HIGH FREQUENCY MULTIFUNCTIONAL WORDS:
ACCURACY OF WORD-CLASS TAGGING

Elaine W. Vine: *School of Linguistics and Applied Language Studies, Victoria University of Wellington <elaine.vine@vuw.ac.nz>*

Abstract
The accuracy of automatic word-class tagging of corpora is tested through a comparison of a manual analysis and the automatic word-class tagging of samples of the occurrences of three high frequency multifunctional words, *as*, *like*, and *so*, in the Wellington Corpora of Spoken and Written New Zealand English. The results of the comparison show rather high error rates in automatic word-class tagging of these sorts of words.

1. Introduction

Automatic word-class tagging (also called part-of-speech tagging or grammatical tagging) of corpora is common and it is often assumed by novice users of corpora, though usually not by experienced corpus linguists, that such tagging of English corpora is reasonably accurate. Machine-readable text is run through a computer program which annotates the text by assigning a word-class tag to each word in the text. In a common approach, tags are assigned on the basis of a lexicon of words (and multiword units) with their possible word-classes. A probability matrix is then used to disambiguate the words in the text where one written form may have several possible word classes.

Word-class tags have played an important role in natural language applications such as speech recognition and information retrieval. They have also been used in educational contexts, where frequency data have contributed to decision-making in language teaching and learning, which is my area of interest. I have been particularly interested in the uses of high frequency multifunctional words in English such as 'like', which can be used as preposition, verb, adjective, conjunction, noun and discourse marker. The importance of such words for learners of English has long been recognised. For example, most of them appear in the "first 2000 words of English" (West, 1953). However, while their frequencies have been well researched, their multifunctionality has been largely overlooked. As Biber (2006) points out, his study of university language is: "one of the first large-scale vocabulary investigations to incorporate part-of-speech distinctions, allowing a more detailed description of vocabulary patterns interacting with POS distributions" (p. 243). The multifunctionality of many high frequency words of course presents a particular challenge for any automatic word-class tagger.

I have been working on analysing the uses of high frequency multifunctional words in the Wellington Corpora of Spoken and Written New Zealand English (WCSNZE and WCWNZE) and in the British National Corpus (BNC XML), and I noticed that the word-class tagging of these words that is available seemed not to be as accurate as I might have hoped.

The WCWNZE has been word-class tagged by an early version of the Constituent Likelihood Automatic Word-tagging System (CLAWS), continuously developed by the University Centre for Computer Corpus Research on Language at Lancaster University since the early 1980s, and both Wellington corpora have been word-class tagged by an unpublished programme developed by Douglas Biber at Northern Arizona University. None of this automatic tagging of the Wellington corpora has been checked or corrected.

The BNC XML has been tagged by a later version of CLAWS (C5), and an overall error rate of 1.15% of all words is stated in the Reference Guide, in addition to an ambiguity rate of 3.75% where the tagger has assigned two possible tags to one word (Burnard, 2007, section 6.1). Because of the large size of the BNC XML, very little checking or correcting of the tagging has been carried out. However, on the basis of a manual analysis of a 50,000 word sample from the corpus, the Reference Guide states that error rates for particular tags can be as high as 17% (Burnard, 2007, section 6.3.2).

Biber has published several large-scale studies which have used versions

of his tagging programme. In earlier studies (Biber, 1988; 1995; Biber et al., 1999), he has explained that and how systematic manual checking and correcting of tagging was carried out. This sort of checking involves very tedious and time-consuming work. More recently, he has assumed that readers are familiar with such procedures and has simply stated that the grammatical features "were analyzed using standard procedures of corpus linguistics based on a tagged corpus" (Biber, 2006, p. 242) without further explanation of the procedures.

I report here some comparisons between the unchecked automatic word-class taggings and a manual analysis of random samples of uses of three high frequency multifunctional words in the Wellington corpora. My focus is on the tagging of particular words, rather than on the accuracy rates for particular tags as reported for the BNC XML above. We might expect the multifunctionality of such words to make them particularly difficult for an automatic tagger to tag accurately. My data give us some insight into whether and to what extent that is indeed the case. It must be pointed out here that the work I am reporting in no way constitutes a 'fair test' of automatic taggers overall, precisely because of the nature of the three words I am focusing on. What I am interested in is how well the taggers cope with these words as examples of high frequency multifunctional words in English.

I carried out the manual analysis myself, using reference resources that are themselves corpus-based (for example: Biber et al., 1999; Carter and McCarthy, 2006; Collins Cobuild, 2006). Where an analysis seemed problematic or controversial in some way, I discussed it with colleagues. If the tagger's analysis was plausible (though I might not agree with it), I counted it as 'correct'. I have provided extensive examples below of what I have counted as 'incorrect' tags, so that readers can judge for themselves whether they agree or disagree with my judgements. My manual analysis included listening to sound files, where necessary, to help disambiguate spoken uses (see, for example, #38 below). The automatic tagging of course did not have the benefit of such a possibility.

I compare the word-class tagging and my manual analysis of random samples of occurrences of three high frequency multifunctional words, *as*, *like*, and *so*, in the two Wellington corpora, as shown in Table 1.

These samples were taken for previous studies, not specifically for this study of accuracy. The 'like' samples include only the form 'like'. They are taken from larger random samples of 300 occurrences which included all inflected and compounded forms of 'like', for example, likes, liked, liking,

Table 1: Corpus samples analysed manually

	WCWNZE	WCSNZE
as	100	100
like	217	286
so	100	100

unlikely, likewise, crab-like. The inflected and compounded forms have been removed, and the remaining 'like' occurrences retained for these samples, 217 written and 286 spoken respectively.

2. Tagging in the written corpus sample

As noted earlier, the written corpus sample has been tagged by both CLAWS and the Biber tagger. Table 2 shows that these words appear to be particularly difficult for both automatic word-classing tagging systems in WCWNZE.

Table 2: % tagging errors in samples from WCWNZE

	CLAWS	BIBER
as	54.0%	78.0%
like	6.5%	33.0%
so	23.0%	53.0%

All these error rates, except the CLAWS tagging of 'like', are well above the highest error rate of 17% stated for the BNC XML (see above), though it must be noted again that the BNC XML figures relate to tags not words, so the rates in Table 2 are not directly comparable with the BNC XML figures. In an early study, Biber noted that "[t]he tagging of some lexical items was so problematic that they were systematically excluded" (1988, p. 216), and he gives 'as' as an example of such a word.

Since both taggers do better with 'like', let us compare the analyses of 'like' in more detail. First, let us note how the taggers perform in comparison with each other. Whether or not they match the manual analysis, overall, the two taggers agree with each other in their analysis of 147 instances of 'like'

and disagree in 70, an inter-tagger reliability of 68%. Second, we compare the automatic tagging with the manual analysis. Table 3 shows the tags for the 217 occurrences of 'like' in the WCWNZE sample, broken down according to whether they match (correct) and do not match (incorrect) the manual analysis.

Table 3: Numbers of correct (boldface) and incorrect (*italics*) tags on 'like' in WCWNZE

	PREPOSITION	VERB	CONJUNCTION	NOUN
Manual analysis	169	39	5	4
CLAWS tags	**prep 165**	**verb 36**	**conj 2**	**noun 0**
	verb 2	*prep 3*	*prep 2*	*adj 4*
	adverb 1		*verb 1*	
	adj 1			
Biber tags	**prep 116**	**verb 28**	**conj 0**	**noun 0**
	verb 53	*prep 11*	*prep 3*	*prep 4*
			verb 2	

According to the manual analysis, preposition uses are by far the most frequent with 169 out of 217 occurrences of 'like', and verb uses are next most frequent with 39 occurrences. CLAWS did reasonably well at distinguishing between preposition and verb uses, with only 2.4% (4 out of 169) and 7.8% (3 out of 39) error rates respectively. The Biber tagger struggled, with 31.4% (53 out of 169) and 28.2% (11 out of 39) error rates respectively. Both taggers failed to cope adequately with identifying conjunction and noun uses of 'like' in this sample.

Following the numbered examples below, correct tags are indicated in **boldface** and incorrect tags in *italics*, as in Table 3 above. CLAWS made only four errors in tagging preposition uses of 'like' in the WCWNZE:

1. and waving the front legs about **like** antennae [J01 145]
   CLAWS (adjective) **Biber (preposition)**
2. 'I heard something **like** it once,' he says [L10 132]
   CLAWS (verb) **Biber (preposition)**

3. How could you be so clumsy! It's not **like** you at all. [K19 157]
   CLAWS (verb) Biber (verb)
4. What was she **like** when you married her, your wife? [K60 009]
   CLAWS (adverb) Biber (verb)

The Biber tagger correctly identified examples 1 and 2 as prepositions, but not 3 and 4. With respect to example 3, one would expect a probability matrix to produce errors in the opposite direction, since 'like' occurs more frequently in both written and spoken English as a preposition than as a verb. With respect to example 4, it is interesting to note that both CLAWS and the Biber tagger identified two other occurrences of 'like' in 'what...like' constructions correctly as prepositions:

5. God only knows what this country will be **like** in a few years to come [B22 071]
6. what will he be **like** at 85 [G59 115]

CLAWS correctly identified the other 165 occurrences of 'like' as prepositions in the sample, while Biber made many errors, all of them misidentifying prepositions as verbs, for example:

7. You see what you want to see, **like** everyone else.[K04 135]
8. I'm going to make a good job of this marriage, not **like** Mum and the old man. [K69 032]
9. You just move quietly around minding your own business. **Like** a sheep. [K94 055]
10. but the facts, **like** the sights, remained wonderfully alike. [L11 094]
11. but there would always be an angel hovering, **like** a motto in winged letters [L19 050]
12. the reception area is expandable, **like** a priest's house. [L24 024]

It is interesting to note that all of the 'verb instead of preposition' errors by the Biber tagger occurred in the K and L sections of the corpus, which are the fiction sections. This could suggest that there is a bug in the Biber tagger programme, but there were also many instances of the Biber tagger correctly

identifying 'like' as a preposition in those sections. It is possible that the Biber tagger's problems have something to do with the ways in which 'like' is used as a preposition in fiction texts.

CLAWS made only three errors in tagging verb uses of 'like' in the WCWNZE, and in all three cases, CLAWS misidentified a verb as a preposition:

13. Whether Catholics **like** it or not, they are going to be involved in this struggle [D11 023]
    CLAWS (preposition) Biber (preposition)
14. For doll selling the articles of faith are that young girls **like** things pretty, they like putting clothes on whatever it is [F40 071]
    CLAWS (preposition) Biber (preposition)
15. Well you **like** me. Do you? [K18 029]
    CLAWS (preposition) **Biber (verb)**

This is the sort of error one might expect through a probability matrix, assuming that it recognised that 'like' occurs more frequently as a preposition than as a verb. Although the Biber tagger did outsmart CLAWS on example 15, there were many more instances where the Biber tagger misidentified verbs as prepositions while CLAWS correctly identified them as verbs, for example:

16. You **like** Mary. Do you? [K18 041]
17. 'Oh, yeah, I'd **like** to do nothing but make speeches for the rest of my life.' [F03 174]
18. I do not **like** that painting. [G50 104]
19. They don't **like** Javanese coming here because they don't trust them. [K42 154]
20. I don't **like** the new teacher. [L22 204]

There were only five occurrences of 'like' as a conjunction in the WCWNZE sample. CLAWS identified two of them correctly:

21. She was bald **like** he was and she didn't have any teeth either. [K09 178] **CLAWS (conjunction)** *Biber (preposition)*
22. I wasn't learning really useful things at school, not **like** I did with Thunderbox. [K68 087] **CLAWS (conjunction)** *Biber (verb)*

However, in the other three cases, CLAWS, like the Biber tagger, misidentified conjunctions as prepositions or verbs, though the two taggers did not always agree with each other in their misidentifications:

23. And with you I feel **like** I have to defend Dad. [K95 151] *CLAWS (preposition) Biber (preposition)*
24. and it features a guitar solo **like** you've never heard. [C07 055] *CLAWS (verb) Biber (preposition)*
25. If I do, I could end up with such pain. **Like** I did with Vanessa. [K74 006] *CLAWS (preposition) Biber (verb)*

The four occurrences of 'like' as a noun in the sample were uses of the fixed phrase 'and the like'. CLAWS plausibly identified them as adjective uses, but the Biber tagger misidentified them as prepositions:

26. such as services, investment, intellectual property, safeguards and the **like** on the negotiating agenda. [F05 166]
27. and numerous others relying on landscape and the **like**. [G44 174]
28. Maori weaving, food preservation, wood carvings and the **like**, which do make an attempt to show [A37 188]
29. for fear of meeting 'Methodist Modernists' and the **like**, could be fortified against [D09 223]

3. Tagging in the spoken corpus sample

The spoken corpus sample has been tagged only by the Biber tagger, which had even greater difficulty with this sample than it had with the written corpus sample, as shown in Table 4.

This could shed some light on why the Biber tagger may have had

Table 4: % errors by the Biber tagger in WCWNZE and WCSNZE

	WCWNZE	WCSNZE
as	78%	89%
like	33%	67%
so	53%	88%

particular difficulties with the fiction sections of the written corpus. Fiction writing includes language use that is more like spoken language than written language, for example, in dialogues. The Biber tagger may not be handling the spoken aspects of fiction writing as well as CLAWS. Biber tagging errors for 'like' occur quite evenly across all the discourse categories in the spoken corpus sample: public speaking, private speaking, monologue and dialogue.

Table 5 shows the details of tags for the 286 occurrences of 'like' in the WCSNZE sample which match (correct) and do not match (incorrect) the manual analysis.

Table 5: Numbers of correct (boldface) and incorrect (*italics*) tags on 'like' in WCSNZE

	DISCOURSE MARKER	PREPOSITION	VERB	CONJUNCTION	QUOTATIVE
Manual analysis	111	101	56	13	5
Biber tags	*prep 60* *verb 51*	**prep 47** *verb 54*	**verb 47** *prep 9*	*prep 8* *verb 5*	*verb 3* *prep 2*

According to the manual analysis, discourse marker uses and preposition uses are the most frequent with 111 and 101 out of 286 occurrences of 'like', and verb uses are next most frequent with 56 occurrences. An obvious issue here is that the Biber tagger has not picked up any of the discourse marker uses of 'like', though the system does have a tag for 'adverb + discourse particle'. The Biber tagger identifies discourse markers as either prepositions, for example:

30. i imagine **like** when my thesis mark comes back through i'm gonna have to take some time off [dgz079]

or verbs, for example:

31. he was working **like** tut one day a week every week for dad [dpc028]

When it comes to prepositions, the Biber tag correctly identifies them as prepositions about half the time, for example:

32. organisations **like** ours are being forced to reduce or even eliminate housing services [dgb022]

and incorrectly identifies them as verbs about half the time, for example:

33. i've heard them talking **like** that [dpc269]

The Biber tagger error rates were 50% or more in all categories except verbs. The error rate of 16% (9 out of 56) on verbs was better, but still unacceptably high. For example:

34. cos i **like** the rolling stones [dpc221]

was correctly identified as a verb, but

35. oh goody i **like** surprises [dpc153]

was incorrectly identified as a preposition.

The Biber tagger failed to identify correctly any of the occurrences of 'like' as conjunction or quotative in the sample. Conjunctions were identified as verbs, for example:

36. to make it sound **like** i do actually have some kind of sport up my sleeve you know [dpc269]

or prepositions, for example:

37. so it wasn't **like** robbie was saying it because he'd heard me talk about it [dpc121]

Quotatives were identified as verbs, for example:

38. doctor ranginui walker was talking and saying that how a lot of old maori people ge. feel really nervous about giving urine samples because they KNOW they stick them in the fridge and it's **like** that's only where food goes you know [dpc240]

This is a case where listening to the sound file was used to help in the analysis. Pausing and a change in voice quality indicate quotative rather than, say, a discourse marker signalling repair.

Quotatives were also identified as prepositions, for example:

39. i'll wander round and they'll be **like** man there's heaps of work for one oh two [dpc331]

4. Conclusion

The analysis presented above shows that CLAWS did better than the Biber tagger with the written corpus sample, and the Biber tagger did better with the written corpus sample than the spoken one. However, the rate of errors and the nature of the errors that are evident in this analysis suggest that we should be very cautious about interpreting word-class data derived from automatic tagging, at least in the case of high frequency multifunctional words such as 'as', 'like' and 'so'. As noted earlier, such words are particularly problematic for automatic taggers, and at times for human analysts as well. Experienced corpus linguists are well aware of this, but novice users of corpora may not be. For educational uses, for example, in the teaching of English as a second or foreign language, high frequency words are recognised as important, but teachers need to be aware that corpus word class data for such words may be unreliable.

References

Biber, Douglas 2006. *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam and Philadelphia, PA: John Benjamins.

Biber, Douglas 1995. *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, Douglas 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas et al. 1999. *Longman Grammar of Spoken and Written English*. Harlow, England: Longman/Pearson Education.

British National Corpus (BNC XML Edition). 2007. Oxford University Computing Services.

Burnard, Lou (ed.). 2007. *Reference Guide for the British National Corpus (XML Edition)*. Oxford University Computing Services. Accessed from http://www.natcorp.ox.ac.uk/docs/URG/ on 08 October 2010.

Carter, Ronald and McCarthy, Michael 2006. *Cambridge Grammar of English: A Comprehensive Guide, Spoken and Written English, Grammar and Usage*. Cambridge: Cambridge University Press.

CLAWS part-of-speech tagger for English. Lancaster University Centre for Computer Corpus Research on Language (UCREL). Accessed from http://ucrel.lancs.ac.uk/claws/ on 12 November 2010.

Collins Cobuild Advanced Learner's English Dictionary (5th edition, 2006). Glasgow: HarperCollins.

Wellington Corpus of Spoken New Zealand English. 1998. Victoria University of Wellington.

Wellington Corpus of Written New Zealand English. 1993. Victoria University of Wellington.

West, Michael 1953. *A General Service List of English Words*. London: Longman.