often lead to highly fragmented sentences. Every interruption entails a new fragment. Neither of these extremes is compatible with a coherent syntactic interpretation of utterances.

We have used a number of methods to overcome the difficulties associated with the parsing of spoken material. The parsing of transcribed speech consists of three subtasks:

1. Identify and abstract each idealised sentence, starting by identifying each segment and determining areas of self-correction. Note that we do not permit the insertion of ellipted material, null elements, etc.

2. Subject these sentences to a parse analysis consistent with the analysis in ICE-GB.

3. Complete the parsing of the tree to include an analysis of self-corrected material.

ICE-GB handles segmentation by treating each speaker track as a potential sequence of sentences and then introducing additional markup to indicate how speakers overlap.

Consider the utterances below (taken from ICE-GB, S1A-006, #144-146). Speaker A overlaps B twice in the same sentence. Different principles of segmentation are used in LLC and ICE-GB.

LLC	**ICE**
<Speaker B> *I mean she fell in love with him*	<Speaker B> *I mean she fell in love with him* [*the*]$_1$ *fifteen-year-old him* [*back*]$_2$ *in time*
<Speaker A> [*Yes*	<Speaker A> [*Yes*]$_1$
<Speaker B> *the*] *fifteen-year-old him*	
<Speaker A> [*Yes*	<Speaker A> [*Yes*]$_2$
<Speaker B> *back*] *in time*	

In the LLC, a sentence with overlaps is broken at each overlap (left), while in ICE-GB the sentence is retained as an entire segment and the overlapping is numbered accordingly. Prior to parsing it is necessary to join such coherent fragments together as much as possible and reorder the overlapping elements.

Overlapping can be in the background or force the termination of the first speaker's utterance (an interruption) and effect a change of speaker turn. In some cases several speakers speak simultaneously. Annotation therefore cannot be performed without addressing this issue manually. We have provided new tools in the ICECUP platform which allow correctors to both break run-on utterances and join interrupted fragments. See also below.

Automatic parsing can only be carried out effectively once sentences have been adequately segmented. The clause-level analysis of such sentences is a difficult task which is further complicated by issues of supplementation, as well as self-correction, both of which are common in speech.

In the project proposal we suggested using the TOSCA/ICE parser developed by the TOSCA research group in Nijmegen, Holland. However this had a number of drawbacks: it not been updated for some time, ran on UNIX workstations, required a manual pre-parsing stage and the annotation scheme it used was slightly inconsistent with ICE-GB. This meant that deciding to use the Nijmegen parser would involve significant manual pre-processing before any trees could be examined and corrected. This 'top-down' parser also has serious difficulties with less 'regular' and long sentences.

We therefore decided to develop a stochastic 'bottom-up' parser, which was trained on ICE-GB and applied to the LLC material. This can handle very large streams of text, and has high coverage. It works by attempting to identify and match phrases in the stream competitively. The advantage of working from the bottom up is that it is far less dependent on the overall completeness of a sentence. The disadvantage is that local variation at the tag level tends to overrule general knowledge.

In order for us to assess the performance of the partial parser, consider the following utterance:

I did do a certain amount I've done I did a certain amount of reading during the last few months <,> and I have been <,> and I went away to did it <,> to do it I went away from home

The partial parser can be applied to this ‘sentence’, producing the tree in Figure 3 on page 9. However, this fragment of spontaneous speech includes a number of false starts. Where the speaker ‘repairs’ their own sentence, it is possible to mark this self-correction. In addition, the editor can choose to break the text unit by inserting ‘↳‘ marks. Corrected material is marked with a strike-through and flagged as ‘ignored’. The replacement material is marked with an insert (‘⇨’) and placed in a box.

The utterance above has been split and self-corrected as follows:

I did do a certain amount

↳ ~~I've done~~ [I did] a certain amount of reading during the last few months <,,> ~~and I have been <,>~~ ⇒ [and I went away] ~~to did it <,>~~ [to do it]

↳ I went away from home

The parsing of each of these is then checked manually. The central fragment, parsed and corrected, is shown in Figure 4.

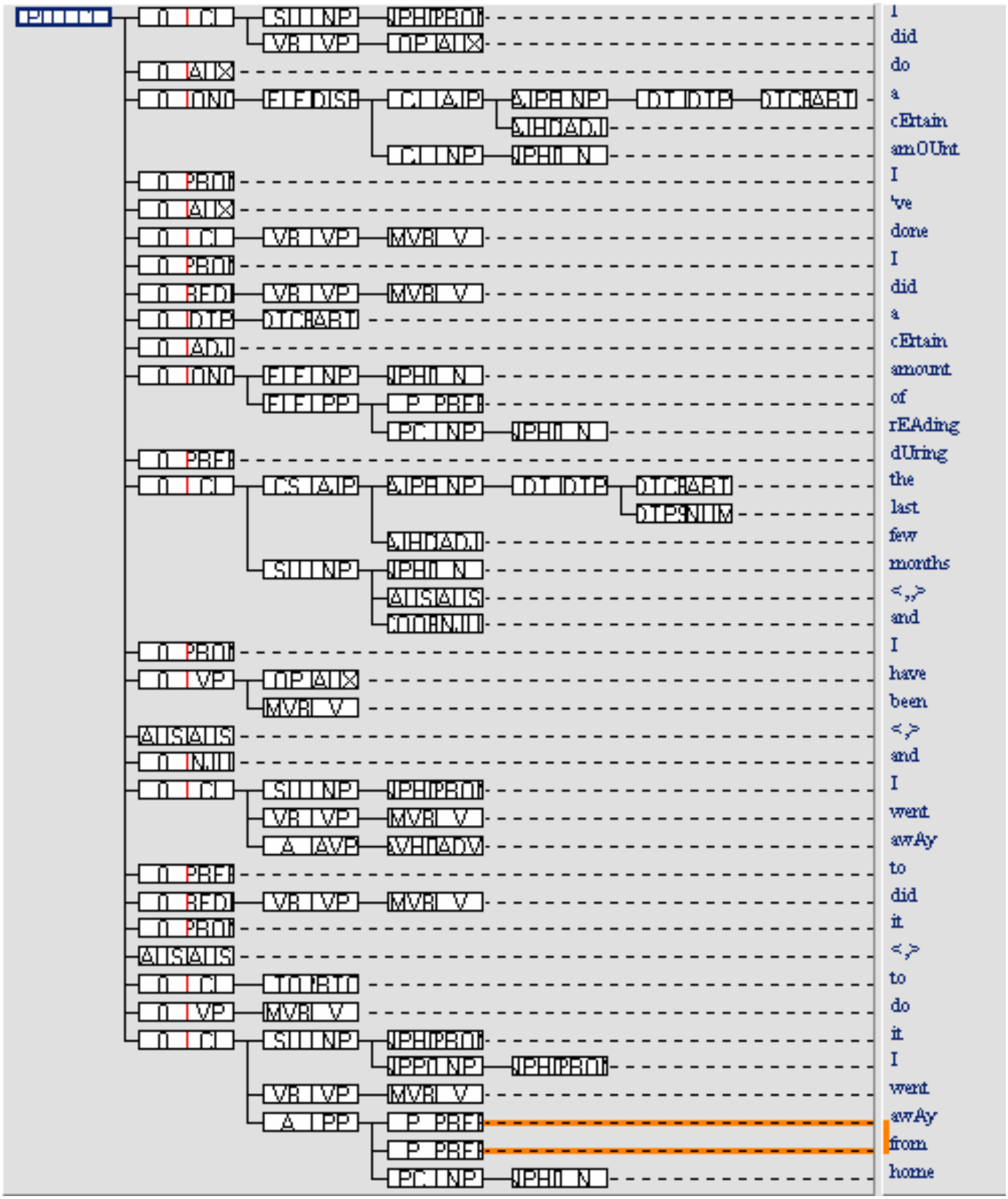

Figure 3: Partially parsed version of the initial utterance

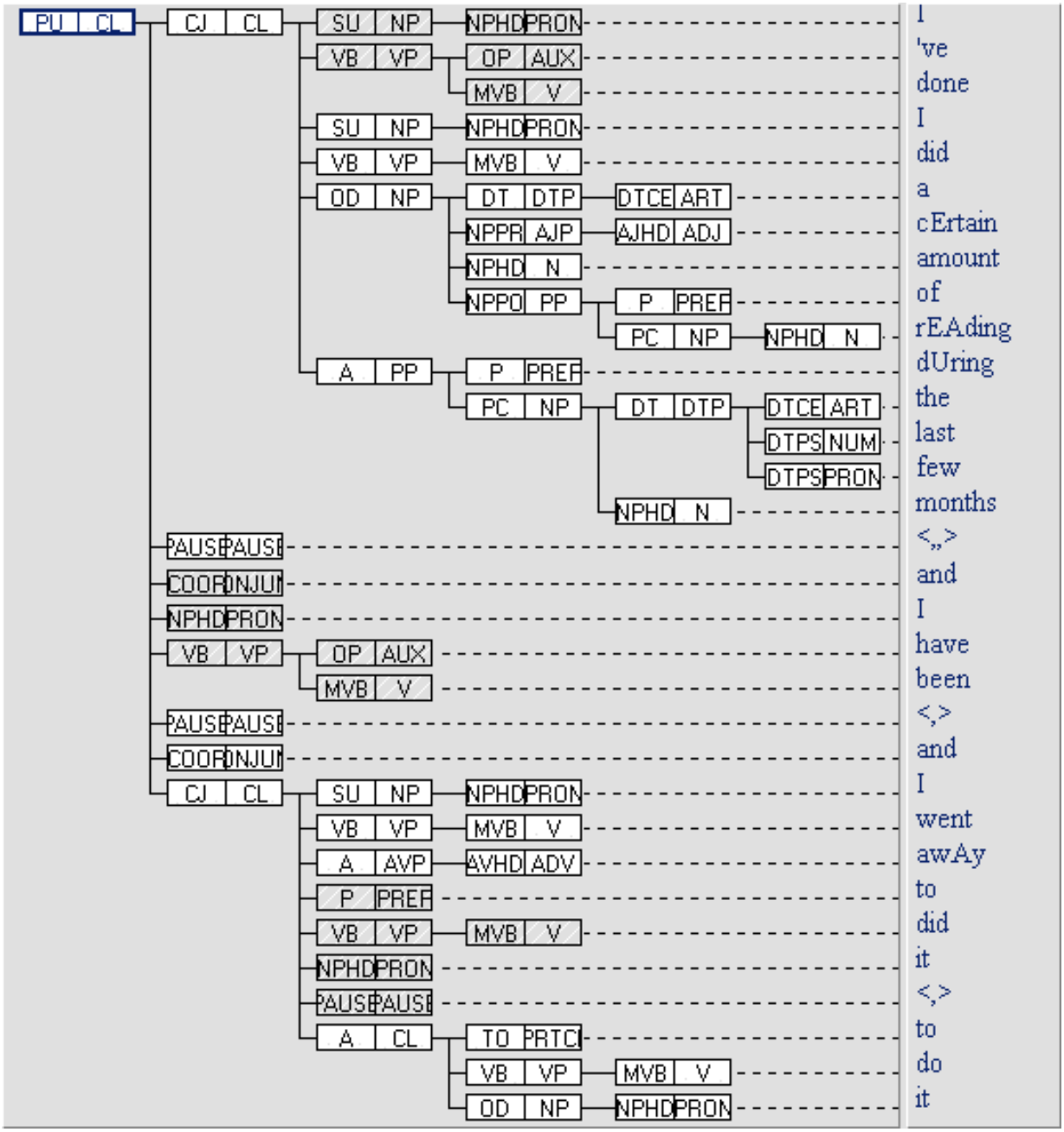

Figure 4: Parsed version of the central fragment, corrected. Greyed nodes have been self-corrected.

Trees were corrected using the ICECUP 3.1 software. This software consists of an underlying corpus maintenance system onto which a user interface, including a number of powerful query and browsing tools, is constructed.

In this project we extended this software to support run-time indexing, parallel searching and a facility for re-integrating corrections carried out off-site. These extensions to the corpus maintenance system were primarily designed to aid our annotators, although parallel searching is a useful facility for researchers. The same software is used to disseminate the corpus, with editing facilities hidden and locked.

ICECUP contains a robust tree editor derived from ICETree 2 (Wallis and Nelson 1997) which was used to complete the parsing of ICE-GB. This editor has been

deployed on approximately 500,000 trees. The editor was extended to include the facility to copy and insert whole branches.

As we stated in the proposal, traditional approaches to correcting and augmenting parse analysis consist of a user carrying out a longitudinal pass through a text sequence, sentence-by-sentence, correcting each tree in turn. This methodology is particularly difficult and prone to inconsistency, because the user is typically presented with a different set of grammatical problems with every tree.

As an alternative, we have been developing a new approach termed *cross-sectional correction* (also referred to as *transverse correction*, Wallis 2003a). Here human effort is directed by selecting specific types of construction, and correcting them systematically across the entire corpus. This has the benefit that it forces the corrector to consider issues of consistency with existing parsed material - a prerequisite in this project. It is also easier.

When we completed ICE-GB, we used this transverse method at a final proofing stage, working on completed (but potentially inconsistent) trees. As shown above, in parsing the LLC we deployed a partial parser, trained on ICE-GB, to work from the sentence up, and then cross-checked and extended this analysis upwards. This is a far more ambitious use of transverse checking.

We used simple *Fuzzy Tree Fragments* (FTFs, see Section 1.3 above) based on lexical item or POS-tag sequences, to cross-check the parsing. The idea is to use what one knows to be reliable as a baseline to check potentially unreliable parsing judgments. Since the approach is cross-sectional, the linguist is presented with a series of similar decision tasks, one after another, resulting in a high degree of consistency. As we note above, we provided a simple structure insertion tool to allow linguists to 'paste in' one of several possible structures.

While cross-sectional checking is useful, it is more difficult to identify the context of a sentence than in a longitudinal exercise – the software allows for this, but it can be distracting. Secondly, correctors may find that focusing on a single problem at a time is repetitive. Nevertheless, we believe that cross-checking has resulted in more reliable parsing, and rough estimates show that the process is around twice as fast as conventional longitudinal checking. It is easier to train correctors to carry it out, and it is useful in the training of correctors. However it is always necessary to apply longitudinal 'proofing' to the corpus to gain complete coverage.

Because it can be difficult to ascertain how much of the corpus has been checked we introduced a mechanism for marking sentences as 'finished'.

We will comment on one final type of cross-checking. We have developed two new integrated 'overview' tools in ICECUP 3.1: a lexicon and a grammaticon.

The lexicon is a simple hierarchy of lexical and lexical/part-of-speech elements which can be organized in a variety of ways by the user. Each element in the lexicon, i.e. each *word*, *word+tag combination*, or *intermediate group*, is a query, and the frequency of each item in the corpus is shown. Clicking on the item lists all matching cases in their sentence context, within a second window. From a cross-checking point of view, the lexicon allows the user to spot implausible lexical items or POS tags, and

to retrieve all cases of that item. Where more than one POS tag is given for a lexical item, one can also identify alternatives for an incorrect tag.

The grammaticon goes further. This tool summarises all the distinct *node patterns* in the corpus, from clausal and phrasal nodes to part of speech nodes. Again, the structure can be organized by the user. Implausible patterns can be detected and the example sentences containing them highlighted. Finally, one can calculate the intersection between the grammaticon or lexicon frequencies and any other query. In particular, we can examine those elements which differ markedly in their frequency between the LLC and ICE-GB subcorpora. This is an extremely useful feature for scholars interested in language change.

4. Results

The collection and annotation of the material for DCPSE is complete. The LLC and ICE-GB materials have been integrated and the exploration software (ICECUP) has been modified to operate with the new resource. As noted above, a small amount of correction still needs to be done. As soon as we have done this we will publish the corpus, together with ICECUP and the documentation we have written. A sample corpus will be published on the web.

5. Activities

The DCPSE project was presented in a number of different places, listed below:

AARTS, B. (2004) Recent developments in corpus linguistics, including a demonstration of DCPSE. Presentations at: the University of Jaén (Spain), The Academy of Korean Studies (Seoul, South Korea), Pusan National University (Pusan, South Korea), King's College London, The Institute of Education, London.

AARTS, B. and S. WALLIS (2005) Recent developments in the syntactic annotation of corpora: a demonstration of ICE-GB and DCPSE. Paper to be presented at the Ninth International Symposium on Social Communication and the Pre-Symposium Seminar on Corpus Linguistics. Santiago de Cuba, Cuba.

WALLIS, S. (2004) Using ICECUP with parsed corpora. Presentation at Roehampton, University of Surrey.

WALLIS, S. and AARTS, B. (2003) Tracking the development of spoken English across the decades: the Diachronic Corpus of Present-day Spoken English. Paper presented at the annual ICAME conference, Guernsey.

6. Outputs

Software

DCSPE is explored using the new release of ICECUP 3.1, which includes a series of refinements and extensions to version 3. ICECUP 3.1 is currently being beta tested by a number of external reviewers who are active in experimental corpus linguistics.

The beta version is available for free download at www.ucl.ac.uk/english-usage/ice-gb/beta, complete with *sample corpora* from ICE-GB and DCPSE.

We have written the second edition of our *Getting Started* booklet for use with the corpus and software (see below).

Datasets

We have agreed with Susan Cadogan, Senior Acquisition Officer at the ESRC Data Archive, the following procedure to assess and archive the data: the SEU will make a provisional deposit ('the beta dataset'), and then provide the Data Archive with a final deposit at a later date. (The reason for this is that we wish to test and revise the resource in an ongoing manner for a short period, i.e. up to six months.)

The ESRC Data Archive will assess the suitability of the onward submission of the corpus to the Oxford Text Archive.

The position we agreed with the OTA, specified in the original ESRC proposal, was as follows:

- The SEU will deposit the primary data, create records, etc. describing the resource, but enquiries and dissemination would be handled by the SEU.
- A number of consequential data outputs, created from the primary data, would be deposited. These would be freely disseminated by the receiving agency. Support issues arising from these data will be handled by the SEU.

We also agreed the following with the Data Archive:

The SEU will send sample files of DCPSE material with the data submission form.

At a later stage the SEU will supply a CD containing the current beta dataset and will agree a record for the ESRC database, cross-referenced it to the SEU's web pages dealing with the published resource.

The final version of the corpus will be published from the SEU on a professionally produced CD, and disseminated, at a licenced cost, by the Survey of English Usage. As outlined in the Research Proposal, we were granted rights by copyright holders to an audited distribution of material for non-commercial research and educational purposes. A 'preservation copy' will be sent to the Data Archive, as well as the consequential datasets abstracted from the primary resource, with descriptions of these to aid dissemination.

DCPSE materials, including a sample corpus, lexicon and grammaticon data will be published at the project website: www.ucl.ac.uk/english-usage/diachronic.

The sample corpus is a 20,000-word balanced subcorpus of DCPSE. 10,000 words have been taken from the 1990s (ICE-GB) part of the corpus, and 10,000 words from the 1960s-70s (LLC) part.

The lexicon data is published from DCPSE and structured into separate tables by ICE wordclass categories. It consists of tables of nouns, verbs, adverbs, etc. and their frequencies in the corpus.

The grammaticon data is also extracted and structured by categories and functions. It consists of tables of grammatical patterns and their corpus frequencies.

Documentation for this data will be available from the website shown above. The sample corpus includes the ICECUP 3.1 software which permits browsing, searching and exploration of the 20,000 word corpus, and contains full help files.

Publications

AARTS, F. and B. AARTS (2002) Relative *Whom*: a 'Mischief Maker'. In: A Fischer G. Tottie and P. Schneider (eds.). *Text Types and Corpora.* Tübingen: Gunter Narr Verlag. 123-130.

WALLIS, S. (2003a) Completing parsed corpora: from correction to evolution. In: Anne Abeillé (ed.), *Treebanks: building and using parsed corpora*. Boston: Kluwer. 61-71.

WALLIS, S. (2003b) Scientific experiments in parsed corpora: an overview. In: Sylviane Granger and Stephanie Petch-Tyson (eds.) *Extending the scope of corpus-based research: new applications, new challenges*. Language and Computers 48. Amsterdam/New York: Rodopi. 27-38.

NELSON, G., S.WALLIS and B. AARTS (2004) *Getting Started with ICECUP, Versions 3.0 and 3.1*. For use with: *The British Component of The International Corpus of English* (ICE-GB) and *A Diachronic Corpus of Present-Day Spoken English* (DCPSE). Second Edition. London: Survey of English Usage.

WALLIS, S. (forthcoming, 2005). Searching treebanks and graphs. In: Anke Lüdeling, Merja Kytö, and Tony McEnery (eds) *Handbook on Corpus Linguistics*. Handbücher zur Sprach- und Kommunikationswissenschaft. Berlin: Mouton de Gruyter.

7. Impacts

As we have stated in the research proposal, there is a growing community of linguists who are interested in *current change* in the English language. We will widely publicise the release of the corpus and its search software and expect the take-up of this resource to be considerable.

8. Future Research Priorities

We envisage the possibility in the future of making the prosodic material of the LLC searchable. Another research aim might be prosodically annotating the ICE-GB part of DCPSE (and also making this material searchable). A further, long-term, project might be the collection and annotation of further spoken material over the decades, thus creating a monitor corpus that can be used to track changes in contemporary English.

9. References

AARTS, B., NELSON, G., and WALLIS, S.A. (1998) Using Fuzzy Tree Fragments to Explore English Grammar. *English Today* 14, 52-56.

AARTS, F. and B. AARTS (2002) Relative *Whom*: a 'Mischief Maker'. In: Andreas Fischer, Gunnel Tottie and Peter Schneider (eds.) *Text types and corpora*. Tübingen: Gunter Narr Verlag. 123-130.

DENISON, D. (1998) Syntax. In: S. Romaine (ed.). *The Cambridge History of the English Language. IV: 1776-1997.* Cambridge. 92-329.

KENNEDY, G. (1998) *An Introduction to Corpus Linguistics*. London.

LEECH, G. (2000) Diachronic linguistics across a generation gap: from the 1960s to the 1990s. Paper read at the symposium *Grammar and Lexis*. University College London Institute of English Studies.

LJUNG, M. (1997)(ed.) *Corpus-Based Studies in English.* Amsterdam.

MAIR, C. (1995) Changing Patterns of Complementation and Concomitant Grammaticalisation of the Verb *help* in Present-Day English. In: B. Aarts, and C.F Meyer (eds.). *The Verb in Contemporary English*, Cambridge. 258-272.

MAIR, C. (1997) Parallel Corpora: a Real-Time Approach to the Study of Language Change in Progress. In: M. Ljung, M. (ed.). 195-209.

MAIR, C. and HUNDT, M. (1995) Why is the Progressive Becoming More Frequent in English? A Corpus-Based Investigation of Language Change in Progress. *Zeitschrift für Anglistik und Amerikanistik* 43.2. 111-122.

MAIR, C. and M. HUNDT (1997) The Corpus-Based Approach to Language Change in Progress. In: U. Böker and H. Sauer, H. (eds.). *Anglistentag 1996*. Dresden.71-82.

NELSON, G., WALLIS, S.A., and AARTS, B. (2002). *Exploring Natural Language*. Amsterdam.

QUIRK, R., GREENBAUM, S., LEECH G., and SVARTVIK, J. 1972. *A Grammar of Contemporary English*. London.

———— 1985. *A Comprehensive Grammar of the English Language*. London.

SMITH, N. AND G. LEECH (2001) Grammatical change in recent written English, based on the FLOB and LOB corpora. Paper read at the ICAME conference. Louvain-la-Neuve, Belgium.

SVARTVIK, J. 1990 (ed.). *The London-Lund Corpus of Spoken English: Description and Research*. Lund Studies in English 82. Lund.

SVARTVIK, J., and QUIRK, R. 1980. *A Corpus of English Conversation*. Lund.

WALLIS, S. (1999) Completing parsed corpora: from correction to evolution. In: A. Abeillé (ed.). *Journées ATALA sur les Corpus Annotés pour la Syntaxe – Treebanks Workshop*. 7-12.

WALLIS, S. and G. NELSON (1997) Syntactic parsing as a knowledge acquisition problem. *Proceedings of 10th European Knowledge Acquisition Workshop*, Catalonia, Spain, Springer Verlag. 285-300.

WALLIS, S. and G. NELSON (2000) Exploiting fuzzy tree fragments in the investigation of parsed corpora. *Literary and Linguistic Computing* **15**, 3: 339-361.

WALLIS, S. and G. NELSON (2001) 'Knowledge discovery in grammatically analysed corpora'. *Data Mining and Knowledge Discovery* **5**. 305-335.

WALLIS, S. (2003a) Completing parsed corpora: from correction to evolution. In: Anne Abeillé (ed.), *Treebanks: building and using parsed corpora*. Boston: Kluwer. 61-71.

WALLIS, S. (2003b) Scientific experiments in parsed corpora: an overview. In: Sylviane Granger and Stephanie Petch-Tyson (eds.) *Extending the scope of corpus-based research: new applications, new challenges*. Language and Computers 48. Amsterdam/New York: Rodopi. 27-38.

Appendix 1: Texts selected from LLC and ICE-GB

A. Face-to-face dialogue (more formal) 2 x 40,000 words

LLC	**8 texts**		**ICE-GB**	**20 texts**	
text code	LLC text code	year of recording	text code	ICE-GB text code	year of recording
DL-A01	S.3.1	1961	DI-A01-04	S1A-001-04	1991
DL-A02	S.3.2	1973-1975	DI-A05	S1A-024	1991
DL-A03	S.3.3	1971	DI-A06-08	S1A-033-35	1992
DL-A04	S.3.4	1971	DI-A09-10	S1A-050-51	1991
DL-A05	S.3.5	1961	DI-A11-12	S1A-059-60	1991,1990
DL-A06	S.3.6	1974	DI-A13	S1A-062	1991
DL-A07	S.6.2	1961	DI-A14	S1A-066	1992
DL-A08	S.6.8	1977	DI-A15	S1A-072	1991
			DI-A16-17	S1A-075-76	1991
			DI-A18-20	S1A-087-89	1991-1992

B. Face-to-face dialogue (more informal) 2 x 180,000 words

LLC	**36 texts**		**ICE-GB**	**90 texts**	
text code	LLC text code	year of recording	text code	ICE-GB text code	year of recording
DL-B01-13	S.1.1-1.13	1963-1976	DI-B01-19	S1A-005-23	1991
DL-B14-25	S.2.2-2.13	1973-1975	DI-B20-27	S1A-025-32	1991
DL-B26-32	S.4.1-4.7	1969-1976	DI-B28-41	S1A-036-49	1990-1991
DL-B33-36	S.5.8-5.11	1971-1976	DI-B42-48	S1A-052-58	1991-1992
			DI-B49	S1A-061	1992
			DI-B50-52	S1A-063-65	1991
			DI-B53-57	S1A-067-71	1991-1992
			DI-B58-59	S1A-073-74	1991
			DI-B60-69	S1A-077-86	1992
			DI-B70	S1A-090	1992
			DI-B71-90	S1B-001-20	1991-1992

C. Telephone conversations (mostly informal) 2 x 20,000 words

LLC	**4 texts**		**ICE-GB**	**10 texts**	
text code	LLC text code	year of recording	text code	ICE-GB text code	year of recording
DL-C01-02	S.7.1-7.2	1961-1975	DI-C01-10	S1A-091-100	
DL-C03-04	S.8.1-8.2	1975-1976			

D. Broadcast discussions (disparates/equals) 2 x 40,000 words

LLC	**8 texts**		**ICE-GB**	**20 texts**	
text code	LLC text code	year of recording	text code	ICE-GB text code	year of recording
DL-D01-07	S.5.1-5.7	1959-1970	DI-D01-20	S1B21-40	1990-1991
DL-D08	S.6.5	1975			

E. Broadcast interviews (disparates/equals) 2 x 20,000 words

LLC	**4 texts**		**ICE-GB**	**10 texts**	
text code	LLC text code	year of recording	text code	ICE-GB text code	year of recording
DL-E01	S.6.1	1966	DI-E01-20	S1B41-50	1990-1991
DL-E02	S.6.3	1974			
DL-E03	S.6.6*	1974			
DL-E04	S.6.7	1971			

*monologue in interview

F. Spontaneous commentary — 2 x ~45,000 words

LLC	**9 texts**		**ICE-GB**	**23 texts**	
text code	LLC text code	year of recording	text code	ICE-GB text code	year of recording
Sports:					
DL-F01-4	S.10.1-10.4	1960-1971	DI-F01-16	S2A-001-07	1990-1991
				S2A-009-10	1991
				S2A-012-18	1991
State funeral:			*Trooping the Colour:*		
DL-F05	S.10.5	1965	DI-F17	S2A-011	1991
Royal wedding:			*The Gulf Ceremony:*		
DL-F06	S.10.6	1973	DI-F18	S2A-019	1991
Mixed					
DL-F07-08	S.10.7-10.8	1960-1976			
Scientific demonstrations:					
DL-F09	S.10.9	1976	DI-F19-23	S2A-051-55	1990-1991

G. Parliamentary Language — 2 x 10,000 words

LLC	**1 text**		**ICE-GB**	**2 texts**	
text code	LLC text code	year of recording	text code	ICE-GB text code	year of recording
DL-G01	S.11.4	1975	DI-G01	S1B-051	1990
DL-G02	S.11.5	1975	DI-G02	S1B-053	1990
			DI-G03	S1B-055	1990
			DI-G04	S1B-058	1990
			DI-G05	S1B-059	1990

H. Legal cross-examination — 2 x ~5,000 words

LLC	**1 text**		**ICE-GB**	**2 texts**	
text code	LLC text code	year of recording	text code	ICE-GB text code	year of recording
DL-H01	S.11.1	1967	DI-H01-02	S1B-061-62	1990

I. Legal cross-examination — 2 x 10,000 words

LLC	**2 texts**		**ICE-GB**	**5 texts**	
text code	LLC text code	year of recording	text code	ICE-GB text code	year of recording
DL-I01-02	S.11.2-11.3	1974, 1961	DI-I01	S2A-021	1991
			DI-I02	S2A-024	1991
			DI-I03	S2A-025	1991
			DI-I04	S2A-032	1991
			DI-I05	S2A-037	1991

J Prepared Speech (mostly monologue) — 2 x 30,000 words

LLC	**6 texts**		**ICE-GB**	**15 texts**	
text code	LLC text code	year of recording	text code	ICE-GB text code	year of recording
DL-J01-06	S.12.1-12.6	1965-1972	DI-J01	S2A-020	1991
			DI-J02-06	S2B-001-05	1990-1991
			DI-J07-11	S2B-021-25	1991
			DI-J12-15	S2B-041-44	1990-1991