Using Fuzzy Tree Fragments to explore English grammar

BAS AARTS, GERALD NELSON, and SEAN WALLIS

Survey of English Usage, Department of English Language and Literature, University College, London

Readers of *ET* may recall two papers, the first by the late Sidney Greenbaum ('ICE: the International Corpus of English,' *ET*7, 1991, 3–7), the second and by Akiva Quinn & Nick Porter ('Investigating English Usage with ICECUP', *ET*10, 1994, pp. 21–24) which introduced the International Corpus of English (ICE) and its search facility ICECUP (the ICE Corpus Utility Programme). The present paper has a two-fold aim: to (re-)acquaint readers with ICE and discuss the latest developments in ICECUP – including its recent release on CD-ROM.

Introduction

The International Corpus of English was initiated by Sidney Greenbaum, whose aim was to set up a number of identically constructed corpora (for the purpose of grammar research) in the world's various English-speaking countries. To date, some 18 research groups are part of the ICE project, among them ICE-GB, ICE-USA, ICE-AUS and ICE-NZ. The map in Figure 1 shows the countries in which ICE groups are based. [Not all these corpora are at the same stage of development. ICE-GB is currently the only completed corpus.]

What do we mean when we say that the various ICE corpora have been 'identically constructed'? It means that each corpus contains an identical number of words, and is constructed by making use of the same text categories. Each of the national corpora is one million words in size and made up of 500 texts of approximately 2,000 words each. They all contain both written and spoken material. The



Figure 1: Countries participating in the ICE project

DOI: 10.1017/S0266078407002052
English Today 90, Vol. 23, No. 2 (April 2007). Printed in the United Kingdom © 2007 Cambridge University Press

https://doi.org/10.1017/S0266078407002052 Published online by Cambridge University Press

Figure 2: Text categories in the International Corpus of English

Spoken (300)	Dialogues (180)	Private (100)	face-to-face conversations (90) phone calls (10)
		Public (80)	classroom lessons (20) broadcast discussions (20) broadcast interviews (10) parliamentary debates (10) legal cross-examinations (10) business transactions (10)
	Monologues (120)	Unscripted (70)	Spontaneous commentaries (20) unscripted speeches (30) demonstrations (10) legal presentations (10)
		Scripted (50)	broadcast news (20) broadcast talks (20) non-broadcast speeches (10)
Written (200)	Non-Printed (50)	Non-professional writing (20)	student essays (10) student examination scripts (10)
		Correspondence (30)	social letters (15) business letters (15)
	Printed (150)	Academic writing (40)	humanities (10) social sciences (10) natural sciences (10) technology (10)
		Non-academic writing (40)	humanities (10) social sciences (10) natural sciences (10) technology (10)
		Reportage (20)	press news reports (20)
		Instructional writing (20)	administrative/regulatory (10) skills/hobbies (10)
		Persuasive writing (10)	press editorials (10)
		Creative writing (20)	novels/stories (20)

texts date from the period 1990–96, i.e. the written material was written or published during this period, and all the spoken material was recorded then. The writers and speakers in the corpus were all over the age of 18, and educated in English. Furthermore, they were born in the country in whose corpus they are included, or moved there at an early point in their life. Samples of speech and writing by both men and women of a wide range of age groups are included. However, this does not mean that men and women are proportionally included in samples of the corpora, simply because in certain professions men are more dominant than women, and vice versa. See Figure 2 for the text categories.

After the corpus was collected we annotated it. This was done at three levels: textual markup, wordclass tagging and syntactic parsing.

The first level includes orthographic transcription, marking overlapping speech, adding discourse features (in the case of spoken texts), as well as sentence boundaries, paragraphs, headings, etc. (in the case of written texts).

Wordclass tagging, as the name suggests, involves assigning a part of speech label to each and every word in the corpus. Naturally, it would have been too laborious to do this manually, so for this purpose a tagger developed at the University of Nijmegen was used. The tagger assigns wordclass labels partly on the basis of a simple lookup in a lexicon, and

https://doi.org/10.1017/S0266078407002052 Published online by Cambridge University Press

partly by applying morphological rules. Word-class labels consist of a main label, such as 'N' for noun and 'V' for verb. In most cases, these are followed by additional features in brackets. For example, in the tagging of verbs, we indicate the transitivity of the verb and its form. The ICE tagset distinguishes 19 different word-classes, and the possible combinations of wordclasses and their features amounts to 262. Figure 3 shows an example of a sentence with added wordclass labels.

The most complex level is the level of syntactic parsing. Again, this was done automatically, using a parser developed at the University of Nijmegen. The parser uses a complex formal grammar to analyse each unit at the word, phrase and clause level. This analysis is then displayed in the form of a tree diagram. An example is shown in Figure 4.

Each node on the tree indicates function and category, and may carry additional features, such as clause type. The ICE-GB corpus contains around 84,000 trees, so it is a very valuable resource for syntactic studies of English. To enhance this value further, we are planning to make available in the near future the original recordings of our spoken texts. These have been computerized and linked to the transcriptions.

Corpora are of little value unless they can be *exploited*. To this end a corpus search programme was developed at the Survey of English

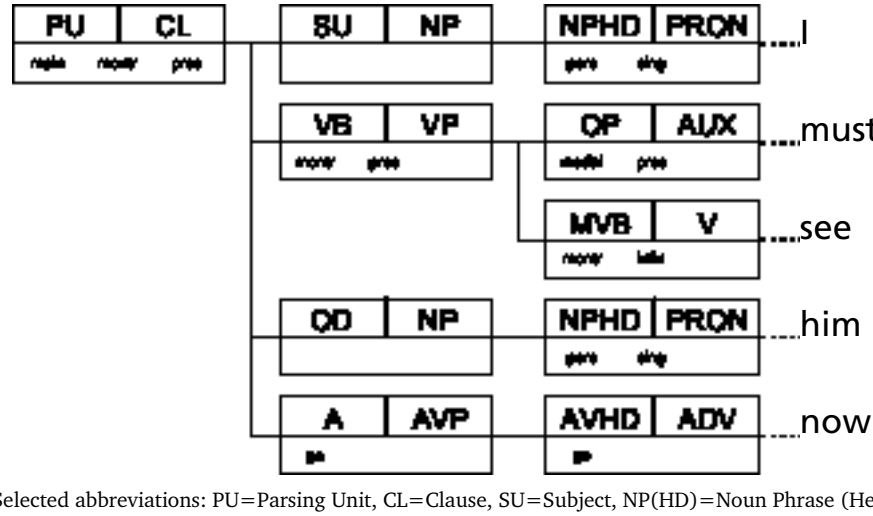
The tagged sentence 'I must see him now' (S1A-045 #132).

I PRON(pers,sing) **must** AUX(modal,pres) **see** V(montr,infin) **him** PRON(pers,sing) **now** ADV(ge)

Figure 3

Usage, which we call ICECUP. What can ICE-CUP do? At the most simple level it can do lexical searches, i.e. search for words specified by the user. For example, one can very quickly retrieve a list of all references to the word *thus*, and it turns out that there are 138 instances in ICE-GB. Naturally, the user can specify whether the search should be conducted across the corpus, or in one or more of the subdomains only, for example only in the spoken material, or only in the newspaper category, etc.

Far more interesting is the possibility of searching for particular syntactic constructions or functions. As an example of the first type of search, consider a user who is interested in cleft constructions, that is, constructions like *It was John who broke the window*. In the old days researchers had to do a manual search of a corpus to find instances of clefts, a very laborious task. Recent corpora allow for automatic searches (e.g. for the word *it*), but users will inevitably find too much material, which will



Selected abbreviations: PU=Parsing Unit, CL=Clause, SU=Subject, NP(HD)=Noun Phrase (Head), PRON=Pronoun, (M)VB=(Main) Verb, OP=Operator, AUX=Auxiliary Verb, A=Adjunct, OD=Direct Object

Figure 4: The parsed sentence 'I must see him now' (S1A-045 #132) in the form of a tree diagram.

https://doi.org/10.1017/S0266078407002052 Published online by Cambridge University Press

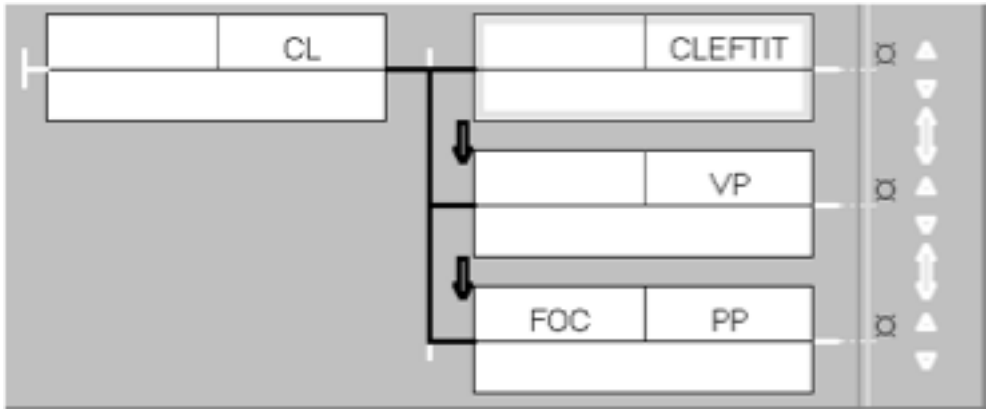
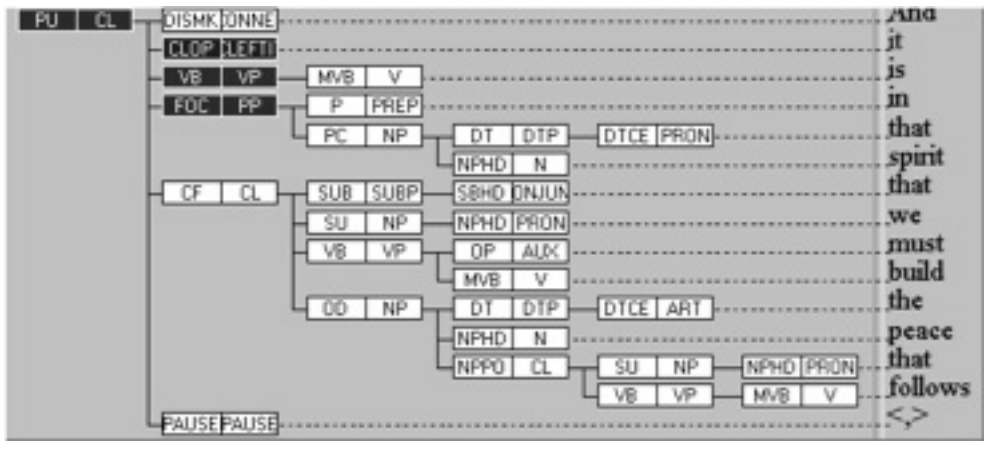


Figure 5: A Fuzzy Tree Fragment specifying a search for cleft it followed by a verb phrase and a prepositional phrase.



Selected abbreviations: PU=Parsing Unit, CL=Clause, CLOP=Cleft Operator, CLEFTIT=Cleft it, (M)VB=(Main) Verb, FOC=Focus, PC=Prepositional Complement, DTP=Determiner Phrase, DTCE=Central Determiner, NP(HD)=Noun Phrase (Head)

Figure 6: Tree diagram representation showing a match for the search pattern in Figure 5.

still be quite a big job to 'clean up'. Alternatively, users can construct complex search queries using Logic. Logic is clear and formal, but it is very difficult to use, in that it requires a reasonably sophisticated knowledge. Of course, users can invest time in learning to create Logic-based queries, but this may not be the most efficient way of using one's time, and it is prone to error. In ICE-GB we have explicitly labeled all instances of cleft *it* as CLEFTIT, so that they can be retrieved almost instantly by simply entering this label.

However, we may wish to construct more complex searches than this. Suppose, for example, that we wish to restrict the search to clefts which have a prepositional phrase in the focus slot, rather than a noun phrase. So we are interested in constructions like *It was in London that we met*, where *in London* is the focus. To support searches of this kind, the survey has developed a sophisticated and novel search system. Instead of writing complicated logical expressions, the system we have devised is based on users drawing a simple model of the tree fragment that they wish to search for. We call such a model a *Fuzzy Tree Fragment* (FTF). The FTF for our complex query is shown in Figure 5.

FTFs such as this can rapidly be constructed with great ease by users with the help of the

https://doi.org/10.1017/S0266078407002052 Published online by Cambridge University Press

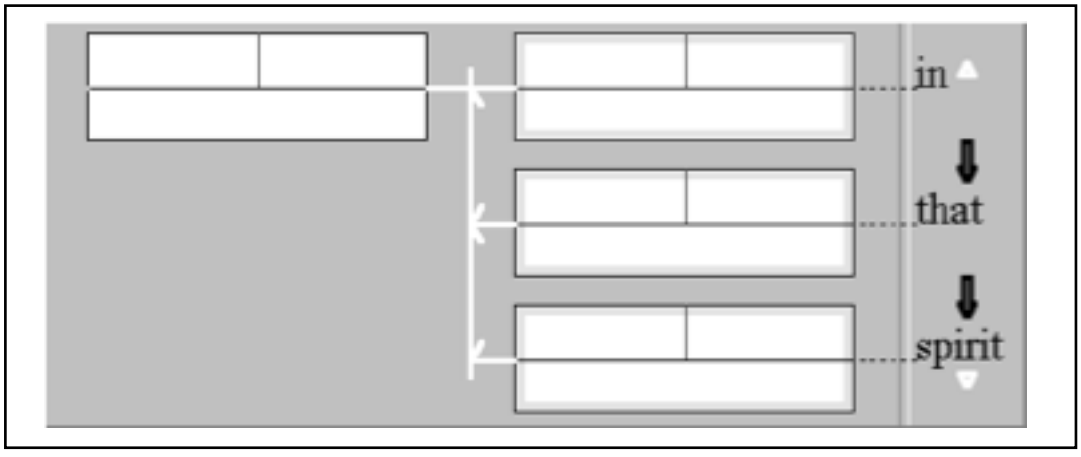


Figure 7: A lexical search string in the form of an FTF

ICE-GB manual (supplied with the corpus). The associated, cumbersome and forbidding search query in Logic would read like this:

_*w*_*x*_*y*_*z*.((*w* = [<unk>,CL]) and (*x* = [<unk>,CLEFTIT]) and (*y* = [<unk>,VP]) and (*z* = [FOC,PP]) and **Parent**(*x*, *w*) and **Parent**(*y*, *w*) and **Parent**(*z*, *w)* and **SiblingAfter**(*x*, *y*) and **SiblingAfter**(*y*, *z*)).

This FTF matches forty-nine constructions in the ICE-GB corpus. Figure 6 shows one of these matches, in which our FTF pattern is highlighted.

FTFs have the advantage of allowing the user to leave information unspecified (hence the term 'fuzzy'). For instance, in Figure 5, we have not specified any details for the verb phrase, such as its tense, or the presence or absence of auxiliaries. However, we can add these and many other details if we wish, by selecting them from pull-down menus of functions, categories, and features.

Words can also be incorporated into FTFs, so that we can use FTFs to search for word (or word and tag) sequences. In this way, we can relate word sequences and tree elements in the same structure. The following FTF is generated by the computer automatically in order to look for the word sequence *in that spirit* (Figure 7).

We believe that Fuzzy Tree Fragments are fairly intuitive, although they can become quite complex. This is because we allow links between nodes to be defined within a range of different 'strengths', from completely unspecified ('unknown'), to strictly specified ('must be directly connected to'). The FTF search facility is an innovative and easy-to-use tool, not just for professional linguists, such as syntacticians, lexicographers, sociolinguists, etc., but for everyone interested in the English language. It requires some knowledge of how to use computers (e.g. Windows-based software), but does not require a knowledge of logic to construct queries. We have recently released the ICE-GB corpus on CD-ROM, together with ICE-CUP, which incorporates the FTF facility. The CD-ROM is supplied with a manual and a library of ready-made tree fragment templates which users can use as they are, or modify, to suit their own needs. ■

Note More ICE-GB information is available at http://www.ucl.ac.uk/english-usage/ and at: Survey of English Usage, Department of English Language and Literature, University College London, Gower Street, London WC1E 6BT <ucleseu@ucl.ac.uk>

https://doi.org/10.1017/S0266078407002052 Published online by Cambridge University Press