Lars Hinrichs
Nicholas Smith
Birgit Waibel

The part-of-speech-tagged 'Brown' corpora

a manual of information, including pointers for successful use

Lars Hinrichs, Nicholas Smith, Birgit Waibel

The part-of-speech-tagged 'Brown' corpora: a manual of information, including pointers for successful use

Department of English, University of Freiburg, Germany

September 2007

For information on the corpora contact:

Christian Mair

Department of English

University of Freiburg

79085 Freiburg

Germany

christian.mair@anglistik.uni-freiburg.de

1 Introduction: the 'Brown family' of corpora

In this manual, the expression "Brown family" is used to refer to the following four corpora:

(1) the *Standard Corpus of Present-Day Edited American English, for use with Digital Computers* (i.e. the Brown corpus proper), compiled by W. Nelson Francis and Henry Kučera of Brown University, Providence, RI, USA, and comprising texts published in 1961;

(2) the LOB corpus (= Lancaster-Oslo/Bergen Corpus of British English), compiled by Stig Johansson, Geoffrey Leech and their co-workers at Bergen and Lancaster respectively and designed to closely match the Brown corpus in size and composition;

(3) the F-LOB (= Freiburg Update of the LOB) corpus, matching LOB in size and composition but comprising texts published in 1991;

(4) the Frown (= Freiburg Update of the Brown) corpus, matching Brown and comprising texts from 1992.

The latter two corpora were compiled by Christian Mair and his team at the English Department of the University of Freiburg, Germany.

Since the 1990s, the Brown family of corpora has become a widely used resource for the computer-driven study of regional and stylistic variation, and recent and ongoing change in Standard English.

To enable studies of variation between the corpora, they were designed to be closely comparable in terms of:

- size: each corpus is composed of 500 text samples of about 2,000 words each, giving a total of roughly a million words per corpus.
- corpus design: each corpus is ordered according to the same structure of textual genres (cf. Appendix A for the corpus structure in tabular view). All of them are written, edited, and published, i.e. "mainstream standard varieties of public, printed text" (Leech & Smith 2005: 86). It is in this qualified sense that the corpora can be called "representative" of the English language.
- compilation technique: the corpora are made up of text samples that were collected according to similar strategies, i.e. beyond a mere match of genres, samples were also taken from publications that were similar in content and style, and, in the case of periodicals, from titles that had a continuous publishing history from the 1960s to the 1990s, e.g. the *Daily Mail* newspaper, and *Amateur Photographer* magazine (cf. Sand & Siemund 1992 on the strategies adopted to match the sample sources for F-LOB with those of LOB).

Figure 1 below illustrates the unique corpus-linguistic working environment provided by the four corpora of the Brown family:

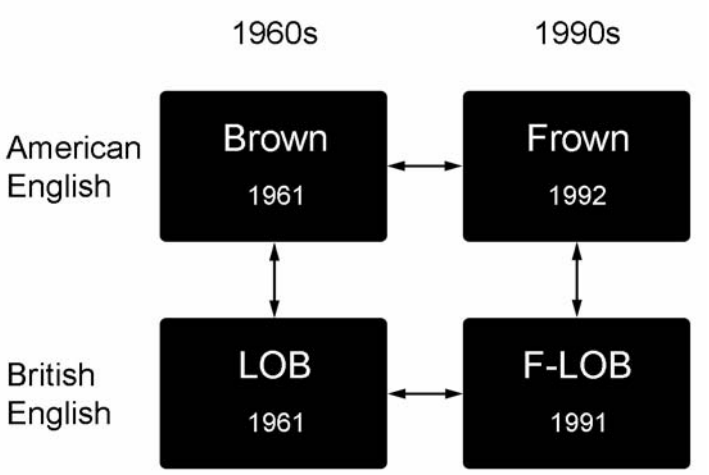

Fig. 1. The Brown quartet of matching corpora of written and published Standard English

At present, these four corpora represent the core of the Brown family. Matching corpora of other regional varieties of English have been created as well, including the Kolhapur Corpus of Indian English (Shastri 1988), the Australian Corpus of English (Collins & Peters 1988), and the Wellington Corpus of New Zealand English (Bauer 1993); they too are frequently considered part of the Brown family. For the remainder of this document, however, the term 'Brown family' will be used to refer only to the two American and the two British corpora.

With the recent completion of the part-of-speech-tagging (POS-tagging) of Frown and F-LOB, a further advance has been made in the provision of resources for studying change in the two largest regional varieties of English during the twentieth century. Previously, the untagged data could be searched for explicit word forms only. For example, one might have searched for all occurrences of the word *catch* in certain contexts, or all words ending in *–ing*. POS-tags add a much needed level of grammatical abstraction to the search. For example, *catch* can now be searched for in either verbal or nominal function (or both), and the search for *–ing*-words can be restricted to verbal forms. (These simple examples are merely for illustration; queries at the grammatical level can be made indefinitely more complex.)

The overall shape of the matching corpus project has been evolving since the initial publication of the Brown corpus in 1963/64.[1] At that time W. Nelson Francis wrote that the corpus could

> certainly be matched by parallel corpora of British English or of English of other periods such as the eighteenth or seventeenth century… But I am quite willing to let someone else prepare the next million words! (Francis 1965: 273, quoted in Leech & Smith 2005: 84)

Some of the current plans for the project have been laid out in recent publications such as Mair et al. (2002) and Leech & Smith (2005). The latter includes a discussion of the considerations that went into the compilation of the corpora and the selection of text samples, and of the kinds of research that the data will ultimately allow. The ultimate basis of the work is the diachronic extension of the synchronic comparative arrangement represented by the original Brown and LOB corpora, which was brought about by Freiburg's decision to move

[1] Francis & Kučera (1979) reported that "Six versions of the Corpus are available," i.e. a non-annotated version and various differently annotated versions. To this count one should add at least the automatically tagged C8-version that was produced at Lancaster in 2002.

text collection forward to the 1990s and Lancaster's subsequent decision to go back in time by sampling the language of the 1930s and the 1900s.

Interesting and important though the subject may be, these theoretical considerations will not be pursued any further here. The orientation of this manual is rather more 'hands-on.' It assembles information that users may find helpful in conducting research on the basis of the POS-tagged Brown family. It is organized as follows:

(1) a brief overview of the history of the corpora;

(2) a description of the POS-tagging that was applied to the corpora and the policies that were followed in post-editing Frown and F-LOB;

(3) a comparative overview of the frequencies of the different word classes, grouped into eleven broad categories, in the four corpora of the Brown family, followed by some suggestions on the use of this information for research;

(4) some concluding remarks.

The appendix contains further information that may serve as reference to users of the data:

(A) the composition of the four corpora, i.e. the fifteen genre categories and numbers of text samples they contain,

(B) the C8 tagset, i.e. a list of the different POS-tags that are assigned to lexical items in the corpora,

(C) a complete table of the frequencies of major POS-tags, in which the fifteen genre categories are grouped into four major genre categories,

(D) association plots showing deviation from independence for the information given in (C), and

(E) an overview of original and revised corpus markup codes.

While Frown and F-LOB were POS-tagged as detailed in this manual, Brown and LOB were originally tagged using different tagsets (Francis & Kučera 1982; Johansson & Hofland 1989 who provide introductions to the respective tagsets used in Brown and LOB as well as comprehensive studies of POS frequencies in the corpora). However, versions of Brown and LOB have now also been produced in the C8 tagset, enabling the kind of four-way comparisons sketched in Figure 1 to be made at the level of grammatical word class (see below for details).

Release of all four corpora, tagged consistently in the C8 tagset, is planned for the third edition of the ICAME CD-ROM. It was in the second edition of this CD (released in 1999) that the F-LOB and Frown corpora were first made generally available, but without the addition of any form of grammatical annotation (ICAME 1999).[2]

Work is still ongoing on the two "prequels" (cf. Leech & Smith 2005) to the British branch of the Brown family: Lancaster1931, already completed at Lancaster University, and Lancaster1901, which is currently being compiled. These two corpora will expand the scope of the suite backward in time to span the twentieth century at four evenly-spaced intervals.

[2] This CD-ROM contains the two older corpora, Brown and LOB, in different versions: without POS-tags as well as with older versions of their POS-tagging.

2 Tagging and post-editing

2.1 Automatic POS-tagging

Figure 2 below shows the main stages involved in producing POS-tagged versions of the Brown family of corpora.

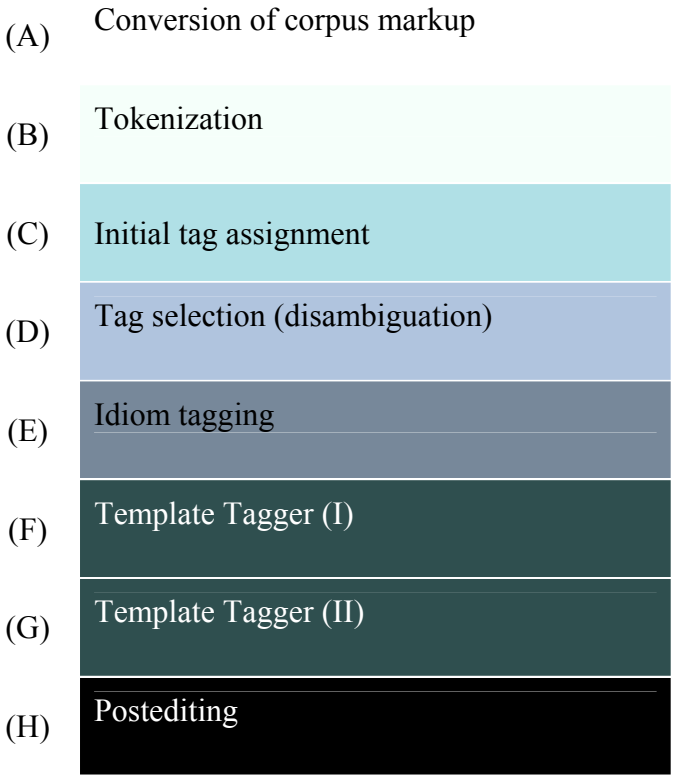

Fig. 2. POS-tagging schema for the Brown family corpora

Stage A is not part of the POS-tagging process proper, but a preliminary phase that enables the tagging software to distinguish ordinary text from features of 'encoding' or 'markup', i.e. codes embedded in the text to represent structural elements such as paragraphs, headings, and chapter divisions, and formatting features such as italics and superscript typeface. Version 1 of the F-LOB and Frown corpora already contained markup to represent such features; however, it was not in a format widely used by linguists or other researchers working with texts, nor was it recognized by UCREL's tagging software.

So our first step was to convert each element of existing markup in F-LOB and Frown to a more standard equivalent; for example, replacing:

```
<h\|>word
```

(i.e. a single-word heading) with:

```
<head>word</head>
```

and:

```
<}><-|>misspelt-word <+|>corrected-form<}/>
```

(i.e. a spelling correction)

with:

```
<reg orig="misspelt-word">corrected-form</reg>
```

A full list of such substitutions is given in Appendix E.

The POS-tagging process proper was handled by two programs operating in tandem: CLAWS4 and Template Tagger. CLAWS has been under continuous development since the early 1980s, for the purpose of tagging corpora such as LOB and the British National Corpus (see Marshall 1983, Leech, Garside & Bryant 1994, Garside & Smith 1997). It performs steps B-E in Figure 2, which can be glossed briefly as follows:

Tokenization: divides up the text or corpus to be tagged into individual (1) word tokens and (2) orthographic sentences.

Tag assignment: assigns to each word token one or more candidate tags. For example, the token *paint* can be tagged as a singular common noun (NN1), a base form verb (VVB), or an infinitive (VVI).

Tag selection (or disambiguation): chooses the most probable tag from any ambiguous set of tags associated with a word token by tag assignment. This stage uses a statistical method of disambiguation, based on the probability of each possible sequence of tags.

Idiom tagging: a matching procedure which operates on lists of patterns which might loosely be termed 'idioms'. Among these are:

- a list of multi-words such as *because of*, *so long as* and *of course*
- a list of place name expressions (e.g. *Mount X* , where *X* is some word beginning with a capital)
- a list of personal name expressions (e.g. *Dr. (X) Y*, where *X* and *Y* are words beginning with a capital)
- a list of foreign or classical language expressions used in English (e.g. *de jure*, *hoi polloi*)

Template Tagging: is like Idiom tagging in CLAWS, but with much more sophisticated pattern-matching. The Template Tagger has two main functions. First, it targets the most error-prone categories introduced (or left unresolved) by CLAWS, "patching" any erroneous tags it finds by using hand-written template rules. A typical rule is the following, which changes the tag on a word like *after* and *before* from conjunction (CS) to preposition (II) if it is not followed by a finite verb within a window of 16 words:

```
#AFTER [CS^II] II, ([!#FINITE_VB])16, #PUNC1
```

The other main function of Template Tagger was first implemented in the tagging of F-LOB and Frown: it is to make the certain POS-tags in the tagged output more discriminatory, and therefore more useful for subsequent linguistic analysis. Using additional hand-crafted rules, Template Tagger distinguishes:

(A) between auxiliary and lexical uses of *be, do* and *have*;

(B) between complementizer and relativizer uses of *that*; and

(C) between relative and interrogative uses of the pronouns *which*, *who*, *whom* and *whose*.

Thus, the set of tags – or "tagset" – applied to F-LOB and Frown is slightly larger than that applied to corpora previously tagged at UCREL. We refer to the new tagset as "C8", to reflect that it is an incremental refinement of the previous tagset, called "C7".[3]

2.2 Post-editing

Correctness of the POS-tags that the UCREL software assigns to natural language corpora varies with genre and quality of the input data; it has also been suggested that earlier versions of the CLAWS tagger worked better with BrE material because the software was originally designed for, and 'trained' on, BrE material. For the language contained in the Brown family, C8 has been found to produce automatic tagging output that is "ca. 98% accurate" overall (Mair et al. 2002: 263); earlier CLAWS versions are reported to have achieved 96.95% accuracy in tagging the British National Corpus (Dickinson & Meurers 2003).

However, as some of the tags and tag sequences which are most interesting to study from a linguistic point of view have rates of correct identification which are considerably below this general average, post-editing by human analysts is indispensable if the corpora are to serve the needs of the wider corpus-linguistic community (Mair et al. 2002). While software exists that performs the task of post-editing POS-tagger output to some success, human post-editing still is the "gold standard" of tagged corpora (Dickinson & Meurers 2003), reaching nearly 100% correctness.

In the 1970s and 1980s the Brown and LOB corpora were tagged using Greene and Rubin's (1971) TAGGIT and CLAWS 1 (Marshall 1983) respectively, and then post-edited. Because the tagging in Brown is so far removed from the current C8 tagset, in respect not only of the delicacy but also of the interpretation of the tags, the corpus was retagged by Nicholas Smith at Lancaster using C8, so as to provide a basis for comparison with the rest of the corpus family. No resources have as yet been available for manual post-editing of Brown, but the data serve the linguist well who wishes to gain preliminary insights into, for instance, broader statistical shifts between Brown and the other three corpora.

In the case of LOB, the original hand-corrected version of the corpus (see Johansson et al. 1986) used a tagset that was historically related to the present one. It was therefore feasible to derive a reliable C8 version without retagging it from scratch. The conversion was done at Lancaster in two stages: (i) a perl script was written to map the original tags in the corpus into the C7 tagset, then (ii) the new grammatical distinctions listed in the C8 tagset were applied using the Template Tagger. Thus, LOB is available in a quality that is clearly more error-free (in fact, nearing 100% correctness) than if it had been automatically tagged in C8 at the outset.

The two newer corpora, F-LOB and Frown, were automatically tagged in C8 at Lancaster and then passed on to Freiburg to be post-edited by human coders. In the procedure adopted for the post-editing of F-LOB and Frown, each of the 500 text samples of each corpus was printed to hard-copy, including POS-annotation, and then read by two different coders in

[3] These most recent additions to the tagging suite's capabilities were implemented by Mike Pacey and Nicholas Smith at Lancaster University.

succession. They marked all errors in the margins, and these corrections were then collated and entered into the computer files of the texts.

In order to manage the considerable workload of hand-correcting all corpus texts and to avoid unnecessary inconsistency, the general guidelines for readers was: "follow the tagger." This is a principle that implies leaving all tagger output uncorrected that is in any way justifiable, including some cases that a human tagger would likely have coded differently. A typical example of follow-the-tagger in practice is the term *White House*. The tagging software failed to recognize the proper name status of this expression, instead tagging it as an adjective followed by a noun. There is no doubt that proper noun tags would be more functionally accurate, but since the tag sequence adjective–common noun is formally (and, one might add, etymologically) correct, follow-the-tagger was observed and the tags left unchanged.

The problems that readers addressed in post-editing, and which clearly required corrections, generally lay in the areas of error, ambiguity, or both.

Correcting erroneous tags is comparatively straightforward. A sequence such as *Southern women at Duke, according to Fiske, are "very conscious of clothes and looks"* (Frown G28) was automatically tagged as follows:

```
<w JJ>Southern <w NN2>women <w II>at <w NN1>Duke<c YCOM>,
<w II>according to <w NP1>Fiske<c YCOM>, <w VVBR>are <quote>
<w RG>very <w JJ>conscious <w IO>of <w NN2>clothes <w CC>and
<w VVZ>looks<c YCOM>, </quote>
```

The plural common noun *looks* is formally identical to a form of the verb *look*, and the occurrence of a conjugated verb at this point in the sentence is not only probable but in fact preferred by the tagger, because of a default bias towards a verb rather than a noun tag in its lexical entry for *looks*. This error was corrected in post-editing.

Other words or phrases might legitimately be tagged in more than one way, and the tagger—which outputs only one tag to each lexical token[4]—might have output a different choice of tag(s) than a human would have selected. For example, *the* and *no* can be considered adverbs in contexts such as *the harder they come* or *it took no less than forty days*. Thus, the tag <w RR> could be justified in these contexts. However, given that the function of *the* and *no* in pre-nominal position is that of an article in the vast majority of cases, it would be equally possible to look at these as atypical uses of the article in special contexts, and at the assignment of the article tag <w AT>, which is sometimes automatically done by the tagger and sometimes not, as simply another possibility. Bearing in mind the potential need of corpus users for consistency in such special cases—after all, it is often these low-frequency phenomena that corpus researchers are interested in—we decided to apply the same decision to each case in the corpora we post-edited. In this case, it seemed reasonable to tag all the cases in which *no* and *the* preceded comparative adjectives as adverbs, with <w RR>.

The overarching, comparatively modest goal in the preparation of POS-tagged Frown and F-LOB was to produce a set of marked-up data that would be largely justifiable and practically free of the most straightforward types of error that occur in automatic tagging. For the following scenarios, however, we decided to go beyond that goal and introduce

[4] The exception to this rule is multiword units, which are lexically identified by the tagger and given only one tag. For instance, the complex preposition *because of* is given only one prepositional tag: `<w II>because of`, rather than tagging *because* as a conjunction and *of* as a preposition. Similarly, *alter ego* is tagged as one singular noun, *for the most part* is tagged as one adverb, *in charge of* is tagged as one preposition, and *in as much as* is tagged as one subordinating conjunction. The tagger recognizes about 700 such multiword unit types. Needless to say, erroneous tag assignments occur here as well, as when in the clause *he was well off the tracks,* the tagger identifies *well off* as a multiword adjective.

consistency in the two corpora, in which we could of course fall back on the rich experience gathered by a previous team involved in post-editing the tagged LOB corpus (cf. Johansson et al. 1986). This decision also accounts for minor discrepancies between the tag frequencies reported here for the current tagged version of F-LOB and those reported in Mair et al. (2002):

Locative nouns. In location names of the pattern *Lombard Street, Rose Park, Chesapeake Bay,* and *Cook Islands*, the second noun was consistently assigned the locative noun tag, <w NNL1> if singular or <w NNL2> if plural.

Frequent alphabetisms and acronyms. Frown contains many alphabetisms and acronyms that the tagger, having been trained on BrE, does not recognize, or which tend to be erroneously tagged for other reasons. While *MP* following a person's name is correctly recognized as the British convention to designate a Member of Parliament, and tagged as <w NNA>, a title such as *MD* (medicinae doctor) is not.

Abbreviation *Dr.* This form is ambiguous. Its correct tags are either <w NNB> (preceding noun of title) or <w NNL1> if it stands for *drive* in the name of a small path or road.

Time can be correctly tagged as either <w NNT1> if it denotes any sense of an expanse in time, or as <w NN1> when used in the meaning of one occurrence of an event, without any duration implied. While in post-editing F-LOB, it was initially decided to follow the tagger, this distinction was now made both in Frown and in F-LOB.

Rock'n' Roll was consistently given only one <w NN1> tag, instead of three tags for all elements of the phrase.

***The blind, the poor, the French,* etc.** Such quasi-nominally employed adjectives were inconsistently tagged as either adjectives, <w JJ>, or as number-neutral common nouns, <w NN>. These cases were normalized to <w JJ>, except for the nationality nouns not marked for plural, which were categorized as <w NN>.

Supposed (to), determined (to), involved (in), known (to), committed (to) had been inconsistently tagged as either adjectives (with the corresponding form of *be* tagged as main verb) or as a participle (with auxiliary *be*). Tags were consistently set to adjective, <w JJ>, and main-verb use of *be*, <w VVB*>, for all occurrences of these five items.

Back occurs in different syntactic functions, which can be tagged in six different ways: noun <w NN1> *(my back hurts)*, adjective <w JJ> *(the back door),* adverb of time <w RT> *(back in the day)*, part of a complex verb construction <w RP> *(to come back)*, verb <w VVI>/<w VV0> *(back out)*, and adverb after nominal head <w RA> *(a few years back)*. The latter is rarely identified correctly by the tagger, and we corrected this in post-editing.

Complex hyphenated forms. Many of these were not recognized by the tagger's lexicon or morphological guesser, and simply tagged as unclassified (<w FU>). We corrected these on a case-by-case basis. A typical error was with premodifying adjectives such as *has-it-all*, in *'a beautiful blond fortyish <hi>wasp</hi> has-it-all knockout* (Frown A12).

Numerals were not consistently distinguished in automatic tagging according to the tagset's provisions for singular, neutral, or plural use, tagged <w MC1>, <w MC> and <w MC2> respectively. We remedied this problem in post-editing.

Henry IV. With a choice of two possible tags for the numeral (<w MC> vs. <w MD>) on the basis of how they are written rather than how they are pronounced we normalized numbers in these contexts to <w MC>.

Gerunds as modifiers in nominal compounds, e.g. *swimming pool, waiting time.* Since the error rate in the automatic tagging of these sequences was rather high (with the tagger being too frequently tricked by the verb-y shape of the first element), we paid special attention to them in post-editing and consistently assigned the tag sequence noun-noun:

```
<w NN1>swimming <w NN1>pool
<w NN1>waiting <w NN1>time
```

Some additional standardisation had, of course, been carried out already at the mark-up stage preceding the tagging process, for example:

Quotations vs. quotation marks. All instances of quotation marks have been checked as to whether they mark quoted language or serve a different function, applying <quote>/</quote>-tags only to the first.

2.3 Summary: the current shape of the corpora

The "Brown family" of corpora looks back on a history of corpus building and development which in some instances extends back more than forty years. Table 1 below summarizes the essential stages in the development of each corpus, so as to enable researchers to assess their current potential and comparability.

Table 1. The evolution of the Brown family of corpora.[5]

	Brown	LOB	Frown	F-LOB
Period sampled	1961	1961	1992	1991
Text samples collected in	1963-64	1970-78	1992-96	1991-96
Text samples collected by	Francis, Kučera and associates	Johansson, Leech, Atwell, Garside and associates	Mair and associates	Mair and associates
Original tagset	"the Brown-tagset"	CLAWS 1	C8	C8
Original tagger	TAGGIT Greene and Rubin (1971)	CLAWS1 (Marshall 1983)	CLAWS4 (Leeech et al 1994) and Template Tagger (Fligelstone et al. 1997)	CLAWS4 and Template Tagger
C8 version produced by*	automatic retagging	automatic mapping of the CLAWS 1-tags onto C8	automatic tagging and manual post-editing	automatic tagging and manual post-editing
Post-editing of C8 version	none	earlier, pre-mapping post-edited version available	completed (Freiburg, 2006)	completed (Freiburg, 2003)

***** All automatically C8-tagged versions of corpora were produced by Nicholas Smith at Lancaster University.

[5] In the tables and diagrams presenting the four corpora at various places throughout this manual, they are listed in different orders according to the purpose of the presentation at hand.

3 Word-class frequencies in the Brown family of corpora

3.1 Global POS-tag frequencies in the Brown family of corpora

To provide a source of reference for linguists using the four Brown corpora in future research, this manual includes a tabular overview of the frequencies of major word classes, based on the frequencies of tags. Table 2 and Figure 3 both give this information, the first for numeric detail and the second for quick and easy graphic reference.

The figures were determined through corpus searches for the tags named in the second column of Table 2, i.e. each search typically contained the first letter of the greater class of tags, complemented by a wildcard.[6] The concordancer software we used was WordSmith 3,[7] and each search was double-checked with Monoconc.

Table 2. Major POS tags in four corpora ('normalized': occurrences per million words)

		LOB		F-LOB		Brown		Frown	
word class	tags included	raw	normalized	raw	normalized	raw	normalized	raw	normalized
adj	J*	75,407	74,660	80,148	79,402	80,810	79,697	83,276	82,322
adv	R*	62,707	62,085	59,435	58,882	56,450	55,672	54,907	54,278
art	A*	112,941	111,821	109,351	108,333	115,429	113,839	107,407	106,177
conj	C*	56,396	55,837	56,033	55,512	57,377	56,587	55,441	54,806
det	D*	31,878	31,562	29,499	29,224	30,532	30,111	27,332	27,019
noun	N*	253,831	251,315	266,083	263,607	269,282	265,572	279,209	276,011
num	M*	15,512	15,358	15,559	15,414	14,012	13,819	15,724	15,544
prep	I*	121,331	120,128	118,039	116,940	121,391	119,719	115,844	114,517
pron	P*, WPR	58,765	58,182	55,391	54,875	55,043	54,285	56,643	55,994
verb	V*	179,900	178,117	178,429	176,768	177,055	174,616	175,244	173,237
misc	Misc Total	41,344	40,934	41,427	41,041	36,588	36,084	40,558	40,094
	TOTAL	1,010,012	1,000,000	1,009,394	1,000,000	1,013,969	1,000,000	1,011,585	1,000,000

Note that the totals given at the end of the "raw" columns can be considered the most exact gauge of the size of each corpus in number of words.[8, 9]

---

[6] Note that the search term for specific POS-terms should begin with a wildcard to allow for an immediately left-aligned tag—a rare but possible case. A search for "all conjunctions" in a C8-tagged corpus using WordSmith is therefore best formulated like this: `*<w C*`

[7] We chose WordSmith 3 over the newer version 4 for consistency and continuity. When tested, version 4 exhibited some problems, unresolved at the time of writing, in handling concordance searches that combined both corpus text and markup.

[8] The aim in corpus compilation was to collect 500 samples of 2,000 words apiece. The fact that the total corpus sizes all exceed one million words is owed to the policy of including the ends of running sentences in the text samples, rather than cutting off at exactly 2,000 words.

The fields of the mosaic plot in Figure 3 represent the number of POS tag classes in each of the four corpora. They allow a rough, preliminary comparison of corpora for selected POS classes. For example, the increase in nouns and adjectives in both BrE and AmE from the 1960s to the 1990s becomes apparent in this visualization. Since word classes differ greatly in the measures of the mean frequencies, however, increases are not equally significant (by a χ^2 measure) for all word classes. Statistically, an increase from 1,000 to 1,020 is much more significant than one from 100 to 102, even though both are 2% increases.

Therefore, a visualization of the statistical significance of discrepancies between the different corpora is provided in Figure 4. More precisely, the association plot—which was produced using the assocplot function of the statistics software package R—indicates deviations from independence for each of the raw frequencies given in Table 2. 'Independence' would be the state in which any differences between the frequencies observed in one corpus and the others are unlikely to be statistically significant, i.e. the variance in the data can be attributed to chance.

For each cell in Table 2 giving an observed (or 'raw') frequency, the association plot in Figure 4 plots one box. Its height is proportional to the cell's contribution to the table's overall χ^2—in other words, box height signals statistical significance. The full area of the box is proportional to the difference between observed and expected frequency for that particular cell.[10]

It should be mentioned that the association plots included in this manual are intended as a first visual orientation only. Corpus users who require numeric values for the statistical significance of any aspect of variance in or among the corpora are encouraged to use the frequencies reported in Table 2 and in Appendix C in computing these, according to their needs and preferences.

[9] Appendix C also gives the sizes of the four genre-based subsections, which will be useful for researchers wanting to compute normalized frequencies for linguistic phenomena in any of the subsections.

[10] 'Observed' = 'raw' frequency. The 'expected' frequency of a cell is essential to the computation of statistical significance measured by χ^2. It is defined as the sum total of the row in which it stands multiplied by the column total, divided by the grand total of all frequencies in the table.

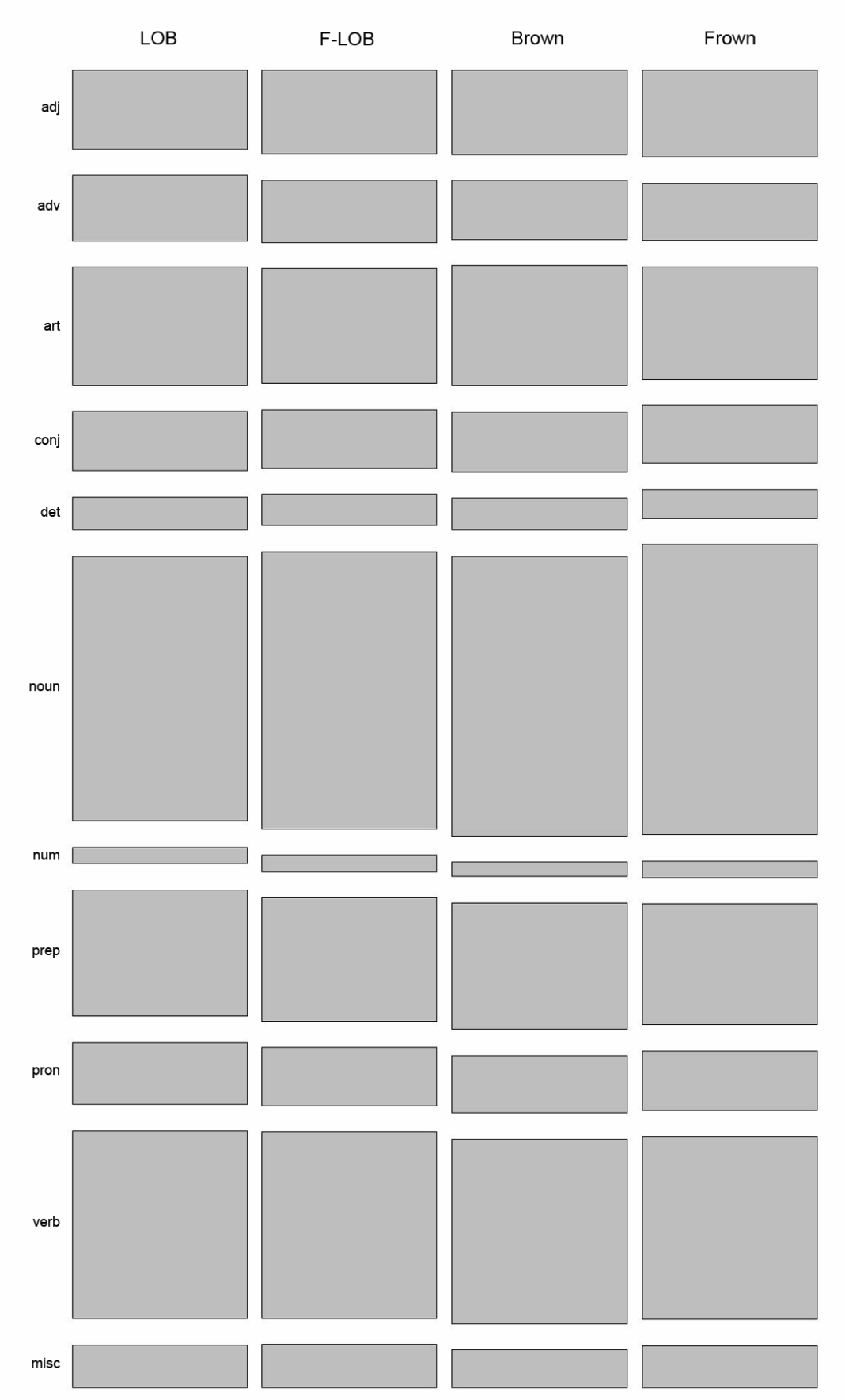

Fig. 3. Raw frequency comparison for global POS-tags in the four corpora

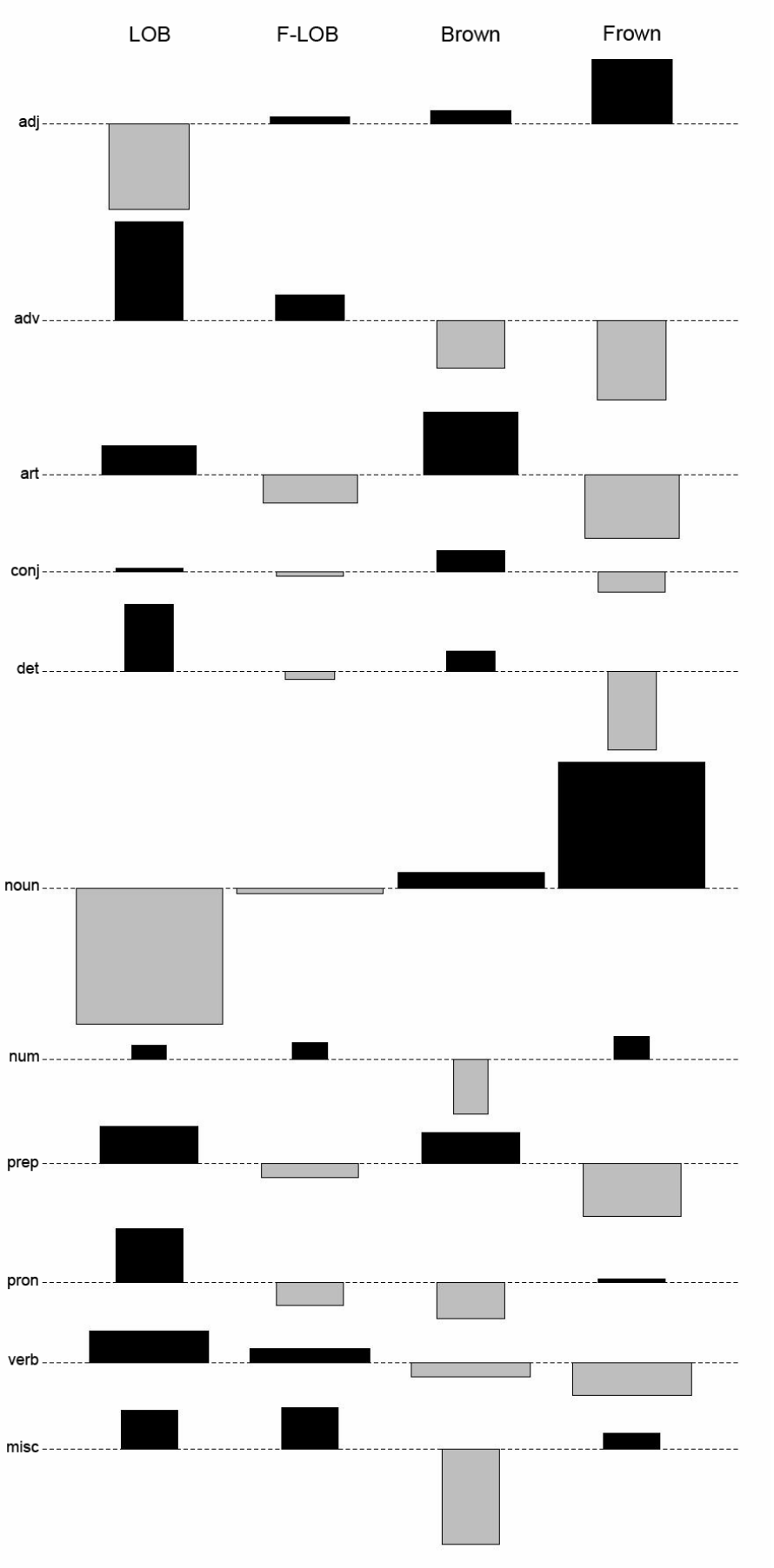

Fig. 4. Association plot for raw frequencies of POS-tags in the four corpora (box sizes indicate deviation of observed frequencies from independence)

3.2 A look at nouns and verbs

The larger shifts in word class frequencies that are suggested by Figures 3 and 4 can be usefully broken up into more detailed views by taking the genre-specific perspective. In numerous publications on the corpora it has proven useful to group the 15 genre categories (cf. Appendix A) into four larger groups as follows: Press: categories A-C, General Prose: D-H, Academic (or Learned): J, Fiction: K-R.

Working with POS-frequencies for these four subgroups, the increase in nouns in both BrE and AmE can be more sensibly investigated. As Figure 5 shows, the shift is pronounced and significant in Press and Academic language, the two 'informational' genre groups. But it is far from global, as General Prose and Fiction actually display the opposite trend. Thus, claims to the effect that English is generally "nominalizing" must be taken with extreme caution. As shown below, however, an investigation of the reasons for the nominalization of *informational* genres is a promising path of research.

Similarly, anyone suspecting that the process of nominalization in Press and Academic writing is complemented by de-verbalization in equal proportions will be proven wrong, at least by the Brown corpora. Figure 6 shows that only AmE Academic writing is de-verbalizing, while BrE actually shows an increase in verbs, as does BrE and AmE press writing.

Figures 5 and 6 show only the association plots for nouns and verbs for genres. Figures 7 and 8 in Appendix D are meant to provide a more generally useful source of orientation in that they show association plots for all tags commonly associated with the noun phrase, split up into genres (Figure 7), and the same for tags commonly associated with the verb phrase (Figure 8).

The next, final section of this manual is a case study showing how a linguist might systematically put the information provided in this manual to use in generating and pursuing a research question.

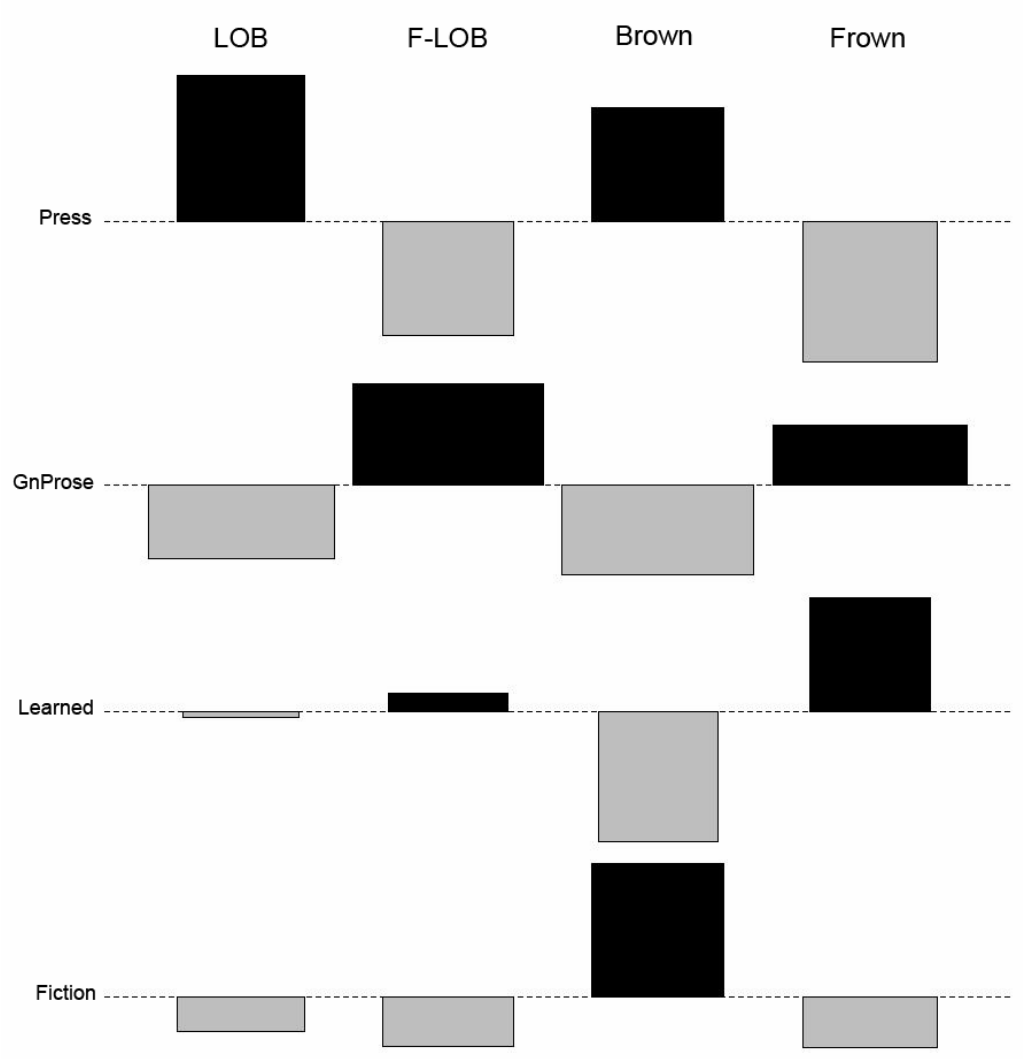

Fig. 5. Nouns in four corpora (<N*>-tags), broken down into genres (deviation of observed frequencies from independence)

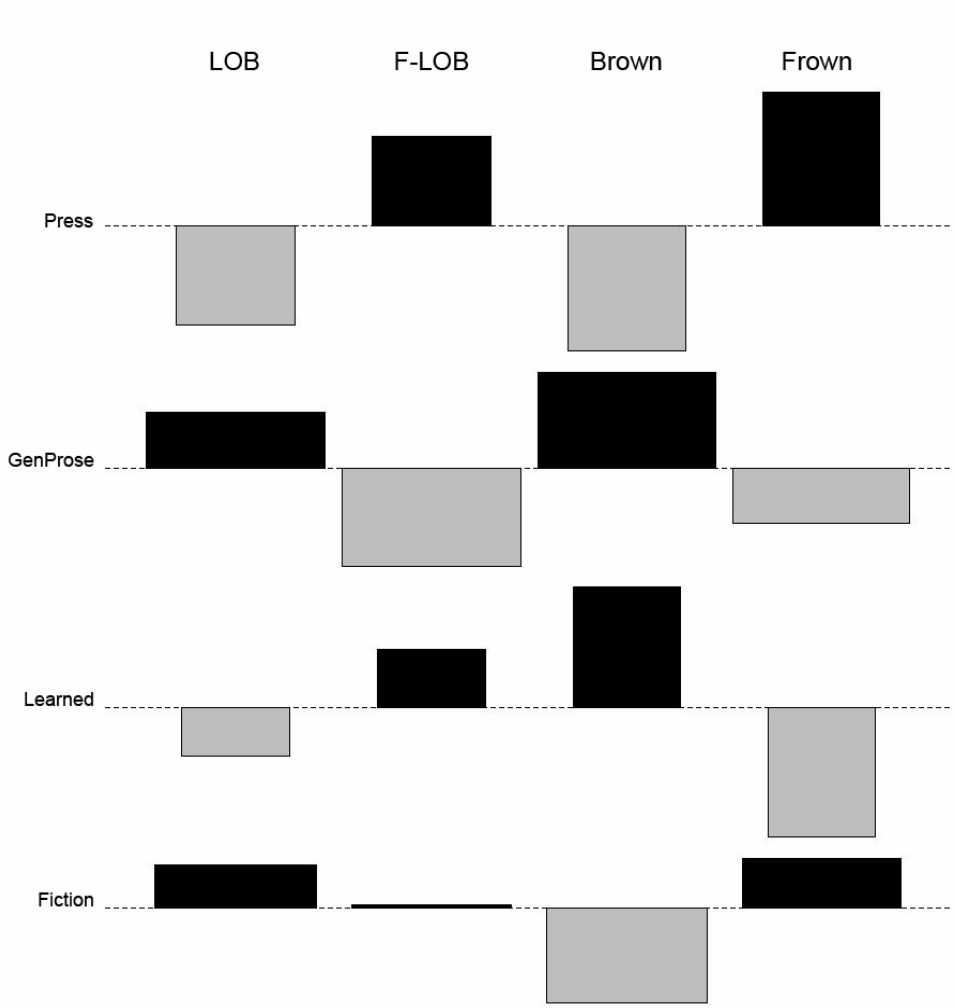

Fig. 6. Verbs in four corpora, broken down into genres (deviation of observed frequencies from independence)

3.3 How to apply this manual: a corpus-linguistic case study

Let us imagine a linguist who intends to translate his[11] fascination with language corpora in general, and with Table 2 of this manual in particular, into a corpus-linguistic investigation of a suitable research question.[12]

As a first step, he might refer to Figure 4, where he will find graphic representations of the significance of the variance displayed in Table 2. The association plot gives a first orientation as to which aspects of variation on the level of POS frequency might be worth investigating. Let us assume that this linguist notes the highly significant increase in nouns from the 60s to the 90s, and decides to examine the phenomenon more closely.

Next, he would turn to Figure 7 in order to ascertain the spread of the observed diachronic increase in nouns across genres, and to relate it to other word classes in the noun phrase. There he will find i) that the increase is only evident in the 'informational' genres, Press and Academic writing, but not in General Prose and Fiction, and ii) that other noun phrase-related content words, viz. adjectives, are increasing parallel to nouns, but that none of the function word classes are increasing, that in fact they are decreasing.

The second observation allows conclusions about the nature of the change related to the nominalization that our linguist initially observed. Informational writers in the 1990s seem to be using noun phrases with more content words than writers in the 1960s—but not a greater number of noun phrases, which would entail an increase in, for example, prepositions and determiners. In other words, noun phrase structure in the informational genres seems to have changed from 1961 to 1991/2, possibly in the direction of compressing more information into longer but not necessarily structurally more complex noun phrases. A trip to the library and review of the relevant literature will confirm that other studies, using different data, have previously found this to be the case. In fact, Biber (2003) writes of the "informational explosion" of the twentieth century that has been exerting pressure on writers of expository prose to package ever more information into ever decreasing amounts of space. He shows that this affects noun phrase syntax in press language, favoring more compact types of noun phrase premodification.

In his search for a suitable research project, our imaginary linguist may therefore find it suitable to turn to aspects of grammar and writing that relate to information density. In particular, a variationist approach seems promising that considers variables in which one variant packages information more densely than the competing one. This is the case with the two genitive forms of English grammar: the *of*-genitive and the genitive with *'s* (or simply *'*). In many contexts these two constructions are interchangeable, but the *s*-construction is the more economical choice: *Jack's house* is more compact than *the house of Jack*.

Having chosen the genitive as his area of investigation, the researcher will put this manual aside until later. He will turn to the data and conduct his own analyses where he requires more specific information than what he will find in the manual. The tags <w GE> and <w IO>, which mark the genitive-*s* and the preposition *of*, respectively, can be retrieved in a concordance search from the data; in this manual their numbers are included in the counts for

[11] This assumes a male linguist for no reason other than the need to make a clear choice in the name of readability.

[12] The project sketched here draws on observations made, among others, in Mair et al. (2002), Biber (2003), and Mair (2006). In particular, Hinrichs & Szmrecsanyi (2007) is a study of genitive variation that further develops most of the questions showcased in this example.

'miscellaneous' tags. More than likely, the tokens will have to be further sorted and coded in order to conduct meaningful research.

Among the possible research questions concerning genitive variation in 1960s and 1990s BrE and AmE is: has the *s*-genitive become more frequent over time? This can be answered by a search for the <w GE> tag in the corpora; in fact the answer will simply be: yes, significantly so. In finding out why this is so, more specific questions concerning the conditioning factors in genitive choice will be interesting, such as:

(1) Given the increase of the *s*-genitive from LOB to F-LOB and from Brown to Frown in expository prose, is there a corresponding decrease in the use of the *of*-genitive? To answer this question, the instances of *of* in the data would need to be further sorted, retaining only *of* in genitival use.

(2) Phonological constraints are known to impact genitive choice in speech as well as in writing (Zwicky 1987; Hinrichs & Szmrecsanyi 2007): if the possessor noun ends in a sibilant, the *s*-genitive is disfavored. But is this constraint as powerful in BrE as in AmE writing? And has it grown stronger from the 1960s to the 1990s? This research question requires even further data reduction and coding. It would have to be pursued in a variationist study in the narrower sense: based on only those instances of *s*- and *of*-genitives that are interchangeable, i.e. only those *s*-genitives for which an *of*-genitive could have been used equally well, and vice versa.

(3) A number of other constraints have been shown to also exert statistically significant influence upon genitive choice, among them semantic factors and discourse-related factors. They all can be analyzed in a variationist study, and data will have to be coded accordingly. One factor that is closely related to the issue of textual economy, whose relation to noun phrase complexity we started out investigating, is the impact of textual density—which can be measured in type-token-ratio (TTR)—on genitive choice. A possible question would be: is the *s*-genitive, the more economical option, more frequently selected in textual environments in which TTR is higher? Data coding for this question would involve determining the TTR for the immediate environment of each genitive token.

These questions relate to our theme of noun phrase structure and density in different ways. For example, question (2) is related to a constraint that has a low correlation with economy and density, while (3) is more strongly related to economy. Questions (2) and (3) might be fruitfully treated in comparison.

But the process of selecting a research question is not the topic of this manual, though we should point out that the information provided here will be helpful again at a later step. Assuming that the researcher has opted for a question like (2) above, then he will eventually produce numeric data that allows a statistical examination of the phonological genitive constraint in the Brown data. For example, his analysis may yield a contingency table like this:

Table 3. Genitive tags in four corpora – hypothetical contingency table

	LOB	F-LOB	Brown	Frown
s-genitives with possessor head nouns ending in sibilants	aa	ab	ac	ad
s-genitives with possessor head nouns not ending in sibilants	ba	bb	bc	bd
of-genitives with possessor head nouns ending in sibilants	ca	cb	cc	cd
of-genitives with possessor head nouns not ending in sibilants	da	db	dc	dd

The individual cells of the table will contain raw frequencies, represented here by letter symbols. While the statistical significance of variation among the four corpora (e.g., *p*-values) can only be computed based on those raw frequencies, it will also be beneficial to compute normalized frequencies for each of the cells in order to facilitate direct comparisons among the different corpora. To compute normalized frequencies of occurrences per one million words (or per another multiple of ten, as the case merits), the total size of the samples in which aa and ab occurred will be needed. This information is given in Appendix C.

4 Concluding remarks

Section 3 has provided suggestions on how the tagged corpora of the Brown family and the general statistical surveys provided in this manual might be used in practice. The 'hypothetical' research project sketched above actually draws heavily on research that is already being carried out. But there is no doubt that the Brown family of corpora will provide an extraordinarily rich environment for investigation of questions of grammatical variation in written English for a good many years to come. We encourage members of the academic community to explore and use the data freely.

F-LOB and Brown were tagged using the same tagset and post-edited by partially overlapping teams, which should bring them close to the gold standard of 100% correctness and perfect comparability. This gold standard is within reach also for the new CLAWS 8 version of LOB, which has been derived from the original post-edited version by a straightforward process. Comparisons of any one or any combination of these three with the uncorrected CLAWS 8 version of Brown should be undertaken with the required caution. For F-LOB and Frown, the Freiburg research team welcomes feedback on any errors found in the POS-tagging as well as all other aspects of the data, e.g. markup features. Thus the quality of the grammatical annotation, which we hope is already high, will be able to be improved even further in future releases. Please direct all correspondence in this matter to Christian Mair at <christian.mair@anglistik.uni-freiburg.de>.

5 Appendix

(A) Text categories in the Brown family of matching 1-million-word corpora of written StE

Genre group	Category	Content of category	No. of texts
Press (88)	A	Reportage	44
	B	Editorial	27
	C	Review	17
General Prose (206)	D	Religion	17
	E	Skills, trades and hobbies	36
	F	Popular lore	48
	G	Belles lettres, biographies, essays	75
	H	Miscellaneous	30
Learned (80)	J	Science	80
Fiction (126)	K	General fiction	29
	L	Mystery and detective Fiction	24
	M	Science fiction	6
	N	Adventure and Western	29
	P	Romance and love story	29
	R	Humor	9
TOTAL			500

(B) UCREL C8 Tagset

(with additions to the previous C7 tagset [used, for example, in the tagging of the British National Corpus sampler] in italics)

APPGE	possessive pronoun, pre-nominal (e.g. my, your, our)
AT	article (e.g. the, no)
AT1	singular article (e.g. a, an, every)
BCL	before-clause marker (e.g. in order (that), in order (to))
CC	coordinating conjunction (e.g. and, or)
CCB	adversative coordinating conjunction (but)
CS	subordinating conjunction (e.g. if, because, unless, so, for)
CSA	as (as conjunction)
CSN	than (as conjunction)
CST	that (as conjunction)
CSW	whether (as conjunction)
DA	after-determiner or post-determiner capable of pronominal function (e.g. such, former, same)
DA1	singular after-determiner (e.g. little, much)
DA2	plural after-determiner (e.g. few, several, many)
DAR	comparative after-determiner (e.g. more, less, fewer)
DAT	superlative after-determiner (e.g. most, least, fewest)
DB	before determiner or pre-determiner capable of pronominal function (all, half)
DB2	plural before-determiner (both)
DD	determiner (capable of pronominal function) (e.g any, some)
DD1	singular determiner (e.g. this, that, another)
DD2	plural determiner (these, those)
DDL	*which, relative*
DDLGE	*whose, relative*
DDQ	wh-determiner (which, what), *interrogative*
DDQGE	wh-determiner, genitive (whose), *interrogative*
DDQV	wh-ever determiner, (whichever, whatever)
EX	existential there
FO	formula
FU	unclassified word
FW	foreign word
GE	germanic genitive marker - (' or's)
IF	for (as preposition)
II	general preposition
IO	of (as preposition)
IW	with, without (as prepositions)
JJ	general adjective
JJR	general comparative adjective (e.g. older, better, stronger)
JJT	general superlative adjective (e.g. oldest, best, strongest)
JK	catenative adjective (able in be able to, willing in be willing to)
MC	cardinal number,neutral for number (two, three…)
MC1	singular cardinal number (one)
MC2	plural cardinal number (e.g. sixes, sevens)
MCGE	genitive cardinal number, neutral for number (two's, 100's)
MCMC	hyphenated number (40-50, 1770-1827)
MD	ordinal number (e.g. first, second, next, last)
MF	fraction,neutral for number (e.g. quarters, two-thirds)
ND1	singular noun of direction (e.g. north, southeast)
NN	common noun, neutral for number (e.g. sheep, cod, headquarters)
NN1	singular common noun (e.g. book, girl)
NN2	plural common noun (e.g. books, girls)
NNA	following noun of title (e.g. M.A.)
NNB	preceding noun of title (e.g. Mr., Prof.)
NNL1	singular locative noun (e.g. Island, Street)
NNL2	plural locative noun (e.g. Islands, Streets)
NNO	numeral noun, neutral for number (e.g. dozen, hundred)
NNO2	numeral noun, plural (e.g. hundreds, thousands)
NNT1	temporal noun, singular (e.g. day, week, year)
NNT2	temporal noun, plural (e.g. days, weeks, years)
NNU	unit of measurement, neutral for number (e.g. in, cc)
NNU1	singular unit of measurement (e.g. inch, centimetre)
NNU2	plural unit of measurement (e.g. ins., feet)
NP	proper noun, neutral for number (e.g. IBM, Andes)
NP1	singular proper noun (e.g. London, Jane, Frederick)
NP2	plural proper noun (e.g. Browns, Reagans, Koreas)
NPD1	singular weekday noun (e.g. Sunday)
NPD2	plural weekday noun (e.g. Sundays)
NPM1	singular month noun (e.g. October)
NPM2	plural month noun (e.g. Octobers)
PN	indefinite pronoun, neutral for number (none)
PN1	indefinite pronoun, singular (e.g. anyone, everything, nobody, one)
PNQO	objective wh-pronoun (whom), *interrogative*
PNL0	objective wh-pronoun (whom), *relative*
PNLS	subjective wh-pronoun (who), *relative*
PNQS	subjective wh-pronoun (who), *interrogative*
PNQV	wh-ever pronoun (whoever)
PNX1	reflexive indefinite pronoun (oneself)
PPGE	nominal possessive personal pronoun (e.g. mine, yours)
PPH1	3rd person sing. neuter personal pronoun (it)
PPHO1	3rd person sing. objective personal pronoun (him, her)
PPHO2	3rd person plural objective personal pronoun (them)
PPHS1	3rd person sing. subjective personal pronoun (he, she)
PPHS2	3rd person plural subjective personal pronoun (they)
PPIO1	1st person sing. objective personal pronoun (me)
PPIO2	1st person plural objective personal pronoun (us)
PPIS1	1st person sing. subjective personal pronoun (I)
PPIS2	1st person plural subjective personal pronoun (we)
PPX1	singular reflexive personal pronoun (e.g. yourself, itself)
PPX2	plural reflexive personal pronoun (e.g. yourselves, themselves)
PPY	2nd person personal pronoun (you)
RA	adverb, after nominal head (e.g. else, galore)

REX	adverb introducing appositional constructions (namely, e.g.)
RG	degree adverb (very, so, too)
RGQ	wh- degree adverb (how)
RGQV	wh-ever degree adverb (however)
RGR	comparative degree adverb (more, less)
RGT	superlative degree adverb (most, least)
RL	locative adverb (e.g. alongside, forward)
RP	prep. adverb, particle (e.g about, in)
RPK	prep. adv., catenative (about in be about to)
RR	general adverb
RRQ	wh- general adverb (where, when, why, how)
RRQV	wh-ever general adverb (wherever, whenever)
RRR	comparative general adverb (e.g. better, longer)
RRT	superlative general adverb (e.g. best, longest)
RT	quasi-nominal adverb of time (e.g. now, tomorrow)
TO	infinitive marker (to)
UH	interjection (e.g. oh, yes, um)
VAB0	*be [base form, finite], auxiliary use*
VABDR	*were, auxiliary use*
VABDZ	*was, auxiliary use*
VABG	*being, auxiliary use*
VABI	*be [infinitive], auxiliary use*
VABM	*am, auxiliary use*
VABN	*been, auxiliary use*
VABR	*are, auxiliary use*
VABZ	*is, auxiliary use*
VAD0	*do [base form, finite , auxiliary use*
VADD	*did, auxiliary use*
VADG	*doing, auxiliary use*
VADI	*do [infinitive] , auxiliary use*
VADN	*done, auxiliary use*
VDZ	*does, auxiliary use*
VAH0	*have [base form, finite , auxiliary use*
VAHD	*had [past], auxiliary use*
VAHG	*having, auxiliary use*
VAHI	*have [infinitive], auxiliary use*
VAHN	*had [past participle], auxiliary use*
VAHZ	*has, auxiliary use*
V*V*B0	be, base form (finite i.e. imperative, subjunctive), *main-verb use*
V*V*BDR	were, *main-verb use*
V*V*BDZ	was, *main-verb use*
V*V*BG	being, *main-verb use*
V*V*BI	be, infinitive (To be or not... It will be ..) , *main-verb use*
V*V*BM	am, *main-verb use*
V*V*BN	been, *main-verb use*
V*V*BR	are, *main-verb use*
V*V*BZ	is, *main-verb use*
V*V*D0	do, base form (finite), *main-verb use*
V*V*DD	did, *main-verb use*
V*V*DG	doing, *main-verb use*
V*V*DI	do, infinitive (I may do... To do...), *main-verb use*
V*V*DN	done, *main-verb use*
V*V*DZ	does, *main-verb use*
V*V*H0	have, base form (finite), *main-verb use*
V*V*HD	had (past tense), *main-verb use*
V*V*HG	having, *main-verb use*
V*V*HI	have, infinitive, *main-verb use*
V*V*HN	had (past participle), *main-verb use*
V*V*HZ	has, *main-verb use*
VM	modal auxiliary (can, will, would, etc.)
VMK	modal catenative (ought, used)
VV0	base form of lexical verb (e.g. give, work)
VVD	past tense of lexical verb (e.g. gave, worked)
VVG	-ing participle of lexical verb (e.g. giving, working)
VVGK	-ing participle catenative (going in be going to)
VVI	infinitive (e.g. to give... It will work...)
VVN	past participle of lexical verb (e.g. given, worked)
VVNK	past participle catenative (e.g. bound in be bound to)
VVZ	-s form of lexical verb (e.g. gives, works)
WPR	*that, relativiser*
XX	not, n't
ZZ1	singular letter of the alphabet (e.g. A,b)
ZZ2	plural letter of the alphabet (e.g. A's, b's)

(C) Major POS groups: totals and classification by genre

	LOB - 1960s BrE		Brown - 1960s AmE		F-LOB - 1990s BrE		Frown - 1990s AmE	
	occ.	**p.m.**	**occ.**	**p.m.**	**occ.**	**p.m.**	**occ.**	**p.m.**
adj Press	13,724	77,198	14,181	79,393	13,949	78,254	14,403	80,537
adj Gen. Prose	32,695	78,862	35,267	84,412	35,361	85,472	37,445	90,123
adj Learned	13,877	86,120	15,508	96,371	15,066	94,029	16,261	101,153
adj Fiction	15,111	58,908	15,854	61,776	15,772	61,322	15,167	59,130
adj all genres	**75,407**	**74,660**	**80,810**	**79,697**	**80,148**	**79,402**	**83,276**	**82,322**
adv Press	9,599	53,995	8,609	48,198	9,442	52,969	8,641	48,318
adv Gen. Prose	24,279	58,562	21,428	51,288	22,032	53,254	20,978	50,490
adv Learned	8,312	51,584	8,016	49,814	8,586	53,586	7,470	46,468
adv Fiction	20,517	79,983	18,397	71,685	19,375	75,331	17,818	69,465
adv all genres	**62,707**	**62,085**	**56,450**	**55,672**	**59,435**	**58,882**	**54,907**	**54,278**
art Press	19,759	111,145	20,136	112,733	19,074	107,005	18,300	102,328
art Gen. Prose	47,696	115,046	48,050	115,009	46,287	111,882	44,567	107,264
art Learned	18,274	113,408	18,379	114,212	16,708	104,277	16,424	102,167
art Fiction	27,212	106,083	28,864	112,470	27,282	106,074	28,116	109,613
art all genres	**112,941**	**111,821**	**115,429**	**113,839**	**109,351**	**108,333**	**107,407**	**106,177**
conj Press	8,763	49,292	8,940	50,051	8,904	49,951	8,853	49,503
conj Gen. Prose	24,507	59,112	24,719	59,165	23,951	57,893	23,865	57,438
conj Learned	9,115	56,567	9,076	56,401	9,273	57,874	9,067	56,402
conj Fiction	14,011	54,620	14,642	57,053	13,905	54,063	13,656	53,239
conj all genres	**56,396**	**55,837**	**57,377**	**56,587**	**56,033**	**55,512**	**55,441**	**54,806**
det Press	5,175	29,110	4,842	27,108	4,561	25,587	4,582	25,621
det Gen. Prose	14,171	34,181	13,621	32,602	12,577	30,400	11,430	27,510
det Learned	5,604	34,778	5,366	33,346	5,347	33,371	4,549	28,297
det Fiction	6,928	27,008	6,703	26,119	7,014	27,271	6,771	26,397
det all genres	**31,878**	**31,562**	**30,532**	**30,111**	**29,499**	**29,224**	**27,332**	**27,019**
noun Press	52,661	296,219	55,588	311,213	53,247	298,714	55,700	311,457
noun Gen. Prose	107,732	259,856	114,144	273,206	114,830	277,559	120,014	288,851
noun Learned	42,067	261,067	43,793	272,141	44,255	276,202	47,096	292,964

noun Fiction	51,371	200,264	55,757	217,260	53,751	208,986	56,399	219,877
noun all genres	**253,831**	**251,315**	**269,282**	**265,572**	**266,083**	**263,607**	**279,209**	**276,011**
num Press	2,986	16,796	3,209	17,966	2,696	15,124	2,830	15,824
num Gen. Prose	6,976	16,827	5,956	14,256	7,059	17,063	6,848	16,482
num Learned	3,930	24,389	3,060	19,016	4,176	26,063	4,048	25,181
num Fiction	1,620	6,315	1,787	6,963	1,628	6,330	1,998	7,789
num all genres	**15,512**	**15,358**	**14,012**	**13,819**	**15,559**	**15,414**	**15,724**	**15,544**
prep Press	21,383	120,280	21,137	118,337	20,288	113,815	19,770	110,548
prep Gen. Prose	52,720	127,164	53,283	127,534	52,461	126,805	50,968	122,670
prep Learned	22,802	141,509	21,795	135,440	20,760	129,566	20,859	129,755
prep Fiction	24,426	95,222	25,176	98,100	24,530	95,374	24,247	94,529
prep all genres	**121,331**	**120,128**	**121,391**	**119,719**	**118,039**	**116,940**	**115,844**	**114,517**
pron, WPR Press	7,606	42,784	6,812	38,137	7,760	43,533	8,048	45,002
pron, WPR Gen. Prose	18,946	45,699	18,643	44,622	16,222	39,211	17,863	42,993
pron, WPR Learned	3,925	24,358	4,177	25,957	3,861	24,097	3,697	22,997
pron, WPR Fiction	28,288	110,277	25,411	99,015	27,548	107,108	27,035	105,398
pron, WPR all genres	**58,765**	**58,182**	**55,043**	**54,285**	**55,391**	**54,875**	**56,643**	**55,994**
verb Press	29,430	165,544	28,766	161,049	30,569	171,491	30,351	169,713
verb Gen. Prose	69,341	167,255	68,689	164,408	67,071	162,119	66,338	159,663
verb Learned	24,880	154,405	25,615	159,178	25,395	158,494	23,695	147,396
verb Fiction	56,249	219,280	53,985	210,355	55,394	215,374	54,860	213,877
verb all genres	**179,900**	**178,117**	**177,055**	**174,616**	**178,429**	**176,768**	**175,244**	**173,237**
Misc Press	6,691	37,637	6,397	35,814	7,764	43,556	7,359	41,149
Misc Gen. Prose	15,520	37,435	13,995	33,497	15,863	38,343	15,172	36,516
Misc Learned	8,349	51,814	6,135	38,125	6,800	42,440	7,591	47,220
Misc Fiction	10,784	42,040	10,061	39,203	11,000	42,768	10,436	40,686
Misc all genres	**41,344**	**40,934**	**36,588**	**36,084**	**41,427**	**41,041**	**40,558**	**40,094**
TOTAL	**1,010,012**	**1,000,000**	**1,013,969**	**1,000,000**	**1,009,394**	**1,000,000**	**1,011,585**	**1,000,000**

(D) Genre-sensitive association plots for noun-phrase and verb-phrase tag groups

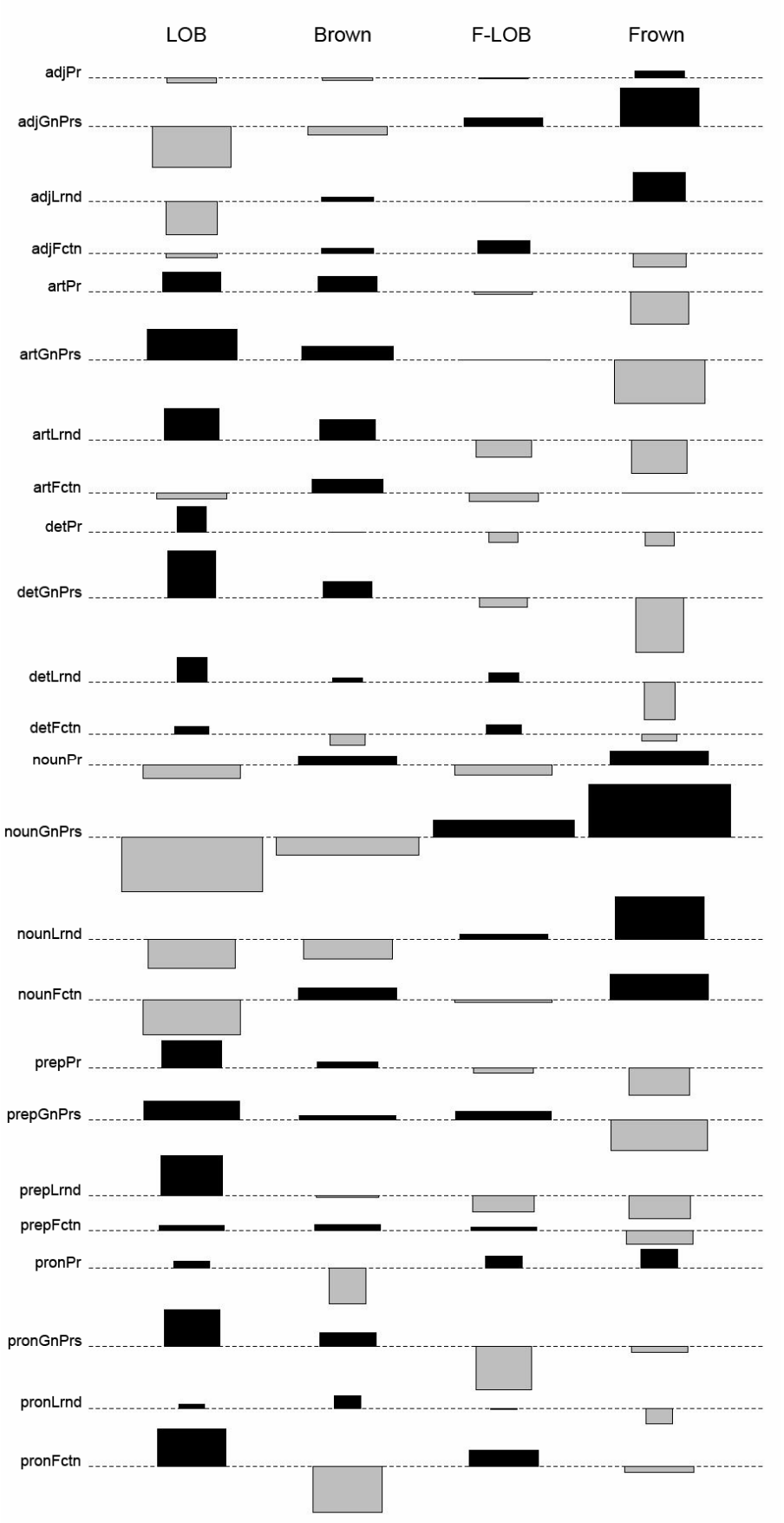

Fig. 7. Noun-phrase-word classes in four corpora, broken down into genres (deviations of observed frequencies from independence)

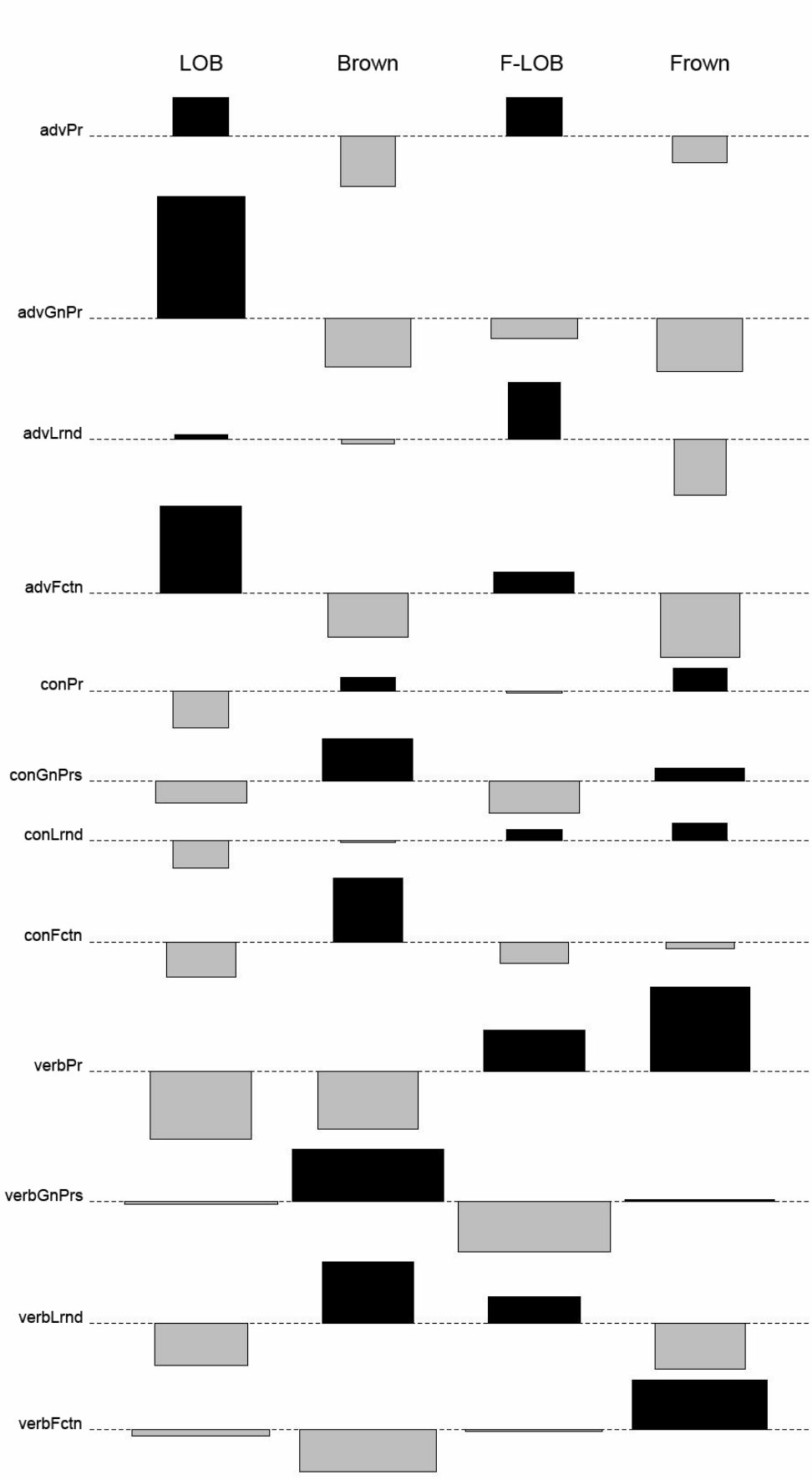

Fig. 8. Verb-phrase-word classes in four corpora, broken down into genres (deviations of observed frequencies from independence)

(E) Markup codes: original and revised representation in F-LOB and Frown

Original markup	Revised markup	Gloss
<#FLOB:([A-Z][0-9]{2})>	<text id=FLOB\1>	filename, e.g. FLOBA01
<p_>		paragraph open
		paragraph close
<h_>	<head>	heading open
<h/>	</head>	heading close
<h|>([w]+)	<head>\1</head>	one-word heading
<quote_>"	<quote>	quotation open
<quote_>([^"])	<quote>\1	quotation close
["]<quote/>	</quote>	quotation close
["]([^"])<quote/>	\1</quote>	quotation close
([^"])([^"])<quote/>	\1\2</quote>	quotation close
<quote|>"([^]+)"	<quote>\1</quote>	one-word quotation
<tf_>	<hi>	typeface shift open
<tf/>	</hi>	typeface shift close
<tf|>([^ ,.)(:]<?!]+)	<hi>\1</hi>	one-word typeface shift
<foreign_>	<foreign>	foreign phrase open
<foreign/>	</foreign>	foreign phrase close
<foreign|>([w]+)	<foreign>\1</foreign>	one-word foreign
<O_>caption&table<O/>	<gap dec="caption and table">	omitted visual material
<O_>diagram&caption<O/>	<gap dec="diagram and caption">	omitted visual material
<O_>figure&caption<O/>	<gap dec="figure and caption">	omitted visual material
<O_>figures&captions<O/>	<gap dec="figure and captions">	omitted visual material
<O_>formula&caption<O/>	<gap dec="formula and caption">	omitted visual material
<O_>graph&caption<O/>	<gap dec="graph and caption">	omitted visual material
<O_>graphs&captions<O/>	<gap dec="graph and captions">	omitted visual material
<O_>table&caption<O/>	<gap dec="table and caption">	omitted visual material
<O_>table&figure&captions<O/>	<gap dec="table and figure and captions">	omitted visual material
<O_>tables&caption<O/>	<gap dec="tables and caption">	omitted visual material
<O_>tables&captions<O/>	<gap dec="tables and captions">	omitted visual material
<O_>([^>]+)<O/>	<gap desc="\1">	other omitted material
<}_><-|>([^>]+)<+|>([^>]+)<}/>	<reg orig="\1">\2</reg>	spelling regularization
([w'-]+)<&_>sic!<&/>	<sic>\1</sic>	sic tag
([w'-]+)<&|>sic!	<sic>\1</sic>	one-word sic
<?_>-<?/>	&rehy;	ambiguous end-of-line hyphen
<TranslitG_>	<note desc="transliterated from Greek">	Greek text open
<TranslitG/>	</note>	Greek text close
<sp_>	<hi rend=hi>	superscript open
<sp/>	</hi>	superscript close
<sb_>	<hi rend=lo>	subscript open
<sb/>	</hi>	subscript close
<sb|>([w]+)	<hi rend=lo>\1</hi>	
<sp|>([w]+)	<hi rend=hi>\1</hi>	
<*_>([A-Za-z])-acute<*/>	&\1acute;	diacritic character
<*_>([A-Za-z])-cedille<*/>	&\1cedil;	diacritic character
<*_>([A-Za-z])-circ<*/>	&\1ring;	diacritic character
<*_>([A-Za-z])-circlet<*/>	&\1ring;	diacritic character
<*_>([A-Za-z])-grave<*/>	&\1grave;	diacritic character
<*_>([A-Za-z])-hacek<*/>	&\1caron;	diacritic character
<*_>([A-Za-z])-stroke<*/>	&\1macr;	diacritic character
<*_>([A-Za-z])-tilde<*/>	&\1tilde;	diacritic character
<*_>([A-Za-z])-umlaut<*/>	&\1uml;	diacritic character
<*_>([A-Za-z])-uml<*/>	&\1uml;	diacritic character
<*_>([A-Za-z])-tilde<*/>	&\1tilde;	diacritic character
<*_>square<*/>	•	bullet
<*_> square <*/>	•	bullet

<*_>black-square<*/>	<gap desc="black square">	omitted graphic character
<*_>black-triangle<*/>	<gap desc="black triangle">	omitted graphic character
<*_>black-circle<*/>	<gap desc="black circle">	omitted graphic character
<*_>circle<*/>	ˆ	omitted graphic character
<*_>bullet<*/>	•	omitted graphic character
<*_>dotted-line<*/>	<gap desc="dotted line">	omitted graphic character
<*_>dot<*/>	·	# ok?
<*_>star<*/>	*	asterisk
<*_> star <*/>	*	asterisk
<*_>square-root<*/>	√	square-root symbol
<*_>infinity<*/>	∞	infinity symbol
<*_>degree<*/>	°	degree symbol
<*_>approximate-sign<*/>	˜	approximation symbol (tilde)
<*_>section<*/>	§	section mark
<*_>arrow<*/>	→	right-arrow symbol
<*_>checkmark<*/>	✓	check symbol
<*_>plus-minus<*/>	±	plus-minus symbol
<*_>pound-sign<*/>	£	pound sterling symbol
pounds([0-9][0-9,]*)([.,)<-;:?!"/])	£\1\2	pound sterling symbol
pounds([0-9][0-9,.]*)(m)([.,)<-;:?!"])	£\1\2\3	pound sterling symbol
pounds([0-9][0-9,.]*)(bn)([.,)<-;:?!"])	£\1\2\3	pound sterling symbol
pounds([0-9][0-9,.]*)(million)([.,)<-;:?!"])	£\1\2\3	pound sterling symbol
pounds([0-9][0-9,.]*)(billion)([.,)<-;:?!"])	£\1\2\3	pound sterling symbol
pound1([ ])	&pound1;	pound sterling symbol
(s)& (S)	\1& \2	ampersand
(s)&$	\1&	ampersand
^&(s)	&\1	ampersand
-	—	long dash
-	—	long dash
•	•	bullet
/ (&unclass;)	/\1	slash (solidus)
#	♯	sharp symbol
<*_>ALPHA<*/>	&Agr;	Greek letter
<*_>BETA<*/>	&Bgr;	Greek letter
<*_>GAMMA<*/>	&Ggr;	Greek letter
<*_>DELTA<*/>	&Dgr;	Greek letter
<*_>EPSILON<*/>	&Egr;	Greek letter
<*_>ZETA<*/>	&Zgr;	Greek letter
<*_>ETA<*/>	&EEgr;	Greek letter
<*_>THETA<*/>	&THgr;	Greek letter
<*_>IOTA<*/>	&Igr;	Greek letter
<*_>KAPPA<*/>	&Kgr;	Greek letter
<*_>LAMBDA<*/>	&Lgr;	Greek letter
<*_>MU<*/>	&Mgr;	Greek letter
<*_>NU<*/>	&Ngr;	Greek letter
<*_>XI<*/>	&Xgr;	Greek letter
<*_>OMICRON<*/>	&Ogr;	Greek letter
<*_>PI<*/>	&Pgr;	Greek letter
<*_>RHO<*/>	&Rgr;	Greek letter
<*_>SIGMA<*/>	&Sgr;	Greek letter
<*_>TAU<*/>	&Tgr;	Greek letter

<*_>UPSILON<*/>	&Ugr;	Greek letter
<*_>PHI<*/>	&PHgr;	Greek letter
<*_>CHI<*/>	&KHgr;	Greek letter
<*_>PSI<*/>	&PSgr;	Greek letter
<*_>OMEGA<*/>	&OHgr;	Greek letter
<*_>alpha<*/>	&agr;	Greek letter
<*_>beta<*/>	&bgr;	Greek letter
<*_>gamma<*/>	&ggr;	Greek letter
<*_>delta<*/>	&dgr;	Greek letter
<*_>epsilon<*/>	&egr;	Greek letter
<*_>zeta<*/>	&zgr;	Greek letter
<*_>eta<*/>	&eegr;	Greek letter
<*_>theta<*/>	&thgr;	Greek letter
<*_>iota<*/>	&igr;	Greek letter
<*_>kappa<*/>	&kgr;	Greek letter
<*_>lambda<*/>	&lgr;	Greek letter
<*_>mu<*/>	&mgr;	Greek letter
<*_>nu<*/>	&ngr;	Greek letter
<*_>xi<*/>	&xgr;	Greek letter
<*_>omicron<*/>	&ogr;	Greek letter
<*_>pi<*/>	&pgr;	Greek letter
<*_>rho<*/>	&rgr;	Greek letter
<*_>sigma<*/>	&sgr;	Greek letter
<*_>tau<*/>	&tgr;	Greek letter
<*_>upsilon<*/>	&ugr;	Greek letter
<*_>phi<*/>	&phgr;	Greek letter
<*_>chi<*/>	&khgr;	Greek letter
<*_>psi<*/>	&psgr;	Greek letter
<*_>omega<*/>	&ohgr;	Greek letter
<*_>unch<*/>	&unclass;	
<*_>unches<*/>	&unclass;	
<*_>([a-z])-([A-Za-z]+)<*/>	&\1\2;	
<*_>([A-Za-z]{2})-ligature<*/>	&\1lig;	diacritic character

5 Acknowledgments

Funding from Deutsche Forschungsgemeinschaft (DFG), Bonn/Germany, which made this project possible, is gratefully acknowledged. In Freiburg, the following individuals have worked at different stages on the compilation and post-editing of POS-tagged F-LOB and Frown: Franziska Becker, Lucas Champollion, Septimius Fericean, Heike Fiedler, Ulf Gerdelmann, Lars Hinrichs, Marianne Hundt, Matthias Kaufmann, Tobias Maier, Christian Mair, Michael Percillier, Stefanie Rapp, Andrea Sand, Silke Scheible, Birgit Waibel, Antonia Walker, Lisa-Maria Wild.

It is gratefully recorded that the research carried out at Lancaster by Geoffrey Leech and Nicholas Smith in connection with the automatic tagging of the four corpora was supported by grants from the Arts and Humanities Research Board, the British Academy, and the Leverhulme Trust. Mike Pacey contributed greatly to developing the Template Tagger software, which was instrumental in implementing the C8-tagging refinements described in this manual.

References

Bauer, L. 1993. *Manual of Information to Accompany the Wellington Corpus of Written New Zealand English*. Wellington: Department of Linguistics, U of Wellington.

Biber, D. 2003. "Compressed noun-phrase structure in newspaper discourse: the competing demands of popularization vs. economy." *New Media Language*, J. Aitchison and D.M. Lewis (eds.), 169-181. London/New York: Longman.

Collins, P. and Peters, P. 1988. "The Australian corpus project." *Corpus Linguistics, Hard and Soft*, M. Kytö, O. Ihalainen and M. Rissanen (eds.), 103-120. Amsterdam: Rodopi.

Dickinson, M. and Meurers, W.D. 2003. "Detecting errors in part-of-speech annotation." *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics* Retrieved 10 October, 2006. From <http://ling.osu.edu/dm/papers/dickinson-meurers-03.html>.

Francis, N. and Kučera, H. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Francis, W.N. 1965. "A standard corpus of edited present-day American English." *College English* 26: 267-273.

Francis, W.N. and Kučera, H. 1979. "Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with digital computers." From <http://khnt.hit.uib.no/icame/manuals/brown/>.

Hinrichs, L. and Szmrecsanyi, B. 2007. "Recent changes in the function and frequency of standard English genitive constructions: a multivariate analysis of tagged corpora." *English Language and Linguistics* 11 (3): 437-474.

ICAME 1999. ICAME CD-ROM Version 2, 1999. K. Hofland. Bergen, International Computer Archive of Modern and Medieval English.

Johansson, S. and Hofland, K. 1989. *Frequency Analysis of English Vocabulary and Grammar. Based on the LOB Corpus*. Oxford: Clarendon.

Leech, G. and Smith, N. 2005. "Extending the possibilities of corpus-based research on English in the twentieth century: A prequel to LOB and F-LOB." *ICAME Journal* 29: 83-98.

Mair, C. 2006. *Twentieth-Century English. History, Variation, and Standardization*. Cambridge: Cambridge UP.

Mair, C., Hundt, M., Leech, G. and Smith, N. 2002. "Short-term diachronic shifts in part-of-speech frequencies: A comparison of the tagged LOB and F-LOB corpora." *International Journal of Corpus Linguistics* 7: 245-264.

Sand, A. and Siemund, R. 1992. "LOB - 30 years on..." *ICAME Journal* 16: 119-122.

Shastri, S.V. 1988. "The Kolhapur Corpus of Indian English and work done on its basis so far." *ICAME Journal* 12 (15-26).

Zwicky, A. 1987. "Suppressing the Zs." *Journal of Linguistics* 23 (1): 133-148.