Probabilistic determinants of genitive variation in spoken and written English

A multivariate comparison across time, space, and genres

Benedikt Szmrecsanyi[*] and Lars Hinrichs[†]
[*]University of Freiburg, [†]University of Texas at Austin

This is a paper about language variation and about language change, investigating the competition between the *s*-genitive and the *of*-genitive in Modern English (written and spoken, British and American) as a case study. Drawing on a range of spoken and written corpora and considering a multivariate envelope of seven major conditioning factors (such as possessor animacy and end-weight), we seek to uncover, first, how the probabilistic preferences of British and American journalists might have changed between the 1960s and 1990s, and, second, how such changes in written English relate to the way *speakers* of English choose between the two genitives. We find that the *s*-genitive is comparatively frequent in both spoken English and contemporary journalistic English thanks to quite different reasons, and that the recent spread of the *s*-genitive in press English is due to a process of economization rather than colloquialization.

1. Introduction[1]

It is fair to say that there is a rich and insightful literature on the competition between the two genitives in English, which is to no small part due to the curious historical fact that while the *of*-genitive (also known as the 'Norman genitive', 'periphrastic genitive', or '*of*-construction'), as in (1), is the long-term incoming form, the *s*-genitive (also known as the 'Saxon genitive'), as in (2), has bounced

[1] Funding from Deutsche Forschungsgemeinschaft (grants no. MA 1652/3-1 and MA 1652/3–2) is gratefully acknowledged. The audience of our talk at the ICAME 27 conference in Helsinki in May 2006 provided very helpful advice on an earlier version of this paper. We also benefited greatly and continuously from discussions with Christian Mair.

back during the Modern English period and has been shown to be spreading right now, especially in press language (for instance, Potter 1969: 105–106; Dahl 1971: 141; Raab-Fischer 1995).

(1) ... this session is helpful for all of us in that it forces us to rethink, to problematize, and to interrogate *the history of American anthropology* ... <CSAE 1034>

(2) While *anthropology's history* is indeed implicated in the scientific construction as race as a biological fact ... <CSAE 1034>

In modern English, the two genitives and in particular the *s*-genitive encode a "grab-bag" (Givón 1993: 264) of semantic and pragmatic relations – for the *s*-genitive alone, Quirk et al. (1985: 321–322) list eight different meanings – yet it is well-known that the two genitives are fairly interchangeable in a considerable number of contexts (cf. Quirk et al. 1985: 321; Jucker 1993: 121); for example, *anthropology's history* (cf. (1)) and *the history of anthropology* are close periphrases and can certainly be considered semantically (near-)equivalent. It is such choice contexts that will take center stage in the present study.

In those contexts where language users can choose between the *s*-genitive and the *of*-genitive, which determinants, or factors, influence the choice? The multiple and regularly conflicting determinants suggested in the literature fall into five major factor groups:

(i) *Semantic and pragmatic factors*. Animate possessors prefer the *s*-genitive, inanimate possessor prefer the *of*-genitive (for instance, Altenberg 1982: 117–148).[2] Further, it has been shown that increased thematicity of the possessor NP makes usage of the *s*-genitive significantly more likely (cf. Osselton 1988). As for higher-level pragmatic factors, givenness of the possessor NP has been claimed to favor the *s*-genitive because the *s*-genitive places the discourse-old possessor first (Quirk et al. 1985: 1282; Biber et al. 1999: 305), but a number of recent corpus studies (for instance, Gries 2002; Hinrichs & Szmrecsanyi 2007) have failed to obtain empirical evidence for this effect.

(ii) *Phonology*. A final sibilant in the possessor NP (for instance, in a plural morpheme) discourages usage of the *s*-genitive (for instance, Altenberg 1982).

(iii) *Factors related to processing and parsing*. In accordance with the principle of end-weight (Behaghel 1909/1910), which according to some authors ultimately

[2] Hawkins (1994: 424) has suggested that animacy might be, in general, epiphenomenal to end-weight (see below) because animate NPs are typically shorter than inanimate NPs; Rosenbach (2005), however, demonstrates that in the case of genitive variation, animacy of the possessor NP is an independent factor and not eliminable.

boils down to "processability" (cf. Kreyer 2003) and parsing efficiency (cf. Hawkins 1994), heavier possessor NPs prefer the *of*-genitive (because the *of*-genitive places the possessor second) while heavier possessums prefer the *s*-genitive (for instance, Quirk et al. 1985: 1282; Biber et al. 1999: 304). It is also known that language users tend to recycle material that they have used or heard before (a phenomenon which is often psycholinguistically motivated): thus, for example, precedence of either genitive construction in discourse (be it written or spoken) increases the odds that the same genitive type will be used next time there is a choice (Szmrecsanyi 2006: 87–107).

(iv) *Economy-related factors.* The bulk of scholarship in this vein has claimed that by virtue of being "more compact" (Biber et al. 1999: 300), the *s*-genitive seems to be preferred in contexts and registers where the "tendency to brevity" (Dahl 1971: 172) is paramount. In this spirit, Hinrichs and Szmrecsanyi (2007) present corpus evidence that journalists favor the *s*-genitive in contexts characterized by comparatively high informational/lexical density.

(v) *External factors*. The more informal the setting, the greater the preference for the *s*-genitive (for instance, Altenberg 1982: 284); because most spoken text types are arguably more informal than most written text types (all other things being equal), we thus expect, as a rule of thumb, the *s*-genitive to be particularly frequent in spoken data (cf. Rosenbach 2002: 39 for a similar argument). In terms of geographic differences, the *s*-genitive is known to occur more frequently in American English than in British English (cf., for example, Rosenbach 2003: 395–396).

How can the present study complement the sizable body of research on English genitive constructions? Given that the *s*-genitive is spreading in Present-Day journalistic English (for instance, Raab-Fischer 1995), the question addressed in the present study is whether or not this shift is due to an increasing importance of those conditioning factors thanks to which the *s*-genitive, as we will see, is overall more frequent in spoken than in written English. In other words, we are interested in the interplay between the internal factors in (i)–(iv) and the external factor 'genre' in an ongoing process of language change. So, as far as the spread of the *s*-genitive is concerned, is it true that we are seeing a "colloquialization of the norms of written English," as Leech & Smith argue (2006, using the terminology of Hundt & Mair 1999)? Or are other forces responsible – for instance, a process of 'economization,' which according to Biber (2003) is especially potent in newspaper language?

On the methodological plane, we depart from the usual practice of much previous corpus-based research to simply crosstabulate raw frequencies. Rather, to explore the determinants of genitive choice in the Brown family of corpora vis-à-vis two spoken English corpora (the Corpus of Spoken American English and the Freiburg Corpus of English Dialects), the present study will offer a fine-grained multivariate analysis of the gradient and probabilistic constraints on genitive choice (in the spirit of, e.g., Bresnan & Hay forthcoming) across different varieties, sampling times, and genres.

2. Data

The present study relies on a quantitative analysis of the following databases:

- *The Corpus of Spoken American English* (CSAE). The release that will be used here is composed of the installments 1 and 2 (Du Bois et al. 2000; Du Bois et al. 2003), spanning in all 41 conversations, each approximately 20–30 minutes in length. Designed primarily for conversation analytic purposes and thus sampling very conversational, unscripted and hence very informal American English, this corpus is a comparatively small one (roughly 166,000 words of running text), though it is large enough for some of the purposes of the present study.
- *The Freiburg Corpus of English Dialects* (FRED). This corpus (cf. Hernández 2006) contains samples (mainly transcribed so-called 'oral history' material) of dialectal speech from a variety of sources. The bulk of these samples were recorded between 1970 and 1990; in most cases, a fieldworker interviews an informant about life, work etc. in former days. The informants are typically elderly people with a working-class background. Speech styles are relatively formal due to the interview situation. The subsample of FRED to be analyzed here spans ca. 1,300,000 words; dialect areas included in the sample are the Hebrides, the Midlands, the North of England, Wales, the Southwest, and the Southeast (the exact composition is not of interest here, as this is not a study in dialectology). What is important is that in comparison to the CSAE, FRED samples spoken English that is at once (i) somewhat less informal, (ii) more British, and (iii) more traditional, thanks to the fact that informants were typically raised before World War I.
- *The A and B sections in the Brown family of corpora* (Brown, LOB, Frown, and F-LOB). These four corpora contain written, edited, and published Standard English. The two older corpora, Brown and LOB, represent, respectively, American and British English from the 1960s, whereas Frown and F-LOB are their 1990s updates. Thus, the quartet covers two varieties and a time span of 30 years. The corpora are all structured according to a set framework of fifteen different genre categories. In total, each corpus contains 500 text samples; at a

sample size of about 2,000 words each, the four Brown corpora thus contain a structured dataset of four million words of running text. We chose to focus on journalistic language and therefore selected the categories 'Reportage' (A) and 'Editorial' (B) from each of the corpora, amounting to 71 samples, or roughly 142,000 words, per corpus, adding up to a total of ~568,000 words. For this study, we used the recently completed part-of-speech-tagged versions of the corpora (on which see Leech & Smith 2005; Hinrichs forthcoming).

3. The linguistic variable

All instances of interchangeable *s*- and *of*-genitives were manually extracted from the above corpora, i.e. each instance of an *s*- or *of*-genitive was classified according to whether the alternative construction could have been used in its place. Altogether, this yielded a data set of *N* = 10,450 interchangeable genitives (CSAE: *N* = 332; FRED: *N* = 1,818; Brown: *N* = 2,204; LOB: *N* = 2,019; Frown: *N* = 2,132; F-LOB: *N* = 1,945).

We retained only those instances of the *s*-genitive which could equally have been expressed as an *of*-genitive by applying a simple conversion rule, without adding or deleting any of the lexemes in the possessor or possessum phrase (except for the optional addition of a determiner to the possessum). Similarly, only those *of*-genitive tokens were retained which could have been expressed using an *s*-genitive construction instead with neither of the noun phrases modified, except for the necessary deletion of any determiner in the possessum phrase. Crucially, the alternative construction would have to leave the meaning of the actual choice unchanged; thus, *the city of Atlanta* was not considered an interchangeable genitive because the alternative, *Atlanta's city*, has a different meaning. A negative list of non-interchangeable genitive types – roughly following the similar lists in Kreyer (2003: 170) and Rosenbach (2006: 622–623) – guided the coders' judgments of interchangeability. While *s*-genitives proved to be interchangeable in the great majority of cases, the following were excluded from the analysis:

(i) any construction in which a noun marked with a genitive *s* is not followed by an explicit possessum phrase (as in *breakfast at Tiffany's*);
(ii) any phrase that has been conventionalized with the *s*-genitive, so that the *of*-genitive is no longer a possible alternative (as in *Murphy's law*);
(iii) 'descriptive genitives' *(men's suits, bird's nest)*, which frequently form an idiomatic unit;

(iv) titles of books, films, works of art, etc. that are premodified with a genitive possessor phrase denoting their creator (as in *John Steinbeck's* Of Mice and Men).

The types of *of*-genitives that were not coded as interchangeable included the following:

(i) *of*-genitive constructions with a possessum that could not possibly be read as definite, since the *s*-genitive always expresses the possessum as definite (as in *a major strategy of his administration*);
(ii) most of those *of*-genitives containing a possessor noun phrase that shows postmodification, since the result of a conversion to the *s*-genitive would be a 'group genitive' (as *in the man I met's girlfriend*);
(iii) measures expressed as *of*-constructions (as in *three liters of beer*);
(iv) as with *s*-genitives, any phrase that has been conventionalized with an *of*-genitive (as in *the President of the United States*).

To establish interrater reliability of our coding scheme, we had a number of random genitive samples independently coded for interchangeability by different coders, who received coder training. After a number of trials on different samples and a series of subsequent refinements to the coding scheme, parallel annotation of a set of $N = 202$ genitives yielded (i) a simple agreement rate of 86% and a "good" (cf. Orwin 1994: 152) Cohen's κ value of .69 for *s*-genitives, and (ii) a simple agreement rate of 89% and an "excellent" Cohen's κ value of .78 for *of*-genitives.

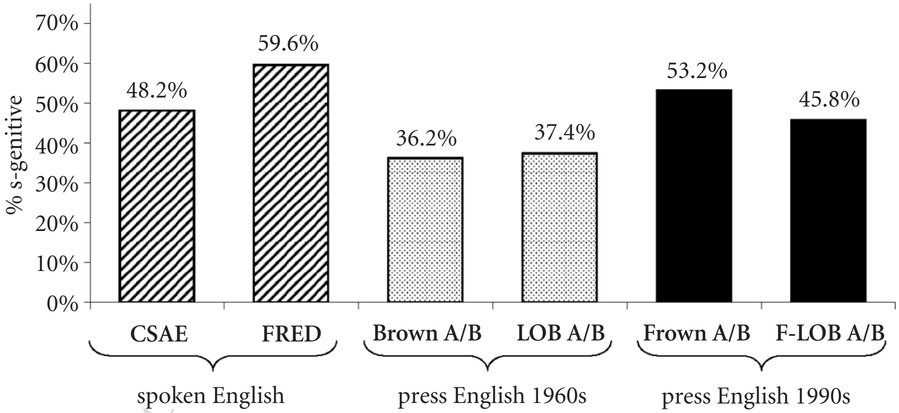

Figure 1. Share of the *s*-genitive among interchangeable genitives across corpora

4. Overall distribution of genitives

Let us start our empirical discussion with a comparatively simple frequency analysis; Figure 1 plots the share of the *s*-genitive of all interchangeable genitives across corpora. This exercise yields four major observations:

1. The *s*-genitive is, on the whole, more frequent in spoken data than in written data, which is in accordance with our working hypothesis: FRED exhibits the highest percentage of the *s*-genitive (59.6%), Brown (36.2%) the lowest. The notable exception to this written-spoken stratification is that the *s*-genitive is more frequent in Frown (53.2%), which samples written American English journalistic prose, than in the CSAE (48.2%), which samples conversational American English.
2. The *s*-genitive is significantly (p < .05) more frequent in FRED, which samples traditional dialects of British English, than in the CSAE, which samples contemporary conversational American English. This result is unexpected, assuming – as we did – that the register sampled in FRED (interviews) is more formal than that sampled in the CSAE (conversation), geographic and/or apparent time differences between the two corpora appear to override the register difference.
3. A look at the datasets from the four Brown corpora shows that the *s*-genitive has become significantly (p < .001) more frequent in press language in the period between the 1960s and the 1990s, which supports claims (for instance, Raab-Fischer 1995) that the *s*-genitive is spreading in real time.
4. As far as contemporary press English – as sampled in Frown and F-LOB – is concerned, the *s*-genitive is significantly (p < 0.01) more frequent in the American data (53.2%) than in the British data (45.8%). This is in accordance with previous claims (cf., for example, Rosenbach 2003: 395–396) that the *s*-genitive is overall more frequent in American English than in British English.

In a nutshell, these frequency figures – one-dimensional as they are – seem to suggest that with regard to genitive frequencies, press language has over time become more similar to spoken varieties of English. In other words, Figure 1 might be interpreted as evidence for a colloquialization of the written norm (Leech & Smith 2006). Our subsequent, more detailed analyses of the conditioning factors working on genitive choice will seek to explore whether such an interpretation is indeed justified.

5. Conditioning factors in genitive choice

Conditioning factors considered in the present study fall into four groups: (i) semantic and pragmatic factors (animacy and thematicity of the possessor),[3] (ii) phonology (i.e. presence of a final sibilant in the possessor), (iii) parsing and processing factors (possessor length, possessum length, and precedence of an identical genitive construction), and (iv) economy (i.e. type-token ratio of a given genitive passage).

5.1 Animacy

A survey of the literature reveals that animacy of the possessor NP is one of the most important determinants of genitive choice. Thus, adopting Rosenbach's (2006: 105) animacy hierarchy (human > animal > collective > inanimate) and drawing on Zaenen et al.'s (2004) general coding scheme for animacy, we sought to operationalize the factor 'animacy' by manually coding each possessor NP in our database according to the following four-way classification: (i) animate possessor NPs, as in (3); (ii) animal possessor NPs, as in (4); (iii) collective possessor NPs, as in (5); and (iv) inanimate possessor NPs, as in (6).

(3) *the emperor*'s family had to call off plans … <Frown A04>

(4) and he'd pick me up and show me, you know, *a little bird*'s eggs … <FRED DEN_001>

(5) Would that the odious discriminatory policy of *the Pentagon* were limited to those two instances. <F-LOB B27>

(6) and it was like on the back bumper of *the Honda*, too. <CSAE 0513>

Interrater reliability of the manual coding procedure was satisfactory: parallel coding of a random subset of $N = 199$ genitive possessors by two trained coders yielded a simple agreement rate of ca. 86% and an "excellent" (cf. Orwin 1994: 152) Cohen's κ value of ca. 0.79.

[3] Information status (i.e., givenness of the possessor or possessum) will not be considered in this study. The reason is that this factor is, across the board, not significant – a finding which dovetails with some previous research on genitive choice (Gries 2002; Szmrecsanyi 2006: 97–107; Hinrichs & Szmrecsanyi 2007).

5.2 Thematicity of the possessor NP

According to Osselton (1988), it is the general topic of a text which determines which nouns in that text can take the *s*-genitive. Thus, according to Osselton, while *sound, soil,* and *fund* will not normally take the *s*-genitive, "in a book on phonetics, *sound* will get its genitive, in one on farming, *soil* will do so, and in a book on economics you can expect to find *a fund's success*" (Osselton 1988: 143). Thus, assuming that increased text frequency of a possessor NP would make the *s*-genitive more likely, we had Practical Extraction and Report Language (Perl) scripts establish, for every individual possessor NP in our database, the *log* text frequency[4] of the possessor NP's head noun in the respective corpus text (measured in frequency per 2,000 words, which is the standard size of texts in the Brown family). Let us illustrate the procedure with the example in (7):

(7) *The bill's supporters* said they still expected Senate approval …
<Frown A02>

In (7), the possessor NP's head noun is *bill*, and *bill* has a text frequency of 32 occurrences in Frown text A02 (which spans about 2,000 words).

5.3 Final sibilants in the possessor NP

Previous scholarship has shown that the presence of a final sibilant in the possessor, as in (8), may discourage the use of the *s*-genitive (cf. Altenberg 1982):

(8) But that is *the sad and angry side of Bush.* <Frown A11>

We operationalized this phonological constraint by having a Perl script identify all possessors that end, orthographically, in <s> (as in *Congress*), <z> (as in *jazz*), <ce> (as in *resistance*), <sh> (as in *Bush*), or <tch> (as in *match*). Possessors ending in <dge> (as in *judge*) are so rare that they were excluded from analysis.

5.4 End weight: Possessor and possessum length

According to the time-honored principle of 'end-weight' (for instance, Behaghel 1909/1910; Wasow 2002), language users tend to place 'heavier,' more complex constituents after shorter ones. Therefore, if the possessor is heavy, there should be a general preference for the *of*-genitive because it places the possessor last; if

[4] Text frequency was modeled logarithmically in order to alleviate the effect of frequency outliers.

the possessum is heavy, we expect a general preference for the *s*-genitive because it places the possessum last. For the purposes of the present study, we sought to approximate the weight of genitive constituents by determining their length in graphemic words, utilizing Perl scripts for automatic coding. For illustration, consider (9):

(9) Latter domain, under *the guidance of Chef Tom Yokel*, will specialize in steaks, chops, chicken and prime beef as well as Tom's favorite dish, stuffed shrimp. <Brown A31>

The possessor phrase in (9) commands three words (*Chef Tom Yokel*) while the possessum spans two words (*the guidance*). Note, however, that if the writer had opted for an *s*-genitive instead, the possessum phrase could not have been determined by an article (**Chef Tom Yokel's the guidance*). Therefore, definite or indefinite articles determining the possessum phrase of an *of*-genitive were not included in the tally (cf. Altenberg 1982: 79–84 for a similar coding procedure); net possessum length of the possessum phrase in (9) is thus exactly one word (*guidance*).

5.5 Persistence

We now move on to a further processing-related constraint on genitive choice, *viz.* precedence of an identical genitive construction in the preceding textual discourse. The basic idea is that usage of, say, an *s*-genitive in a given genitive slot increases the odds that the speaker/writer will use an *s*-genitive again next time she has a choice (Szmrecsanyi 2006: 87–107). Once again we relied on Perl scripts to establish, for each genitive occurrence in our database, whether an *s*-genitive had been used last time there was a genitive choice. (10) exemplifies a context where two subsequent interchangeable genitive contexts (*the continent's river systems* and *the country's Medical Association*) both exhibit *s*-genitives:

(10) … *the continent's river systems* are now infected, making the spread of the disease extremely difficult to control. In Ecuador, *the country's Medical Association* said 100 people had died of a total of 5,000 cases… <F-LOB A14>

5.6 Lexical density and type-token ratios

Hinrichs & Szmrecsanyi (2007) demonstrate that the *s*-genitive is preferred by writers in contexts where informational density is high, i.e. when there is a need to economically code more information in a given textual passage; this is because the *s*-genitive is the more compact and economic coding option of the two (Biber et al. 1999: 99). To check on this factor, we utilized Perl scripts to establish the type-token ratios of the textual passages where the genitive occurrences in our database are embedded; 'textual passage' refers to a context of 50 words before and 50 words after a given genitive construction.

6. Results

6.1 Logistic regression model for individual factors

We will now draw on *binary logistic regression* (cf. Pampel 2000 for an introduction) to quantify the combined contribution of the conditioning factors presented above. As a multivariate procedure, logistic regression integrates probabilistic statements into the description of performance and is applicable "wherever a choice can be perceived as having been made in the course of linguistic performance" (Sankoff & Labov 1979: 151). Predicting a binary outcome (i.e. a linguistic choice, in the case of the present study the choice between an *s*-genitive and an *of*-genitive) given several independent factors (or: predictors), a logistic regression model relies on the following key measures:

– *The magnitude and the direction of the influence of each predictor on the outcome.* This information is provided by *odds ratios*, indicating how the presence or absence of a feature (for categorical factors) or how a one-unit increase in a scalar factor probabilistically influences the odds that some outcome will occur. Odds ratios can take values between 0 and ∞; the more the figures exceed 1, the more highly the effect favors a certain outcome; the closer they are to zero, the more disfavoring the effect (if smaller than 1).
– *Variance explained by (or: explanatory power of) the model as a whole (R^2).* The R^2 value can range between 0 and 1 and indicates the proportion of variance in the dependent variable (i.e. in the outcomes) accounted for by all the factors included in the model. Bigger R^2 values mean that more variance is accounted for by the model. The specific R^2 measure which is going to be reported is the so-called *Nagelkerke R^2*, a pseudo R^2 statistic for logistic regression.
– *Predictive efficiency of the model as a whole.* The percentage of correctly predicted cases vis-à-vis the baseline prediction (*% correct (baseline)*) indicates how accurate the model is in predicting actual outcomes. The higher this percentage, the better the model.

Table 1. Odds ratios in logistic regression. Predicted odds are for the *s*-genitive.

	CSAE		FRED		Brown A/B		LOB A/B		Frown A/B		F-LOB A/B	
possessor animacy												
human	**8.08**	***	**69.66**	***	**9.83**	***	**13.00**	***	**7.08**	***	**14.19**	***
animal	**30.94**	***	**17.75**	***	(n.s.)		(n.s.)		(n.s.)		(n.s.)	
collective	**3.94**	**	**2.77**	**	**3.46**	***	**3.23**	***	**3.11**	***	**5.03**	***
thematicity of p'or	(n.s.)		(n.s.)		**1.30**	***	**1.23**	***	**2.06**	***	**1.63**	***
final sibilant in p'or	**.21**	***	**.36**	***	**.26**	***	**.49**	***	**.22**	***	**.29**	***
possessor length	**.52**	***	**.42**	***	**.40**	***	**.48**	***	**.47**	***	**.41**	***
possessum length	(n.s.)		(n.s.)		**1.30**	***	(n.s.)		**1.51**	***	**1.68**	***
persistence	**3.53**	***	**1.87**	***	**1.67**	***	**1.72**	***	**1.35**	**	**1.29**	*
type-token ratio	(n.s.)		(n.s.)		**2.23**	***	**2.32**	***	**1.86**	***	**1.93**	***
N	332		1,818		2,204		2,019		2,132		1,945	
% baseline	52.0		59.6		63.9		62.9		53.1		54.0	
% correct	70.4		88.8		77.1		76.3		77.1		80.2	
Nagelkerke R^2	.43		.68		.45		.38		.45		.52	

(n.s.) not significant, * significant at $p < 0.05$, ** significant at $p < 0.01$, *** significant at $p < 0.005$

Table 1 displays six logistic regression models, one for each of the corpora subject to analysis. Predictive efficiency is satisfactory: given the conditioning factors considered, the models predict between 70.4% (CSAE) and 88.8% (FRED) of the genitive outcomes accurately. Variance explained ranges between .38 (LOB A/B) and .68 (FRED), which is another way of saying that we can 'explain' between 38% and 68% of the observable variance in the corpora under analysis – the rest of the variance may be due to free variation, or there may exist other conditioning factors not considered in the present study. In all, the system of genitive choice sketched in Table 1 works best for the very traditional dialect speech sampled in FRED, and least well (though still quite satisfactorily) for 1960s British English newspaper language, as sampled in LOB A/B.

What about the effect of individual conditioning factors in the model? First, consider *possessor animacy*: with the exception of animal possessors in the Brown family – animal possessors are a rare species in press coverage and editorials – this factor group is significant throughout and exhibits the theoretically expected effect direction. In the CSAE, for instance, a human possessor, compared to an inanimate possessor (which constitutes the baseline condition in Table 1), increases the odds for an *s*-genitive by a factor of 8.08; an animal possessor increases the odds for an *s*-genitive by a factor of 30.94; and a collective possessor increases the odds for an *s*-genitive by a factor of 3.94.[5] In all, it can be seen that the effect strength of animacy is quite similar in the CSAE and in the Brown family. Having said that, it should be noted that human possessors attract *s*-genitives significantly ($p < .005$) more strongly in the two British press corpora (LOB/F-LOB) than in their American counterparts (Brown/Frown). FRED, then, is truly outstanding among the corpora under analysis: the huge odds ratio of 69.66 associated with human possessors indicates that in traditional British dialects, human possessors – for all intents and purposes – categorically trigger the *s*-genitive. This result is, we believe, unlikely to be due to, e.g., the text type (interviews) sampled in FRED – instead, we would like to claim that what we are seeing here is an older system of genitive choice, given that informants in FRED are elderly people and that many of the traditional dialects sampled in the corpus are rather conservative. Be that as it may, it is not evident that in terms of the importance of possessor animacy Frown and F-LOB are drifting towards the spoken corpora in our database when compared to Brown and LOB.

We had hypothesized that increased thematicity of the possessor – operationalized as the possessor head noun's *log* text frequency in a given corpus text – would make the *s*-genitive more likely, all other things being equal. This hypothesis is indeed borne out in the Brown family of corpora: in Brown, for example, every one-unit increase in a possessor head's *log* text frequency (to illustrate, this would correspond to a frequency differential of, very roughly, 3 occurrences instead of 1 occurrence per corpus text) increases the odds for the s-genitive by a factor of 1.3. Overall, (i) the impact of the factor is strongest in Frown and weakest in LOB, (ii) it is significantly (p < .01) more powerful in the American data than in the British data, and (iii) it is significantly (p < .005) stronger in the 1990s data than in the 1960s data. This longitudinal drift in the written data sources notwithstanding, the predictor is not even selected as significant in the spoken corpora (CSAE and FRED). Hence the written norm and the spoken norm are actually diverging in terms of the effect of the thematicity of the possessor.

A final sibilant in the possessor significantly and reliably discourages usage of the *s*-genitive, as expected: in the CSAE, where the predictor is most muscular, the presence of a final sibilant decreases the odds for an *s*-genitive by 79% (.21). The effect in the written data sources is, on average, weaker than in the spoken data sources, though interestingly the constraint has grown significantly more influential over time in press language. The somewhat ironical fact that a *phonological* constraint should become more influential in press language – a *written*

[5] For reasons of space, we dispense here with a discussion of confidence intervals associated with regression weights.

genre – over time can indeed only be interpreted, we believe, in terms of a colloquialization of the written norm.

As hypothesized, long possessor phrases significantly encourage the *of*-genitive (because this coding option places the possessor second). The effect is strongest in Brown, where every additional word in the possessor NP decreases the odds for an *s*-genitive by 60% (.40), and weakest in the CSAE, where the corresponding figure is 48% (.52). On the whole, differences along the lines of the spoken-written dimension are not obvious here. Further, we also hypothesized that long possessums would favor the *s*-genitive (because the *s*-genitive places the long possessum last). And indeed, where the constraint is significant – in Brown, Frown, and F-LOB – the effect runs in the theoretically expected direction, with odds ratios ranging between 1.30 (Brown) and 1.68 (F-LOB). Crucially, though, the constraint does not seem to be important in any of the spoken corpora, which is another way of saying that possessum length is one of the factors that really make a difference between the spoken and written English system of genitive choice.[6]

What about persistence effects? The predictor is significant throughout, and strongest in the CSAE where precedence of an *s*-genitive in the ongoing discourse increases the odds for another, subsequent *s*-genitive by a factor of 3.53, or 253%; the corresponding figure in F-LOB, where the effect is weakest, is a mere 29% (1.29). In all, it is fairly evident that persistence effects are more important in the spoken data sources than in the written data sources, which hardly comes as a surprise given the effect's deep rootedness in the nature of online processing constraints (Szmrecsanyi 2006). Observe that there is no evidence that in press language the predictor has over time approximated its effect strength in spoken language.

As for lexical density, we assumed that speakers/writers would resort to the more economical *s*-genitive in contexts characterized by high type-token ratios. For writers, this assumption holds true: for every 10-word increase in a given genitive context's type-token ratio (if, say, such a context contains 70 different types, instead of just 60), the odds for an *s*-genitive increase by a factor of between 1.86 (Frown) and 2.32 (LOB). There is a slight but insignificant indication that the predictor has become less important in press language over time. In the spoken corpora, the predictor is not selected as significant, thus the sort of economy implicit in the nature of the predictor appears not to be important in spoken language.[7]

[6] We also tested for interaction effects between possessor length and possessum length: more often than not, such interaction effects were not selected as significant, and even when they were significant, they did not add substantially to the model's overall explanatory power.

[7] We caution readers that these findings for press language should not be taken as representative of all written language; indeed any changes in textual conventions and probabilistic grammar from the 1960s to the 1990s that relate to economy should be seen as indicative

6.2 Multidimensional scaling

So far, we have characterized the mechanism of genitive choice in spoken and written English on the basis of a complex landscape of conditioning factors, yielding six sets (one for each corpus under analysis) of nine discrete odds ratios that characterize this landscape. We have seen that only in the case of the inhibiting effect of final sibilants on usage of the *s*-genitive we are truly seeing a convergence of spoken English and press language. At the same time, we have also discovered some evidence for patterns of divergence – for instance, in terms of the role that thematic possessors play.

Note, now, that comprehensive and fine-grained as the analysis of conditioning factors may be, it is also rather complicated, thanks to its multidimensional nature. This is why we will now endeavor to uncover the 'big' picture of genitive choice, utilizing *multidimensional scaling* (henceforth: MDS), a set of statistical techniques designed to uncover the 'hidden structure' in multidimensional variance (for an introduction to MDS, see Kruskal & Wish 1978); in the linguistic context, MDS is popular in, e.g., dialectometry (see, for instance, Nerbonne et al. 1999). In our case, we will draw on MDS to scale down the 9 original dimensions (each corresponding to one odds ratio in Table 1) by which genitive choice in each corpus in our database is defined. This will enable us to visualize the (dis-)similarities between the corpora in a two-dimensional map. The bonus of this procedure is that because such a map uses the straightforward concepts of space and distance, it can be interpreted fairly intuitively: much as with geographic maps, the further two sampling points are apart, the more dissimilar they are; if two pairs of points are equidistant, the pairs of varieties they represent are equally (dis-)similar. Note that the resulting axes do not have a direct interpretation besides simply indicating relative (dis-)similarity of the data points indexed.[8]

An MDS visualization of the multi-dimensional system of genitive choice is shown in Figure 2. First, we observe that the relationship between the CSAE, FRED, and the Brown family of corpora is best described as triangular – at any rate, the four written corpora form a discrete and clearly identifiable genre of their

only of trends in expository genres. These are most strongly affected by what Biber (2003) has called the "informational explosion" of the twentieth century with its concurrent increases in demands for economic (essentially, space-saving) writing. These findings are therefore not directly transferrable to non-expository genres, although similarities are found on occasion (Mair 2006).

[8] On a technical note, we conducted the MDS analysis using the PROXSCAL (Proximity Scaling) algorithm implemented in SPSS 13.0; the scaling procedure yielded a Normalized Raw Stress value of .00016 and a Tucker's Coefficient of Congruence value of 0.999.

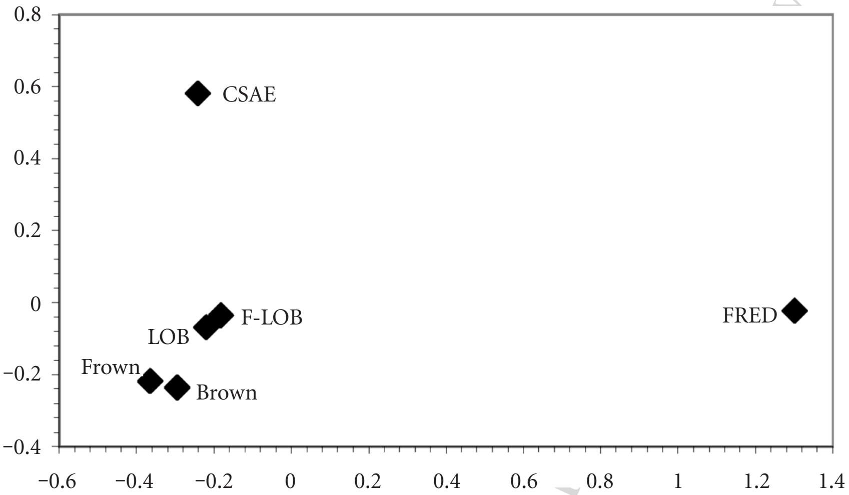

Figure 2. The system of genitive choice in two-dimensional space – MDS visualization of the 9 x 6 odds ratio matrix in Table 1

own. FRED, and hence the traditional dialect speech that the corpus samples, is somewhat more distant from the four Brown corpora than is the CSAE, which samples contemporary conversational American English – recall that we detailed earlier how FRED is very special in terms of the system of genitive choice it exhibits, particularly as regards the (near-)categorical impact that possessor animacy has on genitive choice. Second, note that in the big picture, longitudinal shifts (Brown vs. Frown, LOB vs. F-LOB) are really quite insubstantial: the regional differences (Brown vs. LOB, Frown vs. F-LOB) are clearly more pronounced than any diachronic drift that might have occurred. Third, and most crucially, it is clearly not the case that F-LOB or Frown have shifted substantially in the direction of either of the spoken corpora. Therefore, a "colloquialization of the written norm" (cf. Hundt & Mair 1999) is simply not evident from Figure 2. In a similar vein, it is worth pointing out that the data also do not exhibit a pattern of "Americanization" (cf., for instance, Leech & Smith 2006: 189): this is because explanations along the lines of such a 'follow-my-leader' pattern would predict that F-LOB should have been drifting in the general direction of Brown and/or Frown, and once again such a drift does not emerge from Figure 2.

7. Summary and conclusion

Our analysis of genitive variation in space, time, and across genres has highlighted the fact that raw frequency tallies might be too simplistic to reduce diachronic shifts to hypotheses such as the "colloquialization of the written norm" (cf. Hundt & Mair 1999) conjecture. Our empirical argument, then, boils down to this: while it is true that the *s*-genitive has become roughly as frequent in press English as it is in spoken English, our careful multivariate analysis of seven major conditioning factors has demonstrated that the recent spread of the *s*-genitive in newspaper language is not, in any obvious way, due to a colloquialization of the probabilistic mechanisms of genitive choice. More specifically, we have seen that more often than not, individual factors – for instance, possessor animacy or thematicity of the possessor – have fairly different impacts in spoken and written data. Recall, also, that in the longitudinal perspective (1960s vs. 1990s press English), the spoken and written datasets have become less, rather than more, similar. The increase of the s-genitive, as it emerges from our written data, is due to a dynamics that is rather register-internal, much more so than a phenomenon of the general language.

If colloquialization cannot, as we have argued, account for the spread of the *s*-genitive in journalistic English, how come the *s*-genitive has been spreading in press English nonetheless? We would like to offer that one of the reasons is that journalists have come to increasingly favor the *s*-genitive with thematic possessor NPs (cf. Osselton 1988), arguably because it is simply more economical to use the more compact *s*-genitive with textually recurrent possessor NPs. Second, our regression estimates have indicated that the *s*-genitive is generally favored by journalists in lexically dense environments thanks – once again – to its economy. Notice here that there has been a tendency in the period between the 1960s and the 1990s to cram ever more information (and thus, more lexical material) into a press text of a given length, as we show elsewhere (Hinrichs & Szmrecsanyi 2007). This is a somewhat circumstantial trend which, although *per se* external to the probabilistic system of genitive choice, nonetheless demonstrably favors the *s*-genitive.

In conclusion, then, we offer that it is a process of economization rather than colloquialization which drives the spread of the *s*-genitive in newspaper language. The probabilistic mechanisms behind genitive choice in press English are, and remain, sufficiently distinct from the factors working on genitive variation in spoken English. The *s*-genitive is comparatively frequent in spoken English and in contemporary press English for rather different reasons.

References

Altenberg, B. 1982. *The Genitive v. the Of-Construction. A Study of Syntactic Variation in 17th Century English*. Malmö: CWK Gleerup.

Behaghel, O. 1909/1910. "Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern". *Indogermanische Forschungen* 25.110–142.

Biber, D. 2003. "Compressed noun-phrase structure in newspaper discourse: The competing demands of popularization vs. economy". *New Media Language* ed. by J. Aitchison & D. M. Lewis, 169–181. London & New York: Longman.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Bresnan, J. & J. Hay. Forthcoming. "Gradient grammar: An effect of animacy on the syntax of *give* in varieties of English".

Dahl, L. 1971. "The *s*-genitive with non-personal nouns in modern English journalistic style". *Neuphilologische Mitteilungen* 72.140–172.

Du Bois, J. W., W. L. Chafe, C. Meyer & S. A. Thompson. 2000. *Santa Barbara Corpus of Spoken American English, Part 1*. Philadelphia: Linguistic Data Consortium.

Du Bois, J. W., W. L. Chafe, C. Meyer, S. A. Thompson & N. Martey. 2003. *Santa Barbara Corpus of Spoken American English, Part 2*. Philadelphia: Linguistic Data Consortium.

Givón, T. 1993. *English Grammar. A Function-Based Introduction*. Amsterdam & Philadelphia: Benjamins.

Gries, S. T. 2002. "Evidence in Linguistics: Three approaches to genitives in English". *LACUS Forum XXVIII: What Constitutes Evidence in Linguistics* ed. by R. M. Brend, W. J. Sullivan & A. R. Lommel, 17–31. Fullerton, CA: LACUS.

Hawkins, J. 1994. *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Hernández, N. 2006. "User's Guide to FRED. www.freidok.uni-freiburg.de/volltexte/2489. Freiburg: English Dialects Research Group.

Hinrichs, L. Forthcoming. "The part-of-speech-tagged Brown corpora: A manual of information, including pointers for successful use". University of Freiburg. Ms., 38 pp.

Hinrichs, L. & B. Szmrecsanyi. 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.

Hundt, M. & C. Mair. 1999. "'Agile' and 'uptight' genres: The corpus-based approach to language change in progress". *International Journal of Corpus Linguistics* 4.221–242.

Jucker, A. 1993. "The genitive versus the *of*-construction in newspaper language". *The Noun Phrase in English. Its Structure and Variability* ed. by A. Jucker, 121–136. Heidelberg: Carl Winter.

Kreyer, R. 2003. "Genitive and *of*-construction in modern written English. Processability and human involvement". *International Journal of Corpus Linguistics* 8:2.169–207.

Kruskal, J. B. & M. Wish. 1978. *Multidimensional Scaling*. Newbury Park, London, New Delhi: Sage Publications.

Leech, G. & N. Smith. 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.

Leech, G. & N. Smith. 2006. "Recent grammatical change in written English 1961–1992: some preliminary findings of a comparison of American with British English". *The Changing Face of Corpus Linguistics* ed. by A. Renouf & A. Kehoe, 185–204. Amsterdam & New York: Rodopi.

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

Nerbonne, J., W. Heeringa & P. Kleiweg. 1999. "Edit distance and dialect proximity". *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison* ed. by D. Sankoff & J. B. Kruskal. Stanford: CSLI.

Orwin, R. 1994. "Evaluating coding decisions". *The Handbook of Research Synthesis* ed. by H. Cooper & L. Hedges, 139–162. New York: Russell Sage Foundation.

Osselton, N. 1988. "Thematic genitives". *An Historic Tongue: studies in {English} Linguistics in Memory of Barbara Strang* ed. by G. Nixon & J. Honey. London: Routledge.

Pampel, F. 2000. *Logistic Regression. A Primer*. Thousand Oaks: Sage Publications.

Potter, S. 1969. *Changing English*. London: André Deutsch.

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London & New York: Longman.

Raab-Fischer, R. 1995. "Löst der Genitiv die *of*-Phrase ab? Eine korpusgestützte Studie zum Sprachwandel im heutigen Englisch". *Zeitschrift für Anglistik und Amerikanistik* 43:2.123–132.

Rosenbach, A. 2002. *Genitive Variation in English: Conceptual Factors in Synchronic and Diachronic Studies*. Berlin & New York: Mouton de Gruyter.

Rosenbach, A. 2003. "Aspects of iconicity and economy in the choice between the *s*-genitive and the *of*-genitive in English". *Determinants of Grammatical Variation in English* ed. by G. Rohdenburg & B. Mondorf, 379–412. Berlin & New York: Mouton de Gruyter.

Rosenbach, A. 2005. "Animacy versus weight as determinants of grammatical variation in English". *Language* 81:3.613–644.

Rosenbach, A. 2006. "Descriptive genitives in English: A case study on constructional gradience". *English Language and Linguistics* 10:1.77–118.

Sankoff, D. & W. Labov, W. 1979. "On the use of variable rules". *Language in Society* 8.189–222.

Szmrecsanyi, B. 2006. *Morphosyntactic Persistence in Spoken English: A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis*. Berlin & New York: Mouton de Gruyter.

Wasow, T. 2002. *Postverbal Behavior*. Stanford, CA: CSLI Publications.

Zaenen, A., J. Carlette, G. Garretson, J. Bresnan, A. Koontz-Garboden, T. Nikitina, M. C. O'Connor & T. Wasow. 2004. "Animacy encoding in English: Why and how". *Proceedings of the 2004 ACL Workshop on Discourse Annotation, Barcelona, July 2004* ed. by D. Byron & B. Webber, 118–125.