

Automated retrieval of passives from native and learner corpora: precision and recall

Sylviane Granger

*Centre for English Corpus Linguistics, 1 Place
Blaise Pascal, B-1348 Louvain-la-Neuve, Belgium*

KEYWORDS: retrieval, passive, learner corpus

AFFILIATION: Universite Catholique de Louvain (UCL)

E-MAIL: granger@etan.ucl.ac.be
FAX NUMBER: +32 10 472579
PHONE NUMBER: +32 10 474947

Annotated corpora provide a potentially much more powerful platform for linguistic analysis than raw corpora, not least because they enable “a concordance program, for example, to search for grammatical abstractions (such as instances of the passive voice, of the progressive aspect, of noun-noun sequences, etc.) rather than words” (G. Leech 1991:19). With taggers becoming widely available, the automatic analysis of grammatical features will soon become the rule rather than the exception. In fact, as suggested by J. Sinclair (1992:381) the increasing size of corpora is making full automatization essential. In this light it would seem necessary to be able to assess the reliability of automated retrieval. Very few studies to date have addressed this issue, however. One notable exception is C. Ball (1994), who warns researchers against the pitfalls of automated text analysis. She claims that text processing tools “must be used with a full awareness of their limitations, and should be coupled with or replaced by manual methods when appropriate”. She also stresses the need for “microscopic studies of individual phenomena” before embarking on large-scale macroscopic ones. My paper is directly in keeping with Ball’s line of thinking. It aims to assess the reliability of automated text analysis by providing a microscopic study of one particular grammatical phenomenon, the passive construction. My investigation of the passive is part of the International Corpus of Learner English (ICLE) project based at the University of Louvain. The aim of the project is to identify the differences in grammar, lexis and discourse which distinguish advanced learner writing from native speaker writing (see S. Granger 1993 & 1994). In the first part of my paper I will describe the ICLE corpus, a 1 million+ word computerized learner corpus of argumentative writing by EFL learners from 11 different mother tongue backgrounds (Chinese,

Czech, Dutch, Finnish, French, German, Japanese, Polish, Russian, Spanish and Swedish), and the LOCNESS corpus (Louvain Corpus of Native English Essays), which contains comparable writing from native English writers. Then I will briefly tackle some of the methodological issues which arise from the compilation and analysis of computerized learner corpora. I will demonstrate that (1) the heterogeneity of learner language calls for the adoption of strict corpus design criteria; and (2) learner corpora call for a new type of contrastive approach, called Contrastive Interlanguage Analysis (CIA), which involves comparing native language and learner language as well as comparing learner languages to each other (see S. Granger forthcoming). The current investigation forms part of a project aimed at automating the CIA approach.

In the second part of my presentation I will justify my choice of grammatical variable and discuss the particular difficulties it poses for automated retrieval. One of the fields in which the passive has proved to be a significant variable is that of text typology. The passive is one of the major features in D. Biber’s (1988,1992) automated multidimensional analysis of linguistic variation. Frequent use of the passive is shown to correlate with discourse that is “abstract and technical in content, and formal in style” (D. Biber 1988:111). Another field which stands to benefit from studies of the passive is that of first and second language instruction, which has traditionally presented the passive as an indicator of weak and inefficient writing. This prescriptive approach is still found in most writing textbooks and usage guides and has been adopted by current grammar and style checkers, which systematically flag all instances of passive forms and suggest replacing them by their active counterparts. Several recent studies, however, have begun to discuss the passive in a more positive light. In his investigation of ESL learner writing, P. Kameen (1983) finds a high correlation between incidence of the passive voice and scores assigned to compositions, with ‘good’ writers using significantly more passives than ‘bad’ writers. He concludes that “mean incidence of passive voice seems to be a reliable indicator of both syntactic maturity and rated quality of writing”. The format of the ICLE corpus enables us to look at low/high passive use in learner writing from a different, but complementary angle, namely that of the relationship between frequency of use of the passive and possible influence from the mother tongue, by comparing the frequency of the passive in learner writing from several different mother tongue backgrounds.

The passive thus appears to be a potentially interesting candidate for automated retrieval. It is also a particularly challenging one however, because ‘be + past participle’ (*be Ved*) is a fuzzy structure,

which may display various degrees of adjectivalness and several types of alternation with the active. In a previous study of the passive in spoken English (Granger 1983) I distinguish no fewer than seven different categories of *be Ved*, only one of which displays the two features generally associated with the passive voice: (a) verbal rather than adjectival status; and (b) direct alternation to a semantically equivalent active structure. The seven categories are illustrated in examples (1) to (7). (1) That attitude was maintained by the government in the further nine days of debates in the Lords. (2) I feel we're all faced with this problem. (3) I am very interested in poetry. (4) He's never finished his D.Phil., you see. I mean, it's nearly finished. (5) You're not supposed to kick that. (6) When we knew without doubt that the war situation was very, very complicated, we left the countryside. (7) I feel I'm geared up to working, you know. It is the category of 'true passives', illustrated in the first example, which is of interest in variation studies. The 'non-passives' (cf examples 2-7), which account for approximately one third of the *be Ved* constructions in my study, do not show significant variation across text categories. It is also the category of 'true passives' that is relevant for SLA specialists. P. Kameen (1983:170), for example, explicitly excludes 'stative passives' such as "I am interested in the results" and "My coat is torn". All this shows that in assessing automated retrieval of passives, it is necessary to consider not only the number but also the type of retrieved *be Ved* forms.

The third section of my presentation will be devoted to the microscopic study of the passive. It will be based on a corpus of ca. 50,000 words composed of native and learner data extracted from LOCNESS and ICLE respectively. The analysis will involve two stages: a purely manual one, in which all the 'true passives' will be retrieved from the data, and a fully automated one, in which all the instances of the Aux(pass) tag will be retrieved from the TOSCA-tagged version of the corpus. The results of the two types of retrieval will be compared both quantitatively (number of retrieved forms) and qualitatively (type of retrieved forms). Automated retrieval will be assessed in terms of precision (ie the proportion of retrieved material that is relevant) and recall (ie the proportion of relevant information that was retrieved) (cf. C. Ball 1994:295). One key objective will be to establish whether the differences in passive frequency between the native speakers and the different categories of learners revealed by the manual analysis are also brought out by the automatic analysis. Particular attention will be paid to the passive forms which escape automated analysis. I will demonstrate that a sizeable proportion of these forms belong to some well-defined categories,

notably that of 'elliptical passives' (eg to be created and destroyed; is not dealt with as it should be) and 'complex passives' (eg be allowed/obliged/expected to do something). I will show that the recall rate of the automated analysis can be improved if appropriate 'repair mechanisms' designed to recover these well-defined categories are applied by the analyst at a post-editing stage. In my conclusion I will claim that, for a whole range of grammatical phenomena, the analyst has much to gain from a small-scale preliminary manual investigation. This will ensure that he effectively understands the theoretical underpinnings of the automated analysis. It will also help him assess the program's reliability and devise mechanisms intended to bring the automated analysis in line with his own research requirements.

References

- Ball C. 1994. Automated Text Analysis: Cautionary Tales, *Literary and Linguistic Computing*, Vol 9, Nr 4, 295-302.
- Biber D. 1988. *Variation across Speech and Writing*, Cambridge University Press.
- Biber D. 1992. On the Complexity of Discourse Complexity: A Multidimensional Analysis, *Discourse Processes* 15, 133-163.
- Granger S. 1983. The *be* + past participle construction in spoken English with special emphasis on the passive, *North Holland Linguistic Series* 49, Elsevier: Amsterdam, New York & Oxford.
- Granger S. 1993. The International Corpus of Learner English, in: J. Aarts, P. de Haan & N. Oostdijk (eds) *English Language Corpora: Design, Analysis and Exploitation*, Rodopi: Amsterdam & Atlanta, 57-69.
- Granger S. 1994. The Learner Corpus: A Revolution in Applied Linguistics, *English Today* 39, Vol 10, Nr 3, 25-29.
- Granger S. Forthcoming. From CA to CIA and back: an integrated contrastive approach to computerized bilingual and learner corpora, in K. Aijmer, B. Altenberg & M. Johansson (eds) *Languages in Contrast*, Lund Studies in English, Lund University Press.
- Kameen P. 1983. Syntactic skill and ESL writing quality, in: A. Freedman, I. Pringle & J. Yalden (eds) *Learning to write: First Language/Second Language*, Longman: London & New York, 162-170.
- Leech G. 1991. The State of the Art in Corpus Linguistics, in K. Aijmer & B. Altenberg (eds) *English Corpus Linguistics*, Longman: London & New York, 8-29. Sinclair J.
- 1992. The Automatic Analysis of Corpora, in: J. Svartvik (ed) *Directions in Corpus Linguistics*, Mouton de Gruyter: Berlin & New York, 379-397.