Annotating progressive aspect constructions in the spoken section of the British National Corpus

Andrew Caines[1], Paula Buttery[2]

[1] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, U.K.
[2] University of Cambridge, 9 West Road, Cambridge CB3 9DP, U.K.
acaines@ebi.ac.uk, paula.buttery@cl.cam.ac.uk

Abstract

We present a set of stand-off annotations for the ninety thousand sentences in the spoken section of the British National Corpus (BNC) which feature a progressive aspect verb group. These annotations may be matched to the original BNC text using the supplied document and sentence identifiers. The annotated features mostly relate to linguistic form: subject type, subject person and number, form of auxiliary verb, and clause type, tense and polarity. In addition, the sentences are classified for register, the formality of recording context: three levels of 'spontaneity' with genres such as sermons and scripted speech at the most formal level and casual conversation at the least formal. The resource has been designed so that it may easily be augmented with further stand-off annotations. Expert linguistic annotations of spoken data, such as these, are valuable for improving the performance of natural language processing tools in the spoken language domain and assist linguistic research in general.

Keywords: BNC annotations, progressive aspect, spoken language

1. Introduction

We present a new resource: a set of annotations to accompany all sentences that feature a progressive aspect verb group in the spoken section of the British National Corpus (BNC; Burnard, 2000). This incorporates some ninety thousand sentences of the one million[1] contained in the spoken section of the BNC (sBNC) and includes information on register, properties of the subject, properties of the auxiliary and properties of the clause.

The dataset does not contain any original text from sBNC; that would be a breach of the licence. Instead the annotations may be matched to the original text by the unique document identifier and sentence number. The annotations will only be of any practical benefit to those in possession of, or with legitimate access to, a licensed copy of the BNC therefore. The annotation set is freely available online at http://www.wordiose.co.uk/resources.

The information provided by these annotations is of use to both language researchers (Caines, 2010) and natural language engineers. For instance, the statistical information collated over these annotations has been employed to improve probabilistic parsing of spoken language (Caines & Buttery, 2010). Given that parsers are trained on written language data, they generally perform less well on spoken language data. A set of manually annotated sentences such as these may be used as training data so that parsers perform better in the spoken domain - for sentences containing progressive aspect verb groups, in this instance, at least.

2. Motivation

The motivation for carrying out these annotations relates to the first author's postgraduate research supervised by the second author (Caines, 2010) on the topic of 'zero auxiliary' constructions, a non-standard feature whose occurrence we wished to investigate in progressive aspect constructions such as those exemplified below:

(1) How you feeling now? KBK 3474[2]
(2) You not having any cake? KBW 13888
(3) What you been buying? KPV 5313

We intended to describe the conditions which most often co-occur with a zero auxiliary construction in the progressive aspect, and set out to annotate each and every sentence containing at least one progressive aspect verb group in sBNC. The outcome was more than ninety thousand annotated sentences.

The reason for choosing the BNC for our investigation was that it is a large and accessible resource, rigorously designed and prepared, with data of suitable time period for our purposes. The BNC is a collection of spoken and written documents which was designed to be a snapshot of British English at the end of the twentieth century (Burnard, 2000). The corpus contains 100 million words, including a 10 million word, one million sentence spoken language section. We focused our attention on the spoken subcorpus after a pilot comparison study of the written and spoken sections of the BNC confirmed that the zero auxiliary predominantly occurs in speech.

The spoken section was collated from various sources covering a wide range of formality levels, from sermons and news broadcasts to casual conversation recorded by volunteers in their homes. Even though it was the most costly and labour-intensive material to collect, the conversation subsection makes up four-tenths of sBNC's 10 million words.

[1] To be precise: 1,035,527 sentences, based on assumptions and calculations by Benjamin Van Durme (personal communication and blog post: http://hlplab.wordpress.com/2010/07/01/extracting-speaker-meta-data-from-the-bnc/#comment-1199

[2] For extracts from the BNC: all rights in the texts cited are reserved. As requested by the distributor (Oxford University Computing Services on behalf of the BNC Consortium) each sentence is followed by an alphanumeric text identifier and sentence number.

Volunteers were recruited to broadly represent speaker demographics such as gender, age, social class and region.

3. Annotation method

Annotations were made for every sentence featuring progressive aspect verb group(s) on the basis of seven linguistic properties. Each of these relates to the sentence in various ways, as illustrated in Figure 1.

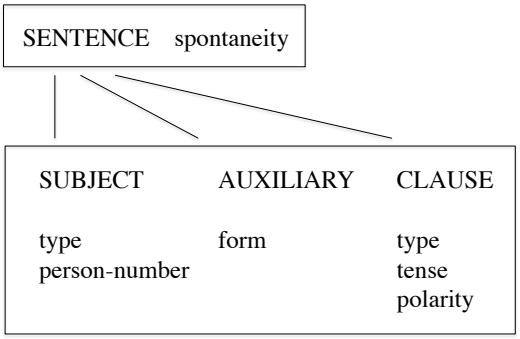

Figure 1: Annotation hierarchy

The diagram shows the seven properties in lower case and four feature types in upper case. These seven properties are explained in further detail below along with their possible values. The numeric annotation given to each value is shown in bold type:

Sentence: spontaneity
Based on the document source [3] and relating to formality of situational context. Besides **0**: unclassified where spontaneity level could not be determined, the three spontaneity levels are **1**: formal and scripted speech; **2**: formal and unscripted speech; **3**: informal speech. Examples of scripted speech include sermons, political speeches and news broadcasts; formal unscripted speech includes business meetings, academic lectures, radio discussions. Informal speech represents the 4 million word conversation section - recordings made by volunteers over several days as they went about their daily lives. The full list of genres are grouped by spontaneity level in Table 1.

Subject: type
Three values for this property - **0**: subject not supplied, or, the occurrence of a 'zero subject'. For example, 'Just been talking to Tracey' KBF 13231. **1**: pronominal subject such as 'he', 'she', 'it'. **2**: other subject types such as noun phrase or clause.

Subject: person-number
1: first person singular ('I'); **2**: second person singular-plural ('you'); **3**: third person singular ('she', 'Norman', 'the penguin'); **4**: first person plural ('we', 'Gio and I'); **6**: third person plural ('they', 'the boys').

[3]Document information gathered from Davies (2006).

Genre	Spontaneity
broadcast documentary	1
broadcast news	1
sermon	1
speech (scripted)	1
broadcast discussion	2
broadcast interview	2
business interview	2
business meeting	2
courtroom proceedings	2
higher education lectures and tutorials	2
medical consultation	2
oral history interview	2
parliamentary proceedings	2
public debates and meetings	2
sales demonstration	2
school lesson	2
sports commentary	2
speech (unscripted)	2
conversation	3

Table 1: Grouping the sBNC genres from Davies (2006) into three spontaneity levels.

Auxiliary: form
1: full form auxiliary verb such as 'are'; **2**: contracted auxiliary verb such as ''re'; **3**: auxiliary verb not supplied, or, the occurrence of a 'zero auxiliary'.

Clause: type
Relating purely to syntactic structure rather than semantic notions of question and statement - **1**: declarative clause; **2**: interrogative clause.

Clause: tense
1: present tense (e.g. 'I am sleeping'), **2**: past tense (e.g. 'I was sleeping'), **3**: present perfect tense (e.g. 'I have been sleeping'), **4**: past perfect tense (e.g. 'I had been sleeping').

Clause: polarity
0: a negated clause, **1**: a positive clause.

Annotations were made manually by a single annotator, the first author. It was deemed unnecessary to cross-validate the annotations since, firstly, the annotated features are on the whole non-subjective in nature. Secondly, where ambiguity exists in any way, classification specifications are clearly defined.
For instance, subject type is unambiguous in that it is either there or it is not, and if it is there it is either a pronoun or not. Mixed, coordinated noun-pronoun subjects such as 'Gio and I' were classified as nominal (other noun).
Person-number is a fairly self-evident feature. The only uncertainty is second person number since the form is *you* for both singular and plural. Thus it is collapsed into the lone category 'second singular-plural'.
In transcription, auxiliary form is indisputable; whether the recordings were transcribed accurately is a general issue for all spoken corpora, not just this one. In the case of the BNC,

preparation and design rigorous; annotators were given consistent training and instructions to represent rather than correct the recordings they heard (Crowdy, 1993; 1994).

The same unambiguity is true of clause type since declarative and interrogative status are derived from word-order alone (rather than statement or question which are pragmatic properties relating to a layer of meaning on top of the form). Tense and polarity are both strictly form based also.

The one area of potential subjectivity is spontaneity. The genres were defined by Davies (2006) based on the setting in which the recording was made. The area of subjectivity comes in then grouping these genres into three spontaneity levels. See Table 1 for how the groups are constituted. By specifying the groups in advance of making the annotations, this part of the task was clearly defined and the possibility for subjectivity removed.

Rather than subjectivity or selectional inconsistency, then, the remaining concern regarding accuracy is human error. There are several points along the pipeline at which error may have accidentally crept in before the annotations even began. For example, were the recordings transcribed entirely accurately to begin with? This is a question we cannot answer, which does not mean we should not be aware of it, but nevertheless we must accept transcription accuracy with good faith in the case of sBNC, as with any spoken corpus.

Annotation accuracy was tested by sampling a random set of 10% of the sentences and re-annotating these, many months after the original work was done. The sample set was found to be more than 99% accurate. The annotations were carried out according to clear guidelines, over a period of several months, working through no more than a thousand sentences in any one session. In the event that readers find any inaccuracies in the annotations, the authors encourage feedback through the online contact form at http://www.wordiose.co.uk/resources.

A summary of counts per feature is set out in Table 2. There is a full analysis of how these properties cross-tabluate in Caines (2010) but a brief analysis is presented for interest's sake in Table 3. The zero auxiliary most often occurs in (a) zero subject progressives and (b) second person progressive interrogatives. The latter construction type occurs in zero auxiliary form 27% of the time and was illustrated in (1), (2) and (3) above. The former construction type, the zero subject as exemplified in (4) and (5), occurs with a zero auxiliary 82% of the time. But since it involves omission not just from the verb group but of a sentence constituent which is supposedly obligatory in English it is certainly a special case.

(4) Just trying to find a place that's a bit more comfortable for you. KE3 323
(5) Yeah, keeping myself busy. KD8 8859

Even though these annotations were prepared for a particular purpose - research into the conditions most favouring zero auxiliaries - it is apparent that they are of potential use to other researchers, and therefore have been made freely available as detailed in the following sections.

4. Online resource

Underlying the investigation alluded to in Table 3 is a new corpus resource. An annotated subset of the spoken section of the BNC analysing features of progressive aspect verb groups.

The annotations are freely available online at http://www.wordiose.co.uk/resources. As well as the main dataset in XML format, there is a Document Type Definition (DTD) and README which fully document the XML and the method underpinning the annotations. The design of the DTD readily allows the addition of further annotations, as shown in Appendix A, whether relating to these same ninety thousand sentences or others in sBNC and the BNC more generally. We ask that others use the online contact form to inform us of the existence of new annotations.

These are stand-off annotations, meaning that the original text is not supplied. Users can match the annotations to the text via document and sentence identifiers. In Appendix B there are three examples of how the annotations are set out in XML format, relating to the following three sentences from sBNC.

(7) What you not talking for Wendy? KR0 185
(8) Was happily pottering about there. KBW 13768
(9) The Council wasn't doing enough for young people. D95 430

It can be seen that appropriate parsing of the BNCDOC and S IDs enable the pairing up of our stand-off XML annotations with the original BNC texts.

5. Summary

In this paper we describe a new resource: a publicly available set of annotations to accompany the ninety thousand sentences in sBNC which feature a progressive aspect verb group. The annotations may be accessed freely at http://www.wordiose.co.uk/resources.

Having written the DTD with scope for further annotations, the possibilities for future addition to this resource are entirely open. Any future work which requires new meta-annotation of sBNC or other sections of the BNC will automatically contribute new data that can be merged into the existing annotations.

We also view further annotation of sBNC as a valuable project in itself. Any such information will in some way help alleviate the problems of spoken text processing with natural language processing tools which were discussed previously.

We hope that others perceive the opportunity to collaboratively build a dataset of annotations to accompany the BNC. In this way the existing resource will become richer, benefitting researchers who use the BNC and in turn improving our understanding of the language used therein.

6. Acknowledgements

The first author was supported by a Arts and Humanities Research Council (UK) doctoral scheme award. The second author was supported by Engineering and Physical Sciences

Value	0	1	2	3	4	6
Spontaneity	unclass	scripted	unscripted	informal		
	8262	3725	39,101	42,165		
Subj type	zero	pronoun	other			
	1509	78,580	13,164			
Subj pers-no	zero	first sg	second	third sg	first pl	third pl
	1509	18,232	17,191	30,687	11,664	13,970
Aux form		full	contr	zero		
		38,015	51,295	3943		
Clause type		declara	interrog			
		83,305	9948			
Tense		pres	past	pres-perf	past-perf	
		70,938	18,986	2956	373	
Polarity	negative	positive				
	7729	85,524				

Table 2: Properties of the 93,253 progressive aspect verb groups in sBNC.

	Zero subj	2nd interrog	Other	Overall
Progressives	1509	4923	86,821	93,253
Full aux	18%	71%	39%	41%
Contr aux	0%	2%	59%	55%
Zero aux	82%	27%	2%	4%

Table 3: Auxiliary form by construction type (selected) in sBNC.

Research Council (UK) grant number EP/F030061/1. Both authors thank the anonymous reviewers for their constructive criticism which was of much assistance in preparing this paper for submission. Finally, both authors sincerely acknowledge the support and guidance of Norman Cobley, Alan Horne, Jyothi Katuri, Barbara Lawn-Jones, Michael McCarthy and Peter Stoehr.

7. References

The British National Corpus, version 3 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/

L. Burnard (ed.) 2000. The British National Corpus User's Reference Guide. http://www.natcorp.ox.ac.uk/docs/userManual/

A.P. Caines 2010 'You talking to me?' Zero auxiliary constructions in British English. Ph.D thesis, University of Cambridge.

A.P. Caines and P.J. Buttery 2010. 'You talking to me?' A predictive model for zero auxiliary constructions. In: *Proceedings of the Workshop on NLP and Linguistics: Finding the Common Ground (ACL 2010)*, 4351. Association for Computational Linguistics.

S. Crowdy 1993. Spoken corpus design. *Literary and Linguistic Computing* 8: 259-265.

S. Crowdy 1994. Spoken corpus transcription. *Literary and Linguistic Computing* 9: 25-28.

M. Davies 2006. VIEW: Variation in English words and phrases (Interface for the British National Corpus). http://corpus.byu.edu/bnc

A DTD

(6)

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT CAINESCORPUS (BNCDOC+)>
<!ELEMENT BNCDOC (S+)>
<!ELEMENT S (ANNOTATIONS)>
<!ELEMENT ANNOTATIONS (PROGRESSIVE?, FUTURE_ANNOTATION?)>
<!ELEMENT PROGRESSIVE (SUBJECT, AUXILIARY, CLAUSE)>
<!ELEMENT SUBJECT EMPTY>
<!ELEMENT AUXILIARY EMPTY>
<!ELEMENT CLAUSE EMPTY>
<!ELEMENT FUTURE_ANNOTATION EMPTY>
<!ATTLIST BNCDOC ID CDATA #REQUIRED>
<!ATTLIST S ID CDATA #REQUIRED>
<!ATTLIST S SPONTANEITY CDATA #IMPLIED>
<!ATTLIST ANNOTATION TYPE CDATA #REQUIRED>
<!ATTLIST SUBJECT TYPE CDATA #REQUIRED>
<!ATTLIST SUBJECT PERSNO CDATA #IMPLIED>
<!ATTLIST AUXILIARY FORM CDATA #REQUIRED>
<!ATTLIST CLAUSE TYPE CDATA #REQUIRED>
<!ATTLIST CLAUSE TENSE CDATA #REQUIRED>
<!ATTLIST CLAUSE POLARITY CDATA #REQUIRED>
```

B XML examples

(7)

```
<BNCDOC ID="KR0">
    <S ID="185" SPONTANEITY="3">
      <ANNOTATIONS>
        <PROGRESSIVE>
          <SUBJECT TYPE="1" PERSNO="2" />
          <AUXILIARY FORM="3" />
          <CLAUSE POLARITY="0" TENSE="1" TYPE="2" />
        </PROGRESSIVE>
      </ANNOTATIONS>
    </S>
  </BNCDOC>
```

(8)

```
  <BNCDOC ID="KBW">
    <S ID="13768" SPONTANEITY="3">
      <ANNOTATIONS>
        <PROGRESSIVE>
          <SUBJECT TYPE="0" PERSNO="0" />
          <AUXILIARY FORM="1" />
          <CLAUSE POLARITY="1" TENSE="2" TYPE="1" />
        </PROGRESSIVE>
      </ANNOTATIONS>
    </S>
  </BNCDOC>
```

(9)

```
  <BNCDOC ID="D95">
    <S ID="430" SPONTANEITY="2">
      <ANNOTATIONS>
        <PROGRESSIVE>
          <SUBJECT TYPE="2" PERSNO="3" />
          <AUXILIARY FORM="1" />
          <CLAUSE POLARITY="0" TENSE="2" TYPE="1" />
        </PROGRESSIVE>
      </ANNOTATIONS>
    </S>
  </BNCDOC>
```