The study of complex grammatical patterns tends to be neglected by corpus linguists, the main reason being that such phenomena are much more difficult to extract from a corpus than simple words or tags. I demonstrate in this article that, although the desirable parsed corpora and appropriate software are not always available, the retrieval of syntactic structures can be automated to a certain extent. A number of corpus-based grammatical analyses, as well as a pilot study of causative structures with make, illustrate the various alternative strategies that can be used to this effect.

**KEYWORDS:** automatic retrieval, grammatical phenomena, syntactic structures, causative structures, parsed corpora
1. Setting off on an adventure

Since corpus linguistics has asserted itself as one of the major trends in the field of linguistic research, it has given rise to many an interesting study, all the more reliable since they are based on authentic language. Many of these studies are concerned with lexical items, both simple (e.g. *well* (Svartvik 1980)) and complex (e.g. *say when* (Minugh 1995)). Such items can be retrieved from corpora very easily thanks to the most basic concordancers, regardless of whether the corpus used is annotated or not. Far fewer linguists, on the other hand, have dared venture into the corpus-based investigation of complex grammatical phenomena such as syntactic structures. This results from the difficulty connected with the automatic retrieval of such items. Some 20 years ago, Olofsson (1981: 14) already observed that

> [l]exical words and constructions based on lexical words can be excerpted with speed and ease, even without pre-editing of the material, at least as long as the problem of homonymy does not complicate the investigation. Grammatical phenomena, however, often cannot be linked with lexical words, which means that an investigation of such matters can only utilize the computer processibility to a very limited extent.

Although techniques have considerably improved since then, it is nonetheless true that, for those who want to avoid a manual, laborious collection of data, the retrieval of complex grammatical phenomena often requires highly sophisticated tools and/or heavily annotated corpora – materials that are not always available – and therefore turns into a task that might remind one of the quest for the Holy Grail.
This article deals with the obstacles involved in automatically retrieving syntactic structures. After emphasising the role of automation in corpus linguistics, it surveys the types of grammatical phenomena that can be investigated, as well as the different methods possible and the parameters that influence this methodological choice. It also takes a snapshot of the strategies prevalent today, thus showing how and why syntactic structures and other complex phenomena tend to be neglected in favour of more simple queries. Through a general description of the state of the art in tagging and parsing and through the case study of causative structures with *make*, it is demonstrated that, although appropriate tools and/or corpora are not always part of the linguist’s armoury of resources, there are ways to retrieve syntactic structures with reasonable precision and recall rates.

2. Setting the scene: automation in corpus linguistics

The analysis of language on the basis of authentic texts is not something new. Poutsma’s (1926) *Grammar of Late Modern English*, for instance, is illustrated with naturally-occurring sentences. What is new, however, is the way we have been working with corpora since the middle of the 20th century. Thanks to the great advances made in information technology, today’s linguist can use his/her own personal computer to store huge bodies of text and search through them automatically in a matter of seconds. As Kennedy (1998: 5) puts it, [c]orpus linguistics is (...) now inextricably linked to computers’ – so much so, actually, that the very meaning of the word ‘corpus’ has come to imply machine-readable (Mason 2000: 4).
The advent of computers among corpus linguists has made it possible to automate a number of tasks that used to be carried out by hand. These tasks can belong to either of two main processes, namely annotation and retrieval. Annotation of a corpus consists in applying different tools to plain orthographic (or ‘raw’) text in order to enhance it with linguistic information. These tools, which Kirk (1994: 19) refers to as ‘intelligent’ tools, include pos-taggers and parsers. Pos-taggers attach to each word of the corpus a label indicating its part of speech (pos). The pos-tagged corpus thus obtained can then be submitted to a parser, which will mark and analyse the syntactic constituents of the text (phrase, clause and sentence structure). Each of these processes can be automated to a certain extent, ranging from automatic annotation with manual post-editing to fully automatic annotation. The success rates vary accordingly and, as a rule, taggers perform better than parsers (see section 3.3). Retrieval, on the other hand, concerns the identification and extraction of a target word or construction. Here too automation is possible, thanks to so-called ‘dumb’ tools (Kirk 1994: 19), concordancers which retrieve the information and can organise it in different ways depending on the researcher’s needs. And here too, the automatic stage may be followed by manual post-editing so as to improve the accuracy of the results.

The focus of this article will be on retrieval, rather than annotation. In other words, corpora will be considered here as finished products and the position taken will be that of a linguist with no knowledge in programming and who therefore has to rely on existing text-retrieval software. This, unfortunate as it may be, still reflects the situation many linguists are in, for, as rightly observed by Mason (2000: 3), ‘the computer, powerful though it is, is not an easy tool to use for someone with a humanities background, and so its use is generally restricted to whatever ready-made programs are available at the moment’.

1
3. The story of linguists’ lives

3.1. Perambulation into the Deep Forest of grammatical phenomena

The various phenomena of the English language do not get the same degree of attention within the field of corpus linguistics. As Kennedy (1998: 88) puts it, corpus-based studies ‘present a rather unsystematic coverage of aspects of English’. More particularly, a number of scholars complain that syntax tends to be neglected in favour of lexis. Oostdijk and de Haan (1994: 41), for instance, point out that

[large-scale quantitative studies of syntactic structures and phenomena are long overdue. While word frequency counts and concordances have been a common good to the linguistic community for quite some time now, corpora that have undergone a detailed syntactic analysis are few, and so are the quantitative studies that are based on these.

In order to understand the origin of this state of affairs, I will start by giving an overview of the different grammatical phenomena and how they can be retrieved from a corpus. Figure 1 presents a threefold distinction between lexically-based, non lexically-based (but form-based) and non form-based grammatical phenomena.

The first category can be referred to as lexico-grammar, that is, grammatical research centred on morphemes or words (see Kennedy 1998: 121-154). Most of the time, an automatic search on a raw corpus with a basic concordancer will suffice. This is the case of an unambiguous grammatical word such as the article an, or the closed class of modals (cf. Mindt 1995), of which all the members can be enumerated and searched for.
Sometimes, however, a grammatical word can be ambiguous and belong to different parts of speech. Several scenarios are then possible. First, the other use(s) of the target word may be very infrequent. A search on the, for instance, will also produce a number of matches where the is adverbial, as in the sooner, the better. But this use is so rare (less than 0.3% of all the occurrences of the in ICE-GB) that manual weeding out is still perfectly feasible – and might even be preferable if one does not want to depend on the (sometimes inaccurate) tagging of a corpus. Second, a linguist may be interested in all the uses of an ambiguous grammatical word. S/he might for example want to study that as a demonstrative determiner, demonstrative pronoun, conjunction of subordination and relative pronoun. Here again, an automatic (lexical) search on a raw corpus will provide all the data needed. This is the approach taken in Altenberg’s (1994) study of the multifunctional word such. Only when the linguist is interested in one particular use of a grammatical word and the other uses are relatively frequent should a pos-tagged corpus be used. Thus, considering that the use of in as an adverb particle represents only 3% of all its occurrences (cf. Meunier 2000: 67), a purely lexical search would involve far too much manual post-editing. The methodology to adopt is therefore a lexical search coupled with a search on pos, e.g. IN_RP will retrieve all the instances of in as a particle (RP = tag for particles in the CLAWS6 tagset used to annotate the BNC Sampler).

When the grammatical phenomenon under investigation is not lexically-based, i.e. is centred on the sentence rather than on morphemes or words (see Kennedy 1998: 154-174), it becomes impossible to list all its members. An automatic search for an open class should therefore involve the use of a tagged corpus (cf. Fang’s (1995) study of the
infinitive). If the corpus has been tagged correctly, such queries should not produce any noise (i.e. irrelevant material).

Things are more difficult for other types of non lexically-based grammatical phenomena. The ‘royal road’ to the automatic retrieval of these phenomena is the use of a parsed corpus, since it contains information about phrases (e.g. NPs), clauses (e.g. relative clauses) and sometimes functions\(^4\) (e.g. subjects). However, for various reasons that will be outlined later, this option is not always available and the linguist may have to make use of a pos-tagged corpus instead. Someone interested in relative pronouns used with a stranded preposition (non-contiguous combination of parts of speech), as in *This is the person that I talked to yesterday*, and who has no access to a parsed corpus can use a tagged corpus and create an algorithm requiring the machine to retrieve all the sentences where a relative pronoun is followed by a preposition within a span of X to Y words (X and Y to be determined by a pilot study).\(^5\) It should be borne in mind, however, that such a query will entail a certain amount of manual post-editing in order to discard the matches that do not correspond to the pattern sought. As we move from syntactic phrases to syntactic structures, it becomes increasingly difficult to carry out an automatic search on a tagged corpus. Moreover, depending on the level of specificity of the query, the results may have a poor precision rate (i.e. ‘the proportion of retrieved materials that are relevant’ (Salton 1989: 248)) and/or a poor recall rate (i.e. ‘the proportion of relevant materials retrieved’ (*ibid.*)).

Let us now have a closer look at the different alternatives available to retrieve syntactic structures automatically, taking the example of the construction ‘*see + NP + infinitive*’, as in *I saw the towers collapse*. Theoretically, as this structure includes a lexical item, it could be retrieved manually from a raw corpus. In the present case, however, such a method would involve far too much weeding out, since this pattern
accounts for only 8.5% of the total occurrences of see in ICE-GB. This leaves us with two possibilities, viz. a search on a tagged corpus and a search on a parsed corpus. With a tagged corpus, the procedure consists of looking for the lemma see, followed by a number of words and an infinitive. Here again, a prior pilot study is indispensable in order to determine the ideal distance between the verb and the infinitive, as well as to assess the precision and recall rates of such a retrieval method. If these results are too poor (or the linguist too lazy!), it might be necessary to turn to a parsed corpus – if such a corpus is available – and require the retrieval software to extract all the ‘see NP infinitive’ sequences. The concordance lines thus obtained should retrieve most of the occurrences of see used with an NP and an infinitive\(^6\) and should not contain too much noise. The automatic retrieval of syntactic structures may sometimes turn out to be even more complicated in cases where the structure has several meanings. This situation is best illustrated by the pattern ‘have + NP + past participle’. As pointed out in Palmer (1988), this pattern can have different readings including the causative and experiential readings. Yet, so long as substantial advances in the field of semantic annotation are not made, such interpretations cannot be disambiguated by computer, so that, whatever the method chosen (search on a tagged or parsed corpus), the data might need a considerable amount of weeding out – unless, of course, one is interested in all the uses of the structure (cf. Ikegami 1989).

Finally, with the highest degree of abstraction, grammatical phenomena encompassed within a non form-based category are certainly the most difficult to retrieve. A semantically annotated corpus can be helpful here, cf. Thomas and Wilson’s (1996) corpus, which has been annotated with the semantic tagger SEMTAG and contains semantic categories such as cause or modality. However, there are so few semantically
annotated corpora that, most of the time, the automatic retrieval of such a category in its entirety is simply beyond the bounds of possibility and that an exhaustive search can only be done by hand (see below, however, for an alternative solution).

What precedes might give the (false) impression that the choice of a methodology is solely dependent on the type of grammatical phenomenon under investigation. In fact, there are a number of additional parameters that influence the method used.

First, the methodological choice may be determined by the availability of corpora. As we have seen, grammatical phenomena that are not lexically-based ideally require the use of a tagged or parsed corpus. However, such corpora are not always available, which explains why some scholars have to opt for a less richly annotated corpus and fall back on a less automatic collection of data. Thus, for his study of that- and zero-object clauses, Rissanen (1991) has chosen a largely manual approach, picking out the occurrences of all verbs that take an object clause with that, so as to find the possible instances of the zero link. The reason for this choice is that the corpus he used, the Helsinki Corpus, was neither tagged nor parsed at the time.

The second parameter to take into account when choosing a methodology is the frequency of the phenomenon investigated. I have already alluded to this criterion in connection with ambiguous grammatical words. Similarly for other types of phenomena, what is manually manageable with a very frequent structure will turn out to be virtually impossible with relatively infrequent structures. Some years ago, for instance, Aarts (1971) carried out a study on NP structures. The NPs (about 8,000) were retrieved manually from a 72,000-word corpus. Although nowadays parsed corpora exist from which NPs can be automatically retrieved, a manual search would still make sense, as NPs are extremely common in language. By contrast, it is a titanic task to
manually look for all the occurrences of a low-frequency structure like AdjP + of + Det. + N, e.g. *how big of a problem* (cf. Trotta and Johansson 2001).

Thirdly, a distinction should be made between identification of occurrences of a construction and extraction of the complete construction. If the goal of the automatic analysis is simply to identify the existence of a particular construction and examine some of its characteristics, a good tagged corpus should do. For instance, the occurrences of prepositional phrases can be retrieved by carrying out a search on the ‘preposition’ pos-tag. All prepositional phrases will then be identified and it will be possible to study, say, the animate or inanimate nature of the (pro)noun following the preposition. If, on the other hand, the goal is to extract the whole construction, including its ending boundary, a parsed corpus is required.

Moreover, the degree of exhaustiveness aimed at can also influence the choice of a particular method. If the linguist does not aim at a fully comprehensive study of his/her research question, s/he may deliberately close an open class, that is, restrict the number of items to investigate. Thus, Blagoeva (2001) studies conjunctions in learner English, but she limits her analysis to 50 words and phrases. Similarly, Løken’s (1997) analysis of the (non form-based) category of possibility is limited to a certain number of verbs (the verbs *can, could, may, might* and their Norwegian translations, as well as Norwegian *kunne* and its translations in English). This selection can be an elegant solution to circumvent the lack of appropriate annotation of the corpus used.

Finally, considerations about the tagging or parsing itself may lead one to turn away from the most comfortable solution. Let us imagine that a scholar wishes to study the behaviour of adjectives ending in *–ed*. Relying on the tagging of a corpus necessarily means accepting the implicit model of language imposed by the tagger (automatic tagger or human annotator/corrector), in this case the distinction between
adjectives ending in –ed and past participles. In this sense, and as rightly observed by Lorenz (1999: 36), the use of an annotated corpus can be said to predetermine the search results to a certain extent. Moreover, tagged and parsed corpora are not always accurate, especially when they have been annotated without any human intervention. So a more laborious manual procedure may sometimes be preferred to an automatic retrieval.

3.2. The easy way out

The types of grammatical phenomena outlined above do not all receive an equal amount of attention. As I will show in this section, there is a general tendency among today’s linguists to address research questions that are easy to investigate and neglect those whose investigation requires more effort in terms of the retrieval of the data. With this aim in mind, I examined the contents of the first five volumes of the *International Journal of Corpus Linguistics* (IJCL) (1996-2001) and the proceedings of the 22\textsuperscript{nd} ICAME conference (see De Cock et al. 2001).

The good news is that there seems to be a gradual shift in focus from lexis to grammar in the field of corpus linguistics. While the volumes of IJCL contain more lexical than grammatical studies (43\% of the articles deal with lexical items and 34\% include the search for a grammatical phenomenon), the talks given at ICAME 2001 show a majority of grammatical studies (16\% of lexis vs. 43\% of grammar). However, it turns out that not all grammatical categories fare equally well.

As appears from Figure 2, the most frequently studied category is that of open part of speech (33\%). Yet, if we group grammatical words and closed pos together, as forming the class of lexically-based phenomena, easily retrievable from a raw corpus, it
becomes the biggest category (35%). Nevertheless, the prevalence of the studies devoted to (open and closed) parts of speech is quite remarkable and is most probably to be explained by the greater availability and reliability of tagged corpora (see section 3.3). The remaining categories (phrases, functions, clauses, structures and non form-based categories), on the other hand, represent a relatively small proportion of the grammatical studies (9%, 4%, 6%, 9% and 4%, respectively). This comes from the difficulty involved in the automatic retrieval of syntactic categories. As noted earlier, it implies either, if one uses a tagged corpus, the creation of a (sometimes complex) algorithm and, incidentally, quite a lot of manual weeding out, or the availability of a parsed corpus and an appropriate query system, materials that are still too rare and not reliable enough (see following section).

Figure 2: Types of grammatical phenomena discussed in IJCL (1996-2001) and ICAME 2001
The types of corpora exploited by the linguists who tackle grammatical issues in IJCL and ICAME 2001 (Figure 3) seem to correlate quite well with the categories of phenomena they look into. Tagged corpora are more frequently used (43%) than raw and parsed corpora (28% and 29%, respectively). That tagged corpora meet with such an enthusiastic reception is no surprise. As a matter of fact, they offer many more possibilities than plain orthographic corpora, especially in the field of syntax. As pointed out by Kennedy (1998: 102-103), ‘they can provide information not only on whether an individual form occurs more often, say, as a noun or a verb, but tagging can also show the frequency and distribution of word classes in a corpus’. Moreover, as we will see below, tagged corpora are both easily available and mostly accurate. By comparison, parsed corpora, which are even more powerful than tagged corpora in that they contain structural information, are less easily available and often not sufficiently reliable and/or detailed. Reliable and detailed parsed corpora do exist (e.g. ICE-GB), but tend to be rather small and are therefore not suitable for the investigation of any phenomenon (cf. infrequent structures).

![Figure 3: Types of corpora used in the grammatical studies of IJCL (1996-2001) and ICAME 2001](image-url)
A cross-tabulation of the two parameters, type of grammatical phenomenon investigated and type of corpus used (see Table 1), brings out an interesting fact, namely there is no one-to-one correlation between the two values. In other words, some linguists use more richly annotated corpora than needed, while others use less richly annotated corpora than their studies would ideally require.

<table>
<thead>
<tr>
<th>Grammatical word → unambiguous</th>
<th>Raw corpus</th>
<th>Tagged corpus</th>
<th>Parsed corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>→ ambiguous</td>
<td>**</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>POS → closed</td>
<td>***</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>→ open → single</td>
<td>*</td>
<td>**</td>
<td>/</td>
</tr>
<tr>
<td>→ multiple → contiguous</td>
<td>**</td>
<td>***</td>
<td>*</td>
</tr>
<tr>
<td>→ non-contiguous</td>
<td>*</td>
<td>*</td>
<td>/</td>
</tr>
<tr>
<td>Syntactic phrase</td>
<td>/</td>
<td>/</td>
<td>***</td>
</tr>
<tr>
<td>Syntactic function</td>
<td>/</td>
<td>/</td>
<td>**</td>
</tr>
<tr>
<td>Clause</td>
<td>/</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Syntactic structure → unambiguous</td>
<td>/</td>
<td>**</td>
<td>*</td>
</tr>
<tr>
<td>→ ambiguous</td>
<td>/</td>
<td>/</td>
<td>*</td>
</tr>
<tr>
<td>Non form-based category</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

/ = no study  
* = 1 – 2 studies  
** = 3 – 5 studies  
*** = 6 – 8 studies

Table 1: Cross-tabulation of the grammatical phenomena discussed and the corpora used in IJCL (1996-2001) and ICAME 2001

A good example of the first tendency is Facchinetti’s (2001) analysis of the modal verb *may*. While as a grammatical word it could be retrieved even from a raw corpus, Facchinetti carried out her study on ICE-GB (parsed corpus). Using a more richly annotated corpus than necessary does not pose any problem for the retrieval. What is more, it can be justified by the composition and characteristics of the corpus. In
this case, ICE-GB has the advantage of being a newly available and well-balanced 1,000,000-word corpus, containing a large proportion of spoken data and including useful sociolinguistic variables such as age or gender of the speaker/writer.

Although the use of a less richly annotated corpus than necessary can also be justified, it is more problematic, for it implies that more work will have to be done by hand. A common practice, then, is to use one of the ‘do-it-yourself’ tricks that can save poorly equipped linguists from getting lost in the forest of data. One such trick is the selection of a number of items. This selection is sometimes arbitrary, but most of the time it is based on frequency (only the most frequent items are taken into account). Valera and Rizo-Rodríguez (1998), for example, investigate adjectives in supplementive clauses (cf. *The teacher, very red in the face, gave Hal a smack*). Since such clauses are not encoded in their corpus (LOB), they selected 671 adjectives with 20 or more occurrences and went through each match in order to determine the adjectives (116) for which examples of supplementive clauses were available. By the same token, Biber’s (1996) analysis of the valency patterns of *tell* and *promise* is based on 200 randomly chosen tokens for each verb. Another DIY-trick is the use of a relatively small corpus (typically between 100,000 and 300,000 words), which can be read through in order to retrieve all the occurrences of the target item (cf. Aarts 1971, based on a 72,000-word corpus, or Meyer 1992, based on three corpora of approximately 120,000 words each). This solution, however, should only be envisaged for high-frequency phenomena, so as to ensure sufficient evidence for one’s claims. Thus, Geluykens (1992), whose study is based on a corpus of 450,000 words, ends up with a mere 149 instances of left-dislocation, which is not many for any authoritative conclusions to be drawn.
Before we leave this section on the methodology applied in current grammatical research, it should be noted that, regrettably, some scholars fail to give a (satisfactory) account of the retrieval method they used. Ball (1994: 296) observes that ‘it is more common in reports of corpus-based research for the search method to be left unspecified’. This, it seems, is particularly the case when the search has been carried out by hand, resulting in what could be called a ‘manual methodological gap’. It is as if, in this technological era, corpus linguists were embarrassed to admit – oh shame! – that they prefer pencil and paper to mouse and computer. However, ‘[h]euristics should be reported along with the findings, and should be treated with skepticism by the reader’ (Ball 1994: 296), for only by examining the researcher’s way of working can the reader truly assess the reliability of the data obtained. As for computerphobes, they should be reassured for three reasons. First, we have seen that the use of a manual method can be justified in some cases, e.g. when the phenomenon is very frequent or when appropriately annotated corpora are not available. Second, even when adequate corpora or tools are not available, there are several ‘tricks of the trade’, such as restricting the number of items analysed, which make it possible to automate the search to some extent (normal heroes always make a detour!). Finally, judging from the progress achieved over the last few years, more powerful and user-friendly tools are likely to come into use in the near future, thus enabling even computer-illiterate linguists to carry out their searches fully automatically.
3. 3. In search of the magical potion

As mentioned previously, the ideal method to retrieve complex grammatical phenomena such as syntactic structures is to make use of a parsed corpus, which has been annotated with useful syntactic information. I will show in this section, however, that parsing has not quite reached a mature level yet and that, in de Mönnink’s words (2000: 41), ‘[t]o date, the availability of [corpora annotated with detailed syntactic information] (…) still leaves much to be desired’. Consequently, it is often necessary to turn to the next best method, namely the use of a tagged corpus which, though usually involving more manual post-editing, still allows for a considerable range of possibilities.

A few years ago, Black (1993: 5) referred to the ‘dismal state of the art in the parsing of English’. Although advances have been made since then, several serious problems remain. To begin with, the number of parsed corpora publicly available is limited. The main representatives are the Nijmegen corpus, the Penn Treebank, the SUSANNE and CHRISTINE corpora, ICE-GB, as well as a couple of historical corpora such as the Penn-Helsinki Parsed Corpus of Middle English (see Souter and Atwell (1994) for details and references on some of these). Most of these corpora present one serious drawback, though, namely their relatively small size. They usually range in size from some 100,000 words to about 1,000,000 words, but as we shall see in the case study of causative structures, even one million words is not always enough to investigate a phenomenon thoroughly. Larger corpora do exist, however they generally use a less refined annotation scheme (cf. first phase of the Penn Treebank, see Marcus et al. 1993), hence a lack of precision and information. As aptly observed by Sampson,9 ‘what the
research community ultimately needs is very large databases of language analysed in
very great detail’.

As for automatic parsers, they have at present ‘not approached the level of
accuracy achieved by automatic word-class taggers’ (Kennedy 1998: 232). Belz (2001)
reports an average recall and precision rate of 82% and 83% respectively for several
rate ranging from 94% to 98% for their TOSCA/ICE parser, but it should be emphasised
that only 48% of the utterances receive a single analysis. For the remaining utterances,
human intervention is needed in order to select the correct analysis. The ENGCG
(English Constraint Grammar) parser, described as a robust system capable of parsing
large amounts of unconstrained text in a limited amount of time, assigns a unique
syntactic label to some 85% of the word-forms, with an error rate of 3% (see Karlsson
1994: 130), but its output is ‘a partial rather than complete parse’ (Leech 1997: 11).

Getting a corpus parsed accurately, however, is only half the battle, as querying
a parsed corpus brings its share of problems too. Souter and Atwell (1994: 154) point
out that ‘[t]here are few standards to ensure that parsed corpora all adhere to the same
format. Consequently the potential for combining corpora, comparing corpora, or
inspecting corpora with a single general-purpose tool is reduced’. More particularly,
there are no accepted standards for label-set or tree structure representation, which has
the effect of ‘preventing many corpora from being easily imported into software
packages’ (p. 155). Thus, ICE-GB is distributed with ICECUP, a program specifically
designed to process and analyse this corpus, but in practice it turns out that ICECUP
cannot (yet) be applied to other parsed corpora. As for the Bank of English, it was
syntactically annotated with the ENGCG parser (see Järvinen 1994), but this parsed
version is very unlikely to be made available, for want of a query system that could handle the syntactic annotation of the corpus (J. Clear, personal communication).

These major drawbacks explain corpus researchers’ reluctance to use parsed corpora. They also suggest that, while the magical potion to retrieve syntactic structures does exist, it usually takes a (computer) wizard to make it truly efficient.

As a consequence, someone embarking on the study of a syntactic structure will often have to make do with a pos-tagged corpus. Though less richly annotated than a parsed corpus, it presents several advantages over the latter. First, there is a large number of ready-made tagged corpora publicly available (e.g. LOB, Brown, BNC, Bank of English), that can be used in conjunction with most concordancers. Second, current automatic taggers achieve a high degree of accuracy. Thus, the ‘generally desirable “target” of 95%’ (Coniam 1998: 242) seems to be reached by most taggers now, with a success rate ‘typically in the region of 96-7 per cent’ (Leech 1991: 15). Karlsson (1994: 130) even reports an error rate not exceeding 0.3% for the 94%-97% of words that are assigned an unambiguous tag. Another advantage is that tagged corpora can be very big. The British National Corpus (BNC) is a 100,000,000-word corpus and the Bank of English contains approximately 450 million words at the moment, but is constantly growing. Such corpora can therefore be used to investigate even infrequent phenomena. Last but not least, one should not underestimate the kinds of analyses that are possible with a good tagged corpus. This has already been hinted at in section 3.1. and will also appear from the case study of causative structures. Indeed, Biber’s (1988) multidimensional study proves that the possibilities of investigation on the basis of tagged corpora are considerable. Thus, Biber studied no less than 67 linguistic features, some of which involved the creation of one or more very complex algorithm(s) combining
lexical items, pos and unspecified strings of words. In order to retrieve present participial clauses (e.g. Stuffing his mouth with cookies, Joe ran out the door), for instance, he carried out the following query (Biber 1988: 233): \(^{12}\)

\[ T#/\text{ALL-P} + \text{VBG} + \text{PREP/DET/WHP/WHO/PRO/ADV} \]

Sophisticated though it is, Biber’s study also reminds us of the limitations of automated text analysis. First, manual intervention is still necessary in some cases. The results of the above algorithm, for example, need to be edited by hand in order to exclude participial forms not having an adverbial function. Second, some features have to be restricted to a number of representatives. The category of downtoners contains a mere 13 items, viz. almost, barely, hardly, merely, mildly, nearly, only, partially, partly, practically, scarcely, slightly and somewhat (Biber 1988: 240). Similarly, because is the only causative adverbial subordinator to be studied, since it is ‘the only subordinator to function unambiguously as a causative adverbial’ (Biber 1988: 236). Third, as convincingly demonstrated by Ball (1994: 296), some of Biber’s algorithms miss part of their target (recall problem or You-Don’t-Know-What-You’re-Missing problem). Ball gives the example of zero complementizers in subordinate clauses and shows that the search criteria as they are stated by Biber (1988: 244), \(^{13}\) viz.

\begin{align*}
(1) \ & \text{PUB/PRV/SUA} + (T#) + \text{demonstrative pro/SUBJPRO} \\
(2) \ & \text{PUB/PRV/SUA} + \text{PRO/N} + \text{AUX/V} \\
(3) \ & \text{PUB/PRV/SUA} + \text{ADJ/ADV/DET/POSSPRO} + (\text{ADJ}) + \text{N} + \text{AUX/V}
\end{align*}

would not have retrieved tokens such as
I think what I’m really seeking all the time is the source of Original Sin in myself. (LOB A19 153)

… I think the Santa Rita or Cabernet would match this food. (LOB E19 112)

I think the way it is going on is very worrying, but nothing more. (LOB A29 122)

These problems and pitfalls, however, should not deter us from using tools of automatic analysis. Rather, they should encourage us to be fully aware of their limitations and to associate them with manual methods when necessary (cf. Ball 1994: 295).

4. The quest: a case study of causative structures

4.1. The object of the quest

Although periphrastic causative verbs (i.e. causatives used with a non-finite clause) have been the subject of numerous studies within a wide range of theoretical frameworks, they have rarely been dealt with from a corpus linguistic perspective. This lack of investigation may be partly due to the difficulty of retrieving such verbs automatically. In fact, the search does not consist in looking for a particular lexical item, but involves finding those occurrences (and, if possible, only those occurrences) where a particular lexical item is used in a particular type of construction. In the case of causative make, I want to exclude instances such as She made no comment or It doesn’t make sense, while retrieving sentences like He made us believe we were late or I was made to stand in front of the class.
As a syntactic structure, there are theoretically three ways to retrieve a causative sentence with *make*. First, one can carry out a lexical search for the lemma *make* on a raw corpus. However, we must take into consideration that the causative use represents only 7.7% of all the occurrences of *make* (ICE-GB) and as a consequence this solution must be discarded since it would require far too much manual post-editing. Second, it is possible to use a tagged corpus and create an algorithm leaving the space between *make* and the non-finite verb unspecified. Such an algorithm is likely to produce quite a lot of noise, since it allows for anything in-between *make* and the non-finite verb. The automatic search will therefore necessarily be followed by a stage of manual post-editing. A third method of retrieval involves the use of a parsed corpus. In this case, the query must contain a form of the lemma *make*, a noun phrase and a non-finite verb. Specifying the nature of the intermediate element (NP) should considerably reduce the amount of manual weeding out and so it appears as the best option. This, however, is not the whole story for, as shown in the preceding section, adequate parsed corpora are not always available. Thus, although ICE-GB is a well-balanced parsed corpus, it contains ‘only’ one million words and does not provide enough instances of periphrastic causative structures for a comprehensive analysis of the phenomenon – 150 instances of causative *make*, 101 of *get*, 77 of *have* and 40 of *cause*. Consequently, it is often necessary to turn to a tagged corpus. Before showing how it can be done, let us first see how causative sentences with *make* are structured.

Causative *make* can be followed by three types of verbal complements, namely (i) a bare infinitive, (ii) a to-infinitive (when *make* is used in the passive) and (iii) a past participle (a marginal and collocationally restricted construction), cf.
(i) *It makes you feel* good when you pull off work like that. <ICE-GB:W1B-010 # 62:2>

(ii) *Esther is made to pay* for the fact that she was illegitimate. <ICE-GB:W1A-010 # 29:1>

(iii) *John-Bishop made his presence felt* with well-taken goals in the 21st and 31st minutes. <BNC:CF9:23967>

In their most basic forms, causative sentences with *make* all share the same kind of structure, viz. a subject, a form of the lemma *make*, a central element, which can be an NP - (i) and (iii) - or the *to*-particle (ii), and a non-finite clause (see Table 2).

<table>
<thead>
<tr>
<th>Subject</th>
<th>MAKE</th>
<th>NP/to</th>
<th>Non-finite clause</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) It</td>
<td>makes</td>
<td>you</td>
<td>feel good</td>
</tr>
<tr>
<td>(ii) Esther</td>
<td>is made</td>
<td>to</td>
<td>pay</td>
</tr>
<tr>
<td>(iii) John-Bishop</td>
<td>made</td>
<td>his presence</td>
<td>felt</td>
</tr>
</tbody>
</table>

Table 2: Basic structure of causative sentences with *make*

However, three caveats should be kept in mind. First, not all sentences with a structure as in Table 2 are causative,\(^{15}\) cf. Table 3. Secondly, some causative structures deviate from the canonical form illustrated in Table 2 and present a different structure, as appears from Table 4.\(^{16}\) An adverb can be inserted before or after the central element (1 and 2), the past participle can directly follow *make* due to the presence of a relative clause (3) or the passivization of the main clause (4), the object (central NP) can be extraposed (5) or there may be no object at all (6) (only with idiomatic expressions, cf. *make do* and *make believe*), and the non-finite clause may be left unexpressed (7). Finally, it is sometimes impossible to distinguish between a causative and non-causative reading, e.g.

*These ‘tunnels’ are made to connect with sites of current or popular importance*  
<BNC:CBB:8507>
May the promise he made us prove reliable.

He helped the student who had made a mistake find the correct answer.

Most people who complain when the Church makes political pronouncements imagine that religion is something to be kept within the four walls of a church. <LOB D16: 159-161>

The motivations which induce a community of speakers to make the change derive from those speakers. <LOB J32: 27-28>

Little effort has been made to find out why. <BNC:H0S:1309>

I am surprised by the little effort they have made to solve the problem.

The scholar made interesting comments based on his research.

She could not make a decision upset as she was by the news.

Table 3: Non-causative sentences with the basic structure of causative sentences

<table>
<thead>
<tr>
<th>Subject</th>
<th>MAKE</th>
<th>NP/to</th>
<th>Non-finite clause</th>
</tr>
</thead>
<tbody>
<tr>
<td>May the promise he made us prove reliable.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>He helped the student who had made a mistake find the correct answer.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Most people who complain when the Church makes political pronouncements imagine that religion is something to be kept within the four walls of a church. | | | <LOB D16: 159-161>
| The motivations which induce a community of speakers to make the change derive from those speakers. | | | <LOB J32: 27-28>
| Little effort has been made to find out why. | | | <BNC:H0S:1309>
| I am surprised by the little effort they have made to solve the problem. | | | |
| The scholar made interesting comments based on his research. | | | |
| She could not make a decision upset as she was by the news. | | | |

Table 4: Causative structures with a non-canonical structure

(1) potholes which made even a Jaguar leap <ICE-GB: W2F – 015 # 33: 1>
(2) a very funny story which made Carlyle absolutely laugh <LOB F07: 191-192>
(3) Er but seriously that phrase doesn't just mean fit for the purpose in the general sense, it also means fit for the purpose in any particular sense you had made known to the shop. <BNC:FUT:4256>
(4) The locations of all fire alarm contact points within the building should be made known to every member of the staff who should be instructed by practical demonstration how to activate the alarm system. <BNC:G0K:15012>
(5) In the event of plans not being funded, the nurse adviser should estimate the effects on patient care and make known her views. <BNC:EVY:17980>
(6) Vocal opposition, such as it is, has come from people who are retired from public life, who have been purged or pushed to one side by the Ceausescu leadership, or who have been forced to make do with a moral posture on key issues, registering their dissent, but no more. <BNC:AAK:9093>
(7) And she said nothing on God’s earth would make me <ICE-GB:S1A-010 # 264:1:B>
This example can be interpreted as being causative (cf. ‘[Someone] made the tunnels connect’) or it may be regarded non-causative and paraphrased as ‘Those tunnels are built in order to connect’. Also, it is not always clear whether or not the complement following make + NP is verbal, cf.

\[
\text{It is necessary, to avoid confusing the issue, to ignore some of the extreme examples of deleterious sounds, those that make telephone operators faint or the jingling of a bunch of keys that sends a mouse into something approaching hysteries.} <\text{LOB F40: 108-111}>\
\]

where faint can be the adjective meaning ‘weak’ or the verb referring to a loss of consciousness.\(^{17}\)

Brought together, all these facts highlight the difficulty of retrieving causative structures and (if possible) only these. In the next section I will show how an automatic search can nonetheless be carried out on the basis of a tagged corpus.

4.2. Conquering the precious vessel

This study is based on a part of the LOB corpus (Lancaster-Oslo/Bergen, written texts dating from 1961), namely the text category ‘Press: reportage’ (LOB[A], approximately 88,000 words), and makes use of X\textsc{kwic} (see Figure 4), a concordance program developed at the University of Stuttgart and part of the IMS Corpus Workbench.\(^{18}\) This tool can be used with raw corpora, but also supports pos-tagged corpora, provided that the corpus is set in the appropriate format (using a conversion program), i.e. one word per line and with each level of annotation separated by a TAB character, cf.
The book seems interesting.

| ART (def) | N (sing) | VB (lex, cop, pres) | ADJ (ge, pos) |

Figure 4: Retrieving causative *make* with XKWIC

The query language of XKWIC is quite complex and needs some getting used to, but it allows for highly refined and specialised searches, in the shape of algorithms describing a succession of words and/or part-of-speech tags. Thus, a possible algorithm\(^\text{19}\) to retrieve causative structures with *make* is:

```plaintext
[word="mak.*|made" & pos="V.*" & genre="LOB[A]"]\{0,4\}
[pos="VB|VBN|BE|BEN|DO|HV|HVN""] \] \* within s;
```
This magic formula asks the program to retrieve from the LOB[A] corpus all the sentences containing a verbal form (V.*, i.e. any pos-tag starting with V, the letter used to tag verbs) of the lemma *make – make/makes/making* (cf. use of the wildcard) or ( | ) *made –* followed by 0 to 4 ({0,4}) unspecified words ( [] ), and then the base form (“VB”) or past participle (“VBN”) of a lexical verb, or one of the forms *be, been, do, have or had* within a sentence (‘within s’). A query should end with a semicolon (;).

By not specifying the nature of the central element(s), the algorithm opens the door to all sorts of possibilities, including the insertion of an adverb (see Table 4, (1) and (2)) and the presence of a relative clause or apposition after the object, as in:

*Such emotive language as this makes us as the reader feel pity for the speaker, it is as though he is trapped on this earth.* <ICE-GB:W1A-018 #52:1>

Moreover, it makes it possible to retrieve passive uses of causative *make* simultaneously since nothing prevents the to-particle from occupying the central position. Finally, the null interval (0) allows for cases where there is no object between *make* and the non-finite complement, a phenomenon due to main clause passivization (see Table 4, (4)), pre- or postposition of the object ((3) and (5)) or idiomatic expression (6).

The first stage of the study consisted in looking for all the occurrences of any form of the verb *make* in the corpus and then manually discarding the non-causative concordance lines. This gave the exact number of causative constructions with *make* in the corpus. The second stage made use of the algorithm given earlier, which provided a number of matches. Finally, the results of the two methods of retrieval (manual and automatic) made it possible to determine the precision and recall rates of the automatic query, as shown in Table 5. The algorithm retrieved 13 causative structures out of the 16 present in the corpus, that is, a reasonable recall rate of 81.25%. Moreover, since the
structures not retrieved all contained an object of five or more words and/or signs (see Table 6), this rate could even be improved by extending the span between *make* and the non-finite complement – although this, obviously, would result in a decrease of the precision rate.

<table>
<thead>
<tr>
<th>Tool</th>
<th>XKWIC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Corpus</strong></td>
<td>LOB, Press: reportage (about 88,000 words) pos-tagged corpus</td>
</tr>
<tr>
<td><strong>Number of causative structures</strong></td>
<td>(manually retrieved)</td>
</tr>
<tr>
<td><strong>Number of matches of the automatic search</strong></td>
<td>- causative structures</td>
</tr>
<tr>
<td></td>
<td>- non-causative structures</td>
</tr>
<tr>
<td><strong>Number of causative structures not retrieved automatically</strong></td>
<td>3</td>
</tr>
<tr>
<td><strong>Precision rate</strong></td>
<td>13/30 = 43.3%</td>
</tr>
<tr>
<td><strong>Recall rate</strong></td>
<td>13/16 = 81.25%</td>
</tr>
</tbody>
</table>

Table 5: Results of the pilot study on *make* with XKWIC

But Sir Roy pointed out that a few months ago Mr. Kaunda said that if 2UNIP did not get its way what would happen would *make* the Mau Mau in Kenya *seem* like a child's picnic." <LOB A02: 171-173>

Conroy is out for the season and the selectors have a problem on their hands in shaping the England attack, which will *make* the more senior members, such as Mr. Harry Lewis and Mr. H. L. Holliwell, think back uneasily to the 1956-57 season. <LOB A08: 148-152>

And Mr. Simpson's lunatic logic has a freshness, a lightness about it that would *make* "Waiting in the Wings* seem* bad even if it weren't. <LOB A19: 90-92>

Table 6: Causative structures not retrieved by XKWIC

As for the precision rate, it is far from being satisfactory, as it amounts to a poor 43.3% (Table 5). As many as 17 concordance lines do not actually instantiate the phenomenon
under investigation (see Table 7). Among these, some have a to-purpose clause (a), in others *make* happens to be followed by a clause containing an infinitive, a past participle or a form such as *do* \(^{23}\) (b) and two have coordinated verbs (c). Of course, an algorithm à la Biber, specifying the internal composition of the central element, could exclude some of these non-instances, but it would also lead to a drop in the recall rate since, as pointed out by Ball (1994: 296), ‘NP structures cannot be represented by a finite set of patterns of this type’.

By comparison, a search on a good parsed corpus would retrieve all the causative structures in Table 6, since they consist of a form of *make* followed by an NP (however long it may be) and an infinitive. Furthermore, it would automatically discard some of the irrelevant materials retrieved by XKwic (Table 7), viz. sentences of type (b) and (c). As for to-purpose clauses, half of them could be excluded ((1), (2), (6), (7)), since the query for active *make* would normally not allow to to precede the infinitive.\(^{24}\) This would result in an improved precision rate. The exact rate would very much depend on the kind of information encoded in the parsed corpus. ICE-GB, for instance, has been encoded with a special feature which makes it possible to distinguish between the causative uses of *make* and the other uses. When used causatively, *make* is said to be a transitive verb, i.e. a verb followed by an NP and a non-finite clause, where the NP can be described as the object of the main verb or the subject of the non-finite clause (see Fang 1996: 145-146). This is also the case when the NP occupies an unusual position or is not expressed.\(^{25}\) In sentences such as those in Table 7, on the other hand, *make* is labelled as a monotransitive verb, one which is complemented by a direct object only. Thanks to such information, ICECUP,\(^{26}\) the ICE Corpus Utility Program specifically designed to process and query the International Corpus of English, rates exceptionally
(a) to-purpose clause

(1) "Why don't you make proposals to legislate in the autumn?" <LOB A06: 20>
(2) "This court is very heavily guarded and King is prepared to give an undertaking that he will make no attempt to escape", he said. <LOB A11: 78-80>
(3) B.E.A. will probably try to resist the strong efforts that will be made in Paris to raise fares, but it may well be obliged to concede something. <LOB A15: 79-81>
(4) In Kuwait plans were being made to evacuate the 3,000 or so Britons who live there. <LOB A21: 64-65>
(5) Steps are in hand to repay the +119,000 of Preference capital and interest in the company's report centres chiefly on what further moves will be made to distribute some of the surplus cash resources. <LOB A25: 186-188>
(6) "But we mean to make a real effort to get the Russians moving again in these negotiations." <LOB A29: 196-198>
(7) The U.S. delegation, led by Mr. Arthur Dean, are under instructions from President Kennedy to make the maximum effort to reach agreement with Russia. <LOB A29: 216-218>
(8) Marlow Urban Council has given the visit every support and appeals have been made for residents to entertain the players. <LOB A42: 124-126>

(b) infinitive/past participle/do- clause

(9) There General de Gaulle had made clear that he would accept Britain into the Common Market only if there were no conditions laid down to meet the Commonwealth and other reservations. <LOB A04: 12-14>
(10) The good beginning made at Vienna must be followed up by new efforts for peace, the Soviet Communist Party newspaper Pravda declared yesterday. <LOB A04: 72-74>
(11) And with all these African politicians making trouble it might blow up into another Congo any day. <LOB A09: 67-69>
(12) The point is often made that Americans have never known modern war on their soil. <LOB A26: 146-147>
(13) The two leaders will discuss a wide range of world problems, although both have made clear there will be no negotiations. <LOB A31: 77-78>
(14) New plans are being made - and they do not include a replacement for Reg Smith, the manager they sacked three weeks ago. <LOB A33: 126-128>
(15) Interviewed, Dixon made a statement which was put in as evidence and the Constable alleged that Cole said that he had a clear conscience. <LOB A43 : 44-46>

(c) coordination

(16) Ind Coope is spending millions to make and market Skol. <LOB A16: 145-146>
(17) I think it is time that the case for the British theatre of today was made, and made loud and clear. <LOB A19: 63-64>

Table 7: Irrelevant materials retrieved by XKWIC
well, with perfect precision of 100% and excellent recall\textsuperscript{27} of 93.75% in the subcorpus of ‘Non-Academic Writing’ (86,643 words). It should be emphasised, however, that ICE-GB has been manually corrected (see Wallis 2002) and that, in the present state of affairs, a corpus parsed with no manual intervention could not possibly reach such a high degree of accuracy and detail.

4.5. Causative structures: an open-ended story

As appears from this study, periphrastic causative structures are difficult to retrieve automatically from a corpus. A search on a parsed corpus yields good precision and recall rates – and even excellent rates in a fine-grained and manually corrected corpus such as ICE-GB – but the state of the art is such that most parsed corpora available today are not sufficiently reliable and/or too small to allow for a comprehensive study of a relatively infrequent phenomenon like causative structures. A search on a tagged corpus, on the other hand, yields a reasonable recall rate, but a precision rate that leaves a lot to be desired. Yet, there is a good case for persisting in using tagged corpora for the automatic retrieval of syntactic structures. First, as demonstrated by Ball (1994: 295-296), poor precision is a lesser evil than poor recall. While irrelevant materials can easily be discarded by hand, ‘it is generally impossible for the analyst to know what has been missed without analysing the entire corpus by hand’ (Ball 1994: 295). Second, even though the recall rate itself might not be perfect for a totally reliable quantitative analysis, it is more than enough for a qualitative analysis, as all the instances retrieved are authentic structures, whose careful investigation is bound to bring out interesting tendencies, as well as counterexamples to some of the claims made in the literature. To
take but one example with respect to causative structures, whereas some grammars consider that the subject of causative *make* can only be an agent (cf. Givón 1993: 9), I found (see Gilquin 1999) that it was predominantly inanimate in the two corpora I used (LOB and BNC Sampler), e.g. *The humiliation made me shudder* <BNC:FU7:6584>. Finally, to date only tagged corpora are sufficiently big to provide large numbers of instances of a particularly rare syntactic structure. Naturally, such a way of working is riddled with traps and so it seems as if the easy way out would be to avoid the study of such phenomena. However, the easy way out is not always the most rewarding one. And anyway, who said corpus linguistics should be easy? Sometimes, one must ‘be willing to march into hell for a heavenly cause’!

5. **Automatic retrieval of syntactic structures: the impossible dream?**

This article has shown that, in the present circumstances, the fully automatic retrieval of syntactic structures with no manual intervention is still something of an impossible dream for want of suitable and/or reliable tools and corpora. This point was highlighted through the case study of causative structures with *make*. Although these structures could quite easily be retrieved from a parsed corpus with very good precision and recall rates, the lack of the ideal parsed corpus (i.e. accurate, detailed and big enough) forces one to turn to a tagged corpus and use a method requiring more manual post-editing and yielding slightly less satisfactory results.

While this paper has emphasised the fact that the numerous obstacles involved in the retrieval of syntactic structures tend to deter scholars from embarking on such studies, it is also meant to be a plea for more research on complex grammatical
phenomena. Linguists with insufficient knowledge in programming to create their own software are encouraged to consider alternative, next best methods, using the tools and data that are available to them, and to ‘run where the brave dare not go’. It has been demonstrated that, not only are there ways to automate the search to a certain extent, but the data collected, even if not perfect from a quantitative point of view, can always form the basis of an accurate and in-depth qualitative analysis. Moreover, there is no limit to what we can hope for, for there is no doubt that, as advances continue to be made in the field of corpus-processing software, our quest will become less of a dream and more of a reality. But that is another story.

Acknowledgements

I acknowledge the support of the Belgian National Fund for Scientific Research, which has offered me a position as a Research Fellow. Also, I sincerely thank Sylviane Granger and two anonymous reviewers for their insightful comments, as well as Nora Condon and Sheila Mugridge for their precious help. Finally, a big thank you to Noëlle Serpollet, who helped me capture, if not the Holy Grail, at least the screen of XKWIC.

Notes

1. It is obvious that relying on available tools closes the door on some research questions, since researchers will tend to investigate what they can easily retrieve, and not what is most interesting or motivated linguistically. Conversely, someone with programming skills will be able to develop their own program or adapt an existing one in order to investigate their particular
research question. Although an introduction book such as Mason’s (2000) can undoubtedly help in acquiring such skills, the long-term goal, though ambitious, would be to introduce training in programming into the curriculum of all future linguists.

2. This, actually, is the case for some of the modals, cf. nominal use of *can*.

3. ICE-GB is the British component of the International Corpus of English (ICE), a project initiated in 1988 by Sidney Greenbaum, University College London, and now coordinated by Gerald Nelson. The idea behind this project is to provide material for comparing different national varieties of spoken and written English. Some twenty countries, with English as a majority first language or an official additional language, are involved in this project, the aim being that each of them produces a 1,000,000-word corpus representative of their national variety of English. All corpora will be designed along parallel lines, so that comparisons will be made possible (see Greenbaum 1996).

4. This is not always the case. Thus, while the SUSANNE annotation identifies the functional roles of clause constituents (cf. Sampson 1995), the ALICE parsing scheme does not (cf. Black and Neal 1996). See Atwell et al. (2000) for an evaluation of several parsers in terms of layers of syntactic annotation.

5. It should be noted that such an algorithm could not retrieve instances of the zero-relative pronoun, since null elements are normally not encoded in tagged corpora. By contrast, a search for relative *clauses* on a parsed corpus would also retrieve relative clauses with a null relative pronoun.

6. Excluded from the matches, however, would be a sentence such as *The man you saw leave the building is my father*, where the infinitive immediately follows the verb *see*.

7. This seems to be a recent development. Thus, in 1997, Granger noted corpus linguists’ reluctance to use pos taggers (p. 365).

8. Some limitations may nonetheless be imposed by the level of granularity of the tagset. Not all tagsets make a distinction between, say, infinitive and base form (CLAWS5 does, but CLAWS1 does not). By the same token, Serpollet (2001) could not rely on tagging to retrieve subjunctives, since these are not pos-tagged. Therefore, her object of research can be classed as a syntactic structure, which she retrieved thanks to a combination of triggering expressions (e.g. *insist, suggest, eager*) and clauses introduced by *that*. 

10. See http://www.lingsoft.fi/cgi-bin/engcg for a demonstration.

11. Biber’s way of working has actually been applied with excellent results in the Longman Grammar of Spoken and Written English (Biber et al. 1999).

12. T# = tone unit boundary, ALL-P = all punctuation, VBG = -ing form of verb, PREP = preposition, DET = determiner, WHP = WH pronoun, WHO = other WH word, PRO = pronoun, ADV = adverb.

13. PUB = ‘public’ verb, PRV = ‘private’ verb, SUA = ‘suasive’ verb (see examples of public, private and suasive verbs in Biber 1988: 242), T# = tone unit boundary, SUBJPRO = I, we, he, she, they (plus contracted forms), PRO = pronoun, N = noun, AUX = auxiliary, V = any verb, ADJ = adjective, ADV = adverb, DET = determiner, POSSPRO = my, our, your, his, their, its (plus contracted forms).

14. The proportions for the other causatives in ICE-GB are: 16.4% (cause), 2.8% (get) and 0.6% (have).

15. The problem is even worse with causative have and, to a lesser extent, causative get. Thus, a minimal change can make the causative interpretation unavailable, cf.

(1) a. I had my watch repaired.
   b. I had my watch stolen.

(2) a. Sherey had George water her plants.
   b. Sherey had George overwater her plants. (Ritter and Rosen 1993: 526)

where the (b) sentences are experiential constructions, i.e. constructions where ‘the subject experiences something, is in some way affected by something’ (Van Roey 1982: 81). Similar structures can also have an existential meaning (e.g. And you had a scientist up there talking about pilgrimages <ICE-GB: S1A-096 # 201:1:A>), a lexical meaning (e.g. Mr Gorbachev has very few cards left to play. <ICE-GB:W2C-008 # 24:1>), express permission (e.g. All that the opposition would have us do was to hand out more and more fish. <ICE-GB:W2B-012 # 120:1>) or obligation (If they don’t support the club now they will only have themselves to blame in the future. <ICE-GB:W2C-004 # 75:3>). A study carried out on ICE-GB (Gilquin 2000) showed that, out of the 181 instances of have + NP + non-finite clause used ‘transitively’ (i.e. where the
NP can be described both as the object of *have* and the subject of the non-finite verb), only 77 (42.5%) were actually causative. For *get*, the ratio was 101/142 (71.1%).


17. This is not to say that constructions with adjectives should not be considered causative. Altenberg and Granger (1998) rightly point out that causative *make* involves three types of structures, namely adjective structures (e.g. *make something possible*), verb structures (e.g. *make someone realise something*) or noun structures (e.g. *make somebody a star*). Only constructions of the second type (verb structures) are taken into account here.

18. See http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/.

19. Using the CLAWS1 tagset (see http://www.comp.lancs.ac.uk/computing/research/ucrel/claws1tags.html). It should be kept in mind that the query syntax has to be adapted to the tagging system of the corpus used. Thus, while a past participle is tagged as VBN in the LOB corpus (CLAWS1 tagset), it is assigned the tag VPN in the BNC Sampler (CLAWS6).

20. The form *done* is tagged as VBN, i.e. past participle of a lexical verb.

21. In the case of *make*, taking such structures into account has little impact on the precision rate. However, the situation is different when dealing with causative *have*. If one wants to retrieve a sentence such as *This is the tunnel he had built [by his slaves]*, it will be necessary to examine all the perfective uses of *have*, resulting in a dramatic drop in the precision rate. Considering that such structures represent only 3 instances out of 77 causative constructions in ICE-GB (3.9%), one might wonder whether the ‘game’ is worth the candle. Similarly, retrieving *get*-sentences with a pre- or postposed object (e.g. *I got repaired the watch that my father had given me on his deathbed*) would involve allowing for all the instances where *get* is used as a passive auxiliary, as in *He got killed during the war*. Such causative constructions, however, seem to be even less frequent than with *have* (not a single example in ICE-GB). In fact, even a query with a span of one to four words will retrieve some instances of perfective *have* and passive auxiliary *get*, notably when an adverb occurs between the verb and the past participle, cf. *I have never seen him* or *He got immediately eliminated*. However, these patterns cannot possibly be overlooked, for many causative constructions actually present a single word between the causative and the non-finite verb, as in *I had it fixed* (38% for *have* and 58% for *get* in ICE-GB).
22. The precision and recall rates are calculated by means of the following formulae:

\[
\text{Precision rate of causative constructions} = \frac{\text{No. of automatically retrieved causative constructions}}{\text{No. of matches of the automatic search}}
\]

\[
\text{Recall rate of causative constructions} = \frac{\text{No. of automatically retrieved causative constructions}}{\text{No. of manually retrieved causative constructions}}
\]

23. CLAWS1 does not make any distinction between *do* used as a base form and *do* used as an infinitive.

24. The question to ask, however, is whether this is a good thing or not. Ball (1994: 296) observes that ‘with perfect precision, we find exactly what we said we were looking for, and no more’. We do not expect causative *make* to be used with a *to*-infinitive in the active. Yet, nothing proves that this kind of structure never occurs in real data.

25. Unfortunately, this feature would not discard the non-causative uses of *have* and *get* alluded to in note 15 (experiential, existential, etc.) for, in all these meanings, *have* and *get* are used ‘transitively’ (in the sense defined above).

26. ICECUP 3.0, together with a 20,000-word sample of ICE-GB, can be downloaded for free from http://www.ucl.ac.uk/english-usage/ice-gb/sampler/download.htm. The full version of the corpus (1,000,000 words) is available on CD-ROM.

27. The only causative structure not retrieved by ICECUP is the following idiomatic expression:

\[
\text{It was therefore preferable, they argued, to make do with an inherited monarch, with an even chance of his being a decent ruler, and to concentrate not on how he achieves power but on how to influence him for the best. <ICE-GB:W2B-014 #59:1>}
\]

which, admittedly, not everybody would consider causative. Yet, for those who want to include such constructions, a lexical search on *make do* and *make believe* should do.
References


Serpollet, N. 2001. “The mandative subjunctive in British English seems to be alive and kicking… Is this due to the influence of American English?”. In Rayson et al., 531-542.


Trotta, J. and M. Johansson. 2001. “How big of a corpus is a big enough corpus?”. In De Cock et al., 105-106.


XKWIC, IMS, Stuttgart ([http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/](http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/)).