The English genitive alternation in a cognitive sociolinguistics perspective

Benedikt Szmrecsanyi

As a corpus-based inquiry into the probabilistic nature of lectal variation, the present study seeks to explore how language-external determinants of linguistic variation – real time, geography, text type – interact with language-internal determinants of linguistic variation, and in so doing shape cognitive and probabilistic grammars. The concrete empirical attention of this study will be directed toward the English genitive alternation as an instructive case study. The evidence suggests that the probabilistic grammar underlying the system of genitive choice is fundamentally the same across sampling times, geographic varieties of English, and text types. This overall qualitative stability notwithstanding, the importance of individual conditioning factors varies across different data sources, and this variability is shown to be mediated by language-external factors.
Keywords: variation, English, genitives, multivariate, real time, text type, standard varieties

1. Introduction

As is well known, English has two grammatically overt means of expressing genitive relations, the *of*-genitive (also known as the 'Norman genitive', 'periphrastic genitive', or 'of-construction'), as in (1), and the *s*-genitive (also known as the 'Saxon genitive'), as in (2):

(1) *... this session is helpful for all of us in that it forces us to rethink, to problematize, and to interrogat**e the history of American anthropology*** ... (Corpus of Spoken American English, text 1034)

(2) *While **anthropology's history** is indeed implicated in the scientific construction as race as a biological fact ...* (Corpus of Spoken American English, text 1034)

In modern English, the two genitives are fairly interchangeable and near-equivalent ways of saying the same thing in a considerable number of contexts (Jucker 1993: 121). For example, *anthropology's history* and *the his-*

tory of anthropology are certainly close paraphrases, and it is such choice contexts that will be in the focus of attention in the present investigation.

Where an *s*-genitive can be paraphrased by an *of*-genitive (or vice versa), which factors bear on language users' choice? Extant research has identified a multitude of parameters affecting the English genitive alternation. The literature suggests four major language-internal factor groups:

(i) *Semantic and pragmatic factors*. Animate possessors attract the *s*-genitive, inanimate possessor attract the *of*-genitive (for instance, Altenberg 1982: 117-148); increased thematicity (i.e. text frequency) of the possessor NP makes usage of the s-genitive significantly more likely (Hinrichs and Szmrecsanyi 2007).

(ii) *Phonology*. A final sibilant in the possessor NP (for instance, in a plural morpheme) attracts the *of*-genitive (for instance, Hinrichs and Szmrecsanyi 2007).

(iii) *Processing and parsing-related factors*. Thanks to the principle of end-weight (Behaghel 1909/1910), longer 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 (cf. Szmrecsanyi 2005a,b, 2006). Thus, 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.

(iv) *Economy-related factors*. By virtue of being "more compact" (Biber et al. 1999: 300), the *s*-genitive is more frequent in contexts and registers where the "tendency to brevity" (Dahl 1971: 172) is pivotal. There is also evidence that journalists favor the *s*-genitive in contexts characterized by comparatively high informational/lexical density (Hinrichs and Szmrecsanyi 2007).

In addition, the genitive alternation is also sensitive to a number of language-external factors:

(v) *External factors*. In historical terms, the *of*-genitive has been the long-term incoming form, yet the *s*-genitive has bounced back during the Modern English period and is claimed to be spreading right now, especially in press language (for instance, Raab-Fischer 1995; Mair 2006). As for genre/text type stratification, more informal settings usually favor the *s*-genitive (for instance, Altenberg 1982: 284) – so, the *s*-genitive should be particularly frequent in spoken data (Rosenbach 2002: 39). In terms of geographic differences, the *s*-genitive is known to be more frequent in American English than in British English (cf., for example, Rosenbach 2003: 395-396).

The univariate impact of each of the factors mentioned above is amply documented in the literature. The aim of the present research, by contrast, is to fit a multivariate logistic regression model describing the probabilistic grammar of genitive choice, with special attention being paid to how the external factors in (v) shape and determine the factor weights of the internal factors in (i) – (iv). To address this particular issue, the present study will rely heavily on visualization techniques such as cluster analysis and multi-dimensional scaling. An interesting issue along these lines is the cultural-cognitive motivation driving the on-going spread of the *s*-genitive, especially in press language: is this a text-type-interdependent process such that we are witnessing a 'colloquialization of the norms of written English' (Leech and Smith 2006; Hundt and Mair 1999)? Alternatively, are we seeing a geography-related process of 'Americanization' (such that the *s*-genitive would become more frequent in British English because it is frequent in American English')? Or are we rather dealing with a process of "economization" (Hinrichs and Szmrecsanyi 2007: 469), such that journalists have to increasingly convey ever more information in ever less paper space, a constraint that would favor the more compact *s*-genitive?

In addressing these issues, and thus sketching a more complete, more realistic, and thoroughly usage-based picture of linguistic variation in space, time, and across text types, the overarching objective of this study is to explore the gradient interaction between language-internal and language-external factors as a *cognitive* and *cultural* phenomenon that comes within the remit of cognitive sociolinguistics.

2. Method and Data

The present study will re-examine the database and the coded dataset drawn upon in Szmrecsanyi and Hinrichs (2008). This section will loosely paraphrase their methods section.

2.1. Data

The database taps the following corpora sampling naturalistic language data:

– 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 (see Hernández 2006; Szmrecsanyi and Hernández 2007) contains samples of dialectal speech (mainly transcribed so-called 'oral history' material) from a variety of sources. The bulk of these samples was 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. 1.3 million 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).

– 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 contain a structured dataset of four million words of running text. The present study will focus on journalistic language and therefore explore 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, relying on the recently completed part-of-speech-tagged versions of the corpora (see Leech and Smith 2005; Hinrichs et al. 2007).

In sum, the database to be explored here comprises material from different sampling times (1960s [LOB, Brown] vs. 1990s [F-LOB, Frown] press English), different geographic varieties (American English [CSAE, Brown, Frown] vs. British English [FRED, LOB, F-LOB]), and different text types (spoken [FRED, CSAE], written press reportage [Brown-A, LOB-A, Frown-A, F-LOB-A], written press editorials [Brown-B, LOB-B, Frown-B, F-LOB-B]).

2.2. Method

All occurrences of interchangeable *s*- and *of*-genitives were manually identified in the database, 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. This procedure yielded a dataset 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).

While Szmrecsanyi and Hinrichs (2008) provide a detailed description of the coding scheme for interchangeability, suffice it to say here that the coding procedure only considered 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;

consequently, 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.[1]

3. A first overview: text frequencies

To provide a first impression of the degree of variation exhibited in the dataset, Figure 1 presents the relative frequency of the *s*-genitive (as a percentage of all interchangeable genitives) across the 10 (sub)corpora studied. As can be seen, there is a good deal of frequency variation: the share of the *s*-genitive ranges from 29.8% in LOB-B to 59.6% in FRED, and while the mean share of *s*-genitive, across all (sub)corpora, is 44.4%, the standard deviation associated with this mean value is a very considerable 10 per cent points.

[insert Figure 1]

[insert Table 1]

Table 1 elucidates how this variation in text frequency is sensitive to language-external factors. Firstly, as far as sampling time is concerned, the *s*-genitive has become a good deal more frequent in press language in the period between the 1960s and the 1990s, which supports claims in the literature (for instance, Raab-Fischer 1995) that the *s*-genitive is spreading in real time. Secondly, the *s*-genitive is overall a tad more frequent in the American data than in the British data. While the differential is not statistically significant, it nonetheless dovetails 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. Observe, however, that while the *s*-genitive is actually substantially more frequent in American press English than in British press English (mean shares: 43.8% vs. 40.2%), the situation is just the reverse in the spoken data sources, CSAE and FRED (mean shares: 48.2% vs. 59.5%). Notice also that in contemporary press English (as sampled in Frown and F-LOB), the *s*-genitive is significantly more frequent in the American data (mean share: 53.2%) than in the British data (mean share: 45.8%), while the difference between Ameri-

can English and British in the 1960s is marginal (mean shares: 35.0% vs. 36.8%). Thirdly, a fairly neat text-type continuum emerges: the *s*-genitive is most frequent in spoken data and least frequent in press editorials, while press reportage covers the middle ground. This suggests that press reportage is, compared to press editorials, the more 'colloquial' text type.

The above discussion of overall text frequencies – one-dimensional as they are – has indicated that genitive variation indeed seems to be sensitive to language-external factors. In what will follow, this study will treat text frequencies as epiphenomenal to the probabilistic and cognitive mechanics which underlie the multidimensional system of genitive choice, with a special interest in the role that external factors play in this system.

4. Conditioning factors in genitive choice

Following the methodology of Szmrecsanyi and Hinrichs (2008), the present study aims to model genitive frequencies as a function of seven major language-internal conditioning factors. These fall into four groups: (i) semantic and pragmatic factors (animacy and thematicity of the possessor), (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).

4.1. Possessor animacy

Animacy of the possessor NP is commonly claimed to be the chief determinant of genitive choice. Adopting Rosenbach's (2006: 105) animacy hierarchy (human > animal > collective > inanimate) and drawing on Zaenen et al.'s (2004) general coding scheme for animacy, each possessor NP in the dataset was manually annotated according to the following four-way classification: (i) human 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).[2]

(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)

4.2. 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. So, 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). Assuming, in this spirit, that increased text frequency of a possessor NP would make the *s*-genitive more likely, the *log*-transformed text frequency 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) was established for every individual possessor NP in the dataset. The example in (7) will illustrate the basic idea:

(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 (*log* value: 1.5) in Frown text A02 (which spans about 2,000 words).

4.3. Final sibilants in the possessor NP

A final sibilant in the possessor NP, as in (8), may discourage usage of the s-genitive (cf. Altenberg 1982):

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

All possessors in the dataset ending, orthographically, in <s> (as in *Congress*), <z> (as in *jazz*), <ce> (as in *resistance*), <sh> (as in *Bush*), or <tch> (as in *match*) were identified and annotated.[3]

4.4. End weight: possessor and possessum length

The time-honored principle of 'end-weight' (for instance, Behaghel 1909/1910; Wasow 2002) postulates that language users tend to place 'heavier,' more complex constituents after shorter ones, yielding a constituent ordering that might facilitate parsing (see, for example, Hawkins 1994). Hence, if the possessor is heavy, there should be a general preference for the *of*-genitive because it places the possessor last. If the possessum is heavy, a general preference for the *s*-genitive is expected because it places the possessum last. The present study seeks to approximate the weight of genitive constituents by determining their length in graphemic words (see Szmrecsanyi 2004 for an empirical argument that vis-à-vis other measures, counting graphemic words approximates syntactic weight surprisingly well). 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, though, 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*).

4.5. 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. We hypothesize 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 (see Szmrecsanyi 2006: 87-107). So, each genitive occurrence in the dataset was annotated according to 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 .... In Ecuador,* ***the country's Medical Association*** *said 100 people had died of a total of 5,000 cases…* (F-LOB A14)

4.6. Lexical density and type-token ratios

Hinrichs and Szmrecsanyi (2007) demonstrate that the *s*-genitive is attracted by 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 (Biber et al. 1999: 99). To check on this factor, Perl scripts established the type-token ratios of the textual passages (50 words before and 50 words after a given genitive construction) where the genitive occurrences in the dataset were embedded.

5. Results

5.1. A regression model of genitive choice

We will now draw on *binary logistic regression* (see Pampel 2000) 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 and Labov 1979: 151). In predicting a binary outcome (i.e. a linguistic choice, in the case of the present study whether speakers/writers will choose an *s*-genitive over an *of*-genitive) on the basis of 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 (also known as *factor weights*). This information is provided by *odds ratios* (ORs), which indicate 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 (in our case: choice of an *s*-genitive) 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 (if smaller than 1), the more disfavoring the effect.

– Variability accounted for by (or: explanatory power of) the model as a whole (R^2). The R^2 value can range between 0 and 1 and gauges 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 in the present study 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 (*% correct*) vis-à-vis the baseline prediction (*% baseline*) indicates how accurate the model is in predicting actual outcomes. The higher this percentage, the better the model.

[insert Table 2]

Rather than fitting a one-size-fits-all regression model on the entire dataset and modeling the effect of external factors via interaction terms, the present investigation fits 10 independent regression models – one for each of the (sub)corpora under analysis – on the language-internal factors discussed in section 4 above.[4] The results are provided in Table 2. Predictive efficiency of the models is satisfactory: on the basis of the conditioning factors considered, the models predict between 70.4% (CSAE) and 88.8% (FRED) of the genitive outcomes accurately. Variance explained (R^2) ranges between .34 (LOB-B) and .68 (FRED), which is another way of saying that we can account for between 34% and 68% of the observable variability in the (sub)corpora under analysis – the remainder of the variability may be due to free variation, or to other conditioning factors not considered in the present study. In all, the system of genitive choice sketched in Table 2 works

best for the very traditional dialect speech sampled in FRED, and least well (though still somewhat satisfactorily) for 1960s British English press editorials, as sampled in LOB-B. There is, moreover, a tendency for those models on spoken data to have a better fit than models on written data (mean R^2 spoken data: .56, mean R^2 spoken data: .45), which may suggest that in written data, other factors not considered here (stylistics, prescriptivism, etc. …) might have more weight than in spoken data.

Let us next discuss individual factor groups and their effect on genitive choice. As for semantic and pragmatic factors, consider animacy. The models reported in Table 2 take inanimate possessors (*the Honda*, *a rock*, etc.) as the default category and quantify the effect that human/animal/collective possessors have on the odds that an *s*-genitive will be chosen. The effect of human and collective possessors is statistically significant throughout, while animal possessors are significant in the spoken corpora only (the simple reason for this being that animal possessors are a very rare species in press material). The factor also has the theoretically expected effect direction: as a generalization, the more animate a possessor is, the greater the odds that an *s*-genitive will be chosen. Take, for instance, Brown-A: if the possessor is animate (e.g. *the emperor*, *John*) instead of inanimate (e.g. *the Honda*, *a rock*), the odds that an *s*-genitive will be chosen increase by a factor of 8.53. If the possessor is a collective noun (e.g. *the Pentagon*, *the police*), the odds for an *s*-genitive increase by a factor of 3.40. Notice now that there is a general tendency for human possessors to attract s-genitives more strongly in the British data (mean OR: 25.65) than in the American data (mean OR: 8.72), suggesting that the *s*-genitive is cognitively more strongly associated with human possessors in British English than in American English. FRED is an extreme case: 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 is unlikely to be due to, e.g., the text type (interviews) sampled in FRED. Instead, what we are seeing here is probably 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. Notice here that this line of reasoning does not contradict the fact that the *s*-genitive is becoming more frequent in Present-Day English – the contemporary expansion of the *s*-genitive in press English is actually quite unrelated to the animacy constraint.

As detailed above, the literature suggests that increased thematicity of the possessor – operationalized as the possessor head noun's *log* text fre-

quency in a given corpus text – makes the *s*-genitive more likely. In the written data sources, this hypothesis is indeed borne out: for 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 per corpus text instead of 1 occurrence per corpus text), the odds for the *s*-genitive increase by a factor of between 1.20 (Brown-A, LOB-A) and 2.14 (Frown-A). Overall, the factor appears to be somewhat more powerful in the written American data (mean OR: 1.70) than in the British data (mean OR: 1.39). It is also stronger in 1990s texts (mean OR: 1.80) than in 1960s texts (mean OR: 1.29). By contrast, the factor is not even selected as significant in the spoken corpora (CSAE and FRED). In other words, possessor thematicity is characteristic of written, not spoken, language.

Turning to phonology, a final sibilant in the possessor significantly and reliably discourages usage of the *s*-genitive, as expected: the presence of a final sibilant decreases the odds for an *s*-genitive by between 46% (LOB-B) and 79% (CSAE). There is hardly any difference between the written (mean OR: .32) and the spoken data sources (mean OR: .29), though interestingly the constraint has become significantly (cf. Hinrichs and Szmrecsanyi 2007) more influential over time in press language (mean OR 1960s: .28, mean OR 1990s: .25). The somewhat curious fact that a phonological constraint should become more influential in press language (a written genre) over time advertises itself to be interpreted in terms of a "colloquialization of the norms of written English" (Leech and Smith 2006; Hundt and Mair 1999).

What about factors relating to parsing and processing? As hypothesized, longer possessor NPs significantly and consistently disfavor the *s*-genitive (because this coding option places the possessor second): for every additional word in the possessor NP, the odds for an *s*-genitive decrease by between 62% (Brown-A) and 37% (LOB-B), an effect which, among the written data sources, is stronger in press reportage material (mean OR: .40) than in press editorials material (mean OR: .51). Conversely, longer possessum NPs significantly attract the *s*-genitive in six of the ten data sources studied: thus, for every additional word in the possessum NP, the odds for an *s*-genitive increase by between 19% (Brown-A) and 97% (F-LOB-B). In this connection it should be noted that possessum length does not seem to be important in the spoken data sources, which is another way of saying that the factor is a characteristic of the written, not spoken, English system of genitive choice.

The factor 'persistence' is significant in six of the ten (sub)corpora studied (it is not significant in 1990s press English), and has the theoretically expected sign throughout: among the data sources where the factor is significant, precedence of an *s*-genitive in the ongoing discourse increases the odds for another, subsequent *s*-genitive by a factor of between 1.44 (LOB-A) and 3.53 (CSAE). 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 (on this point, cf. Szmrecsanyi 2005a,b, 2006).

We finally move on to the economy-motivated factor in the variable portfolio, *viz*. lexical density as approximated by the type-token ratio of a given genitive passage. Recall that we assumed that speakers/writers would resort to the more economical *s*-genitive in contexts characterized by high type-token ratios and thus high lexical (or: informational) density. For writers (though not for speakers), this hypothesis is borne out: 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 1.58 (F-LOB-B) and 2.55 (LOB-B). Because the predictor is not even selected as significant in the spoken data sources, the sort of economy implicit in the nature of the predictor appears not to be important in spoken language.

By way of an interim summary, the most important finding of this portion of the analysis is that the grammar of genitive choice is *qualitatively* (that is, in terms of the effect direction of the factors studied) very similar in all of the ten (sub)corpora under investigation. Where significant, more animate and thematic possessors, longer possessum phrases, precedence of an *s*-genitive, and higher type-token ratios all attract the *s*-genitive. Final sibilants and long possessor phrases, in turn, attract the *of*-genitive. At the same time, we have seen that the *magnitude* of the effect of individual predictors may vary, statistically, as a function of a number of language-external factors – time, geography, and text type. In an attempt to see the wood for the trees, it should be worthwhile to invoke this quantitative variance to establish aggregate similarities (and dissimilarities) between the cognitive and probabilistic grammars of genitive choice. It is to this task that I next turn.

5.2. Aggregate similarities between genitive choice systems

Thus far, we have sought to characterize the cognitive and probabilistic grammar of genitive choice in English on the basis of a complex system of conditioning factors, yielding ten sets of nine discrete odds ratios – one for each data source under analysis – which characterize this system. Note, now, that fine-grained and instructive as the analysis of conditioning factors may be, its multidimensional nature makes it rather difficult to spot overarching tendencies and patterns relying merely on one's eyeballs. This is why we will now abandon our earlier focus on individual factors and their probabilistic weights, turning instead to two non-parametric statistical analysis and visualization techniques (*cluster analysis* and *multidimensional scaling*) to uncover the 'big' picture of genitive variation in time, geography, and across text types.

[insert Figure 2]

Cluster analysis is a cover term for a set of techniques designed to objectively group a given number of cases (in this study, probabilistic grammars of genitive choice) into a smaller number of discrete and meaningful clusters on the basis of some sort of similarity – in our case, similarities between probabilistic factor weights – in order to establish higher-order patterns in an objective way (for an introduction to the technique from the social scientist's perspective, see Aldenfelder and Blashfield 1984). Data clustering can be visually represented using tree diagrams, also known as *dendrograms*, which work in essentially the same way as family trees. The dendrogram deriving from this study's dataset (more specifically, from the probabilistic factor weights in Table 2) can be seen in Figure 2.[5] In this dendrogram, the first and most basic split occurs between the written and the spoken (sub)corpora under investigation. Further down the road, the written cluster regroups into two subclusters, yielding a three-cluster solution at a (statistically comparatively robust) cophenetic distance of 1.0, as indicated by the dotted vertical line in Figure 2. The first of the two written subclusters contains the British press reportage subcorpora (LOB-A, F-LOB-A), the 1960s British press editorials subcorpus (LOB-B), and the two American subcorpora (Frown-B, Brown-A). The second of the two written subclusters is more homogeneous, containing 1990s British editorials (F-LOB-B) as well as the remainder of the American material (Brown-B, Frown-A). In all, Figure 2 makes amply clear that the most fundamental

split – as indicated by the distance from the leaves to the encompassing node – in the dataset occurs between the written and the spoken material, which testifies to the paramount importance of the written-spoken distinction for the exact quantitative shape of a given system of genitive choice. This distinction overrides all other language-external factors.

[insert Figure 3]

The dendrogram in Figure 2 has provided us with a first impression of the similarities and dissimilarities between genitive choice systems as exhibited in our dataset. For the remainder of this section, we will rely on *multidimensional scaling* to visualize the hidden structure of genitive variation in time, space, and across text types (for an introduction to multidimensional scaling, see Kruskal and Wish 1978). This means that we will scale down the original nine dimensions (i.e. probabilistic factor weights) by which every genitive choice system in our dataset is characterized to two dimensions, an exercise which will make it possible to visualize the aggregate (dis-)similarities between these systems in two-dimensional maps. The advantage of such perceptual maps is that these can be interpreted fairly intuitively: much as with geographic maps, the further two points are apart, the more dissimilar (in geographic terms, distant) they are. If two pairs of points are equally close or distant, the pairs of genitive choice systems they represent are equally (dis-)similar.[6] The resulting visualization is given in Figure 3; also shown in this figure are cluster memberships as derived from hierarchical agglomerative clustering (see Figure 2).

We observe, first and foremost, that the relative distance between the spoken material in FRED and the CSAE (cluster 1) and the written macro cluster (clusters 2 and 3) is considerable. So, in a bird's eye perspective, the written (sub)corpora clearly form a genre of their own, which is different from the spoken material. What is happening within the written text types, though? To begin to address this question, consider the position of the data points relative to the vertical axis: high values (as in cluster 1) are associated with spoken material, so the vertical axis may be considered indicative of increasing levels of orality, i.e. *colloquiality*. Assuming that this interpretation is correct, the material in cluster 3 is least colloquial, while the material in cluster 2 covers the middle ground. It turns out, therefore, that cluster analysis has grouped the material in the dataset according to increasing levels of colloquiality as, once again, the most important external parameter working on genitive choice systems. What is the interpretation of

the horizontal axis? Observe that all data sources yielding negative scores on the horizontal axis sample British material, while all the data sources yielding positive values comprise American material. The horizontal axis may thus be considered being indicative of increased '*Americanness*'.

The *colloquiality* vs. '*Americanness*' dimensions underlying the plot in Figure 3 yield an additional four-way classification of the material in our dataset: the upper left-hand quadrant in Figure 3 is the *colloquial/British* quadrant, the upper right-hand quadrant is the *colloquial/American* quadrant, the lower right-hand quadrant is the *written/American* quadrant, and the lower left-hand quadrant is the *written/British* quadrant. Having so at once taken care of the external parameters 'text type' (spoken vs. written) and 'geography', we will now go on to a discussion of drifts, among the written material, in real time. Recall from the literature review that Hinrichs and Szmrecsanyi (2007: 469) have shown that in written English in particular, it is primarily a process of 'economization' that drives the spread of the *s*-genitive in real time. Szmrecsanyi and Hinrichs (2008) – not differentiating between the written genres (press editorials vs. press reportage) that are subject to differentiation in the present study – likewise suggest that press language *as such* cannot be said to have substantially colloquialized. In the light of the present study's more fine-grained distinction between press reportage and press editorials, and on the basis of Figure 3 (consider the arrows indicating diachronic drifts), these claims can be restated more precisely in the following way:

– *British press reportage* (LOB-A → F-LOB-A) exhibits a modest drift towards less colloquiality ('de-colloquialization') as well as modest Americanization;

– *British press editorials* (LOB-B → F-LOB-B) attest a considerable drift towards more colloquiality ('colloquialization') as well as modest shift away from the American sector of the diagram;

– *American press reportage* (Brown-A → Frown-A) shows a medium-scale drift towards more colloquiality ('colloquialization') and slight Americanization (to the extent, of course, that a *per se* American genre can become even more American);

– *American press editorials* (Brown-B → Frown-B) are characterized by a modest shift towards less colloquiality ('de-colloquialization') and medium-scale Americanization (cf. the caveat above).

This exercise in drift tracing has suggested that curiously – as far as the direction of the drifts (and not the respective endpoints) are concerned – British press reportage aligns with American press editorials, and British press editorials somewhat align with American press reportage. In sum, the data reveal that while consistent with extant literature there is no such thing as a robust *overall* pattern of colloquialization or Americanization in press English, the two processes are arguably still somewhat involved in diachronic drift, depending on text type and geographic variety. The mediating factor that very likely accounts for this interpretatorial twilight is *economization*, *viz*. the differential importance, depending on text type, of the "tendency to brevity" (Dahl 1971: 172) and of the need to save paper space by opting for more compact coding options (such as the *s*-genitive) instead of more explicit coding options (such as the *of*-genitive). Because such pressures are arguably more acute in press reportage than in editorials, we see differential drift directions (a more detailed discussion of this issue is provided in Szmrecsanyi and Hinrichs 2008).

6. Concluding remarks

The foregoing analysis leads to two principal conclusions about the alternation between the *s*-genitive and the *of*-genitive in English in a cognitive sociolinguistics perspective. For one thing, we have seen that while there is a good deal of variation in text frequencies, the probabilistic grammar underlying the system of genitive choice is fundamentally the same across sampling times, geographic varieties of English, and text types: animate possessors are cognitively associated with the *s*-genitive, long possessor NPs trigger the *of*-genitive, and so on. On the other hand, however, the magnitude of the effect that individual conditioning factors may have on genitive choice can vary substantially across different data sources, and this statistical variance is demonstrably mediated by language-external factors. By aggregating individual factor weights to an aggregate measure of distance between genitive choice systems and by subsequently partitioning and visualizing the resulting variance, this study has sought to demonstrate that the most important language-external factor working on the English

genitive alternation is the written/spoken text-type distinction, and that the real-time drift of written genitive choice systems – depending on their exact genre and on whether they are British or American – may be differentially impacted by cultural phenomena such as colloquialization, Americanization, or economization. On more methodological grounds, this study highlights the fact that by exploring how language-external and cultural factors leave their mark on the quantitative footprint of probabilistic grammars, and thus on the cognitive factors that motivate linguistic choices, we can learn a lot about how language variation is more patterned and predictable than one might perhaps think. In exactly this spirit, further study may wish to continue this line of inquiry to explore, e.g., how genuinely sociological variables such as age, gender, and social class interact with probabilistic grammars.

Notes

1. As for interrater reliability, parallel annotation of a set of $N = 202$ genitives by two trained coders 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. Hinrichs and Szmrecsanyi (2007: section 3) provide more detail.
2. Interrater reliability of animacy coding 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 .79. Hinrichs and Szmrecsanyi (2007: section 5.1.1) provide more detail.
3. Possessors ending in <dge> (as in *judge*) are so rare that they were excluded from analysis.
4. Note that this is mainly for expository purposes – interaction terms can be notoriously hard to interpret. Also notice that the analysis techniques drawn on in Section 5.2. (cluster analysis and multidimensional scaling) will draw on the discrete odds ratio vectors presented in Table 2. See Hinrichs and Szmrecsanyi (2007) for a uniform model of genitive choice in the Brown family of corpora that models the effect of language-external factors as interaction terms. I should also like to point out that in the present study's dataset, there are no statistically significant and/or substantially interpretable interaction effects *between* the language-internal factors considered here (say, between animacy and thematicity).
5. Technically, the set of 10×9 odds ratios in Table 2 was first *log*-transformed (in order to alleviate the effect of outliers) and then converted into a distance matrix using Euclidean distance as an interval measure. On the basis of this distance matrix, a hierarchical agglomerative clustering algorithm (specifically, Ward's Minimum Variance method) subsequently partitioned the (sub)corpora in the dataset into clusters. Note that because simple clustering can be unstable (see, for instance, Nerbonne et al. 2007), the robustness of the dendrogram in Figure 2 was assessed by also running three other common clustering algorithms – Weighted Average (WPGMA), Group Average (UPGMA), and Complete Link – on the dataset. Since the exact same dendrogram as reported in Figure 2 also emerged in two of the three additional runs (with only the Complete Link algorithm yielding a slightly different clustering outcome), the dendrogram in Figure 2 can be considered fairly reliable.
6. The scaling procedure was conducted using the Proxscal algorithm implemented in SPSS, on the basis of the same distance matrix (derived from

Euclidean distances in the *log*-transformed set of 10 × 9 odds ratios) used as input to the cluster analysis (see previous footnote). The resulting two-dimensional scaling solution yields a normalized raw stress value of .0012, a dispersion-accounted-for value of .99, and a Tucker's coefficient of congruence value of .99.

References

Aldenfelder, Mark, and Roger Blashfield
1984 *Cluster Analysis*. Newbury Park/London/New Delhi: Sage Publications.

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

Behaghel, Otto
1909/10 Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25: 110-142.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan
1999 *Longman grammar of spoken and written English*. Harlow: Longman.

Dahl, Lisa
1971 The *s*-genitive with non-personal nouns in modern English journalistic style. *Neuphilologische Mitteilungen* 72: 140-172.

Du Bois, John W., Wallace L. Chafe, Charles Meyer, and Sandra A. Thompson
2000 *Santa Barbara corpus of spoken American English*, Part 1. Philadelphia: Linguistic Data Consortium.

Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, and Nii Martey
2003 *Santa Barbara corpus of spoken American English*, Part 2. Philadelphia: Linguistic Data Consortium.

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

Hernández, Nuria
2006 *User's Guide to FRED*. Available online at http://www.freidok.uni-freiburg.de/volltexte/2489 (last accessed: 3/13/2008). Freiburg: English Dialects Research Group.

Hinrichs, Lars, Birgit Waibel, and Nicholas Smith
2007 *The POS-tagged, postedited F-LOB and Frown corpora: a manual, including pointers for successful use.*

Hinrichs, Lars, and Benedikt 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, Marianne and Christian Mair

1999 'Agile' and 'uptight' genres: the corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4: 221-242.

Jucker, Andreas

1993 The genitive versus the *of*-construction in newspaper language. In *The Noun Phrase in English. Its Structure and Variability*, Andreas Jucker (ed.), 121-136. Heidelberg: Carl Winter

Kreyer, Rolf

2003 Genitive and *of*-construction in modern written English. Processability and human involvement. *International Journal of Corpus Linguistics* 8(2): 169-207.

Kruskal, Joseph, and Myron Wish

1978 *Multidimensional Scaling.* Newbury Park/London/New Delhi: Sage Publications.

Leech, Geoffrey, and Nicholas 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.

2005 Recent grammatical change in written English 1961-1992: some preliminary findings of a comparison of American with British English. In *The Changing Face of Corpus Linguistics*, Antoinette Renouf and Andrew Kehoe (eds.), 185-204. Amsterdam/New York: Rodopi.

Mair, Christian

2006 Inflected genitives are spreading in present-day English, but not necessarily to inanimate nouns. In *Corpora and the History of English: Festschrift für Manfred Markus*, Christian Mair (ed.). Heidelberg: Winter.

Nerbonne, John, Peter Kleiweg, Franz Manni, and Peter Heeringa

2007 Projecting dialect differences to geography: Bootstrap clustering vs. noisy clustering. In: *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classiffcation Society*, Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt, and Reinhold Decker (eds.). Berlin: Springer.

Orwin, Robert

1994 Evaluating Coding Decisions. In *The Handbook of Research Synthesis*, Harris Cooper and Larry Hedges (eds.), 139-162. New York: Russell Sage Foundation.

Osselton, Noel

1988 Thematic Genitives. In *An Historic Tongue: studies in English linguistics in memory of Barbara Strang*, Graham Nixon and John Honey (eds.). London: Routledge.

Pampel, Fred

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

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik

1985 *A Comprehensive Grammar of the English Language*. London, New York: Longman.

Raab-Fischer, Roswitha

1985 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, Anette

2002 *Genitive variation in English: conceptual factors in synchronic and diachronic studies*. Berlin/New York: Mouton de Gruyter.

2003 Aspects of iconicity and economy in the choice between the *s*-genitive and the *of*-genitive in English. In *Determinants Of Grammatical Variation in English*, Günter Rohdenburg and Britta Mondorf (eds.), 379-412. Berlin, New York: Mouton de Gruyter.

2006 Descriptive genitives in English: a case study on constructional gradience. *English Language and Linguistics* 10(1): 77-118.

Sankoff, David and William Labov

1979 On the use of variable rules. *Language in Society* 8: 189-222.

Szmrecsanyi, Benedikt

2004 On Operationalizing Syntactic Complexity. In Purnelle, Gérard, Cédrick Fairon, and Anne Dister (eds.), *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis. Louvain-la-Neuve, March 10-12, 2004*. Louvain-la-Neuve: Presses universitaires de Louvain, 1032-39.

2005a Language users as creatures of habit: a corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1(1): 113-150:

2005b Never change a winning chunk. *Recherches Anglaises et Nord-Américaines* 38: 21-34.

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.

Szmrecsanyi, Benedikt, and Nuria Hernández

2007 *Manual of Information to accompany the Freiburg Corpus of English Dialects Sampler ("FRED-S")* Available online at http://www.freidok.uni-freiburg.de/volltexte/2859/ (last accessed: 3/13/2008). Freiburg: English Dialects Research Group.

Szmrecsanyi, Benedikt, and Lars Hinrichs

2008 Probabilistic determinants of genitive variation in spoken and written English: a multivariate comparison across time, space, and genres. In *The Dynamics of Linguistic Variation: Corpus Evidence on English Past and Present*, Terttu Nevalainen, Irma Taavitsainen,

Päivi Pahta and Minna Korhonen (eds.), 291-309. Amsterdam: Benjamins.

Wasow, Thomas

2002 *Postverbal behavior*. Stanford, CA: CSLI Publications.

Zaenen, Annie, Jean Carlette, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor, and Thomas Wasow

2004 Animacy encoding in English: why and how. In *Proceedings of the 2004 ACL workshop on discourse annotation, Barcelona, July 2004,* Donna Byron and Bonnie Webber (eds.), 118-125.

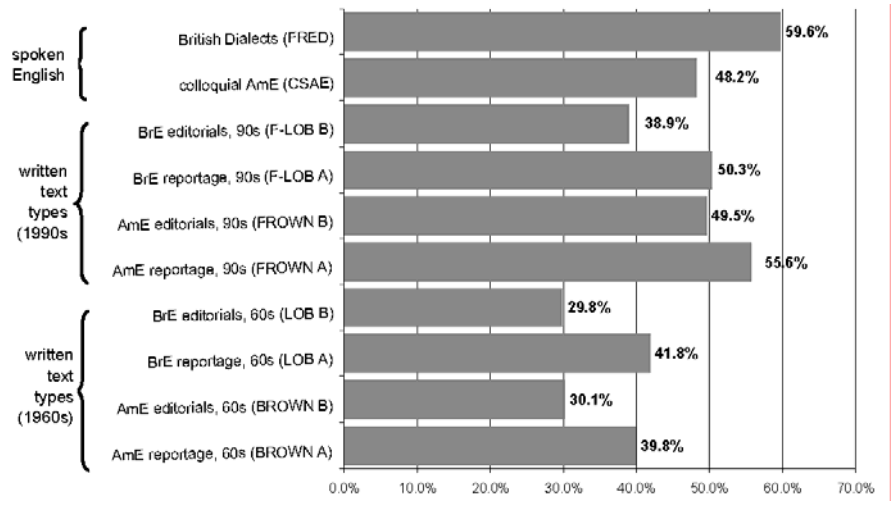

Figure 1. Share of the *s*-genitive among interchangeable genitives across (sub)corpora

Table 1. Mean share of the *s*-genitive among interchangeable genitives according to sampling time, geography, and text type.

	mean share of the s*-genitive*
sampling time	
1960s press English	35.4%
1990s press English	48.6%
Geography	
American English	44.6%
British English	44.1%
text type	
Spoken	53.9%
press reportage	46.9%
press editorials	37.1%

Table 2. Odds ratios (ORs) in logistic regression. Predicted odds are for the *s*-genitive. Significant ORs ($p < .05$) are in bold.

		CSAE	FRED	Brown-A	Brown-B	LOB-A	LOB-B	Frown-A	Frown-B	F-LOB-A	F-LOB-B
possessor animacy (default category: inanimate)	human	**8.08**	**69.66**	**8.53**	**13.00**	**11.01**	**18.40**	**7.25**	**6.76**	**13.84**	**15.36**
	animal	**30.94**	**17.75**	.00	.00	99	6.56	.00	1.55	99	.00
	collective	**3.94**	**2.77**	**3.40**	**3.35**	**3.16**	**3.91**	**3.63**	**2.69**	**4.86**	**5.53**
thematicity of possessor		.90	.99	**1.20**	**1.50**	**1.20**	1.25	**2.14**	**1.95**	**1.82**	**1.29**
final sibilant in possessor		**.21**	**.36**	**.24**	**.25**	**.50**	**.54**	**.22**	**.22**	**.30**	**.27**
possessor length		**.52**	**.42**	**.38**	**.44**	**.42**	**.63**	**.41**	**.55**	**.40**	**.42**
possessum length		1.00	.96	**1.19**	**1.45**	1.16	.95	**1.45**	**1.57**	**1.54**	**1.97**
persistence		**3.53**	**1.87**	**1.66**	**1.51**	**1.44**	**2.11**	1.31	1.34	1.21	1.38
type-token ratio		.90	.95	**2.42**	**1.77**	**2.36**	**2.55**	**1.68**	**2.10**	**2.23**	**1.58**
N		332	1,818	1,329	804	1,241	707	1,244	816	1,138	736
% baseline		52.0	59.6	60.0	70.1	58.4	70.7	55.5	50.6	50.6	61.1
% correct		70.4	88.8	75.7	80.8	75.2	78.8	77.7	76.1	80.5	79.2
Nagelkerke R^2		.43	.68	.45	.46	.42	.34	.48	.42	.54	.49

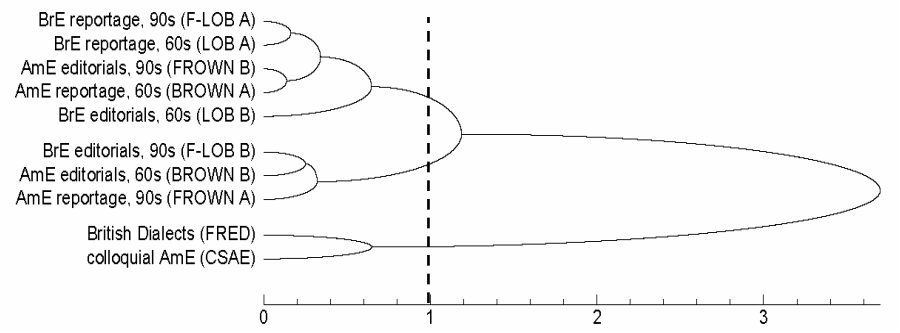

Figure 2. Dendrogram derived from hierarchical agglomerative cluster analysis (cluster algorithm: Ward's method) of the *log*-transformed 10 × 9 odds ratio matrix in Table 2.

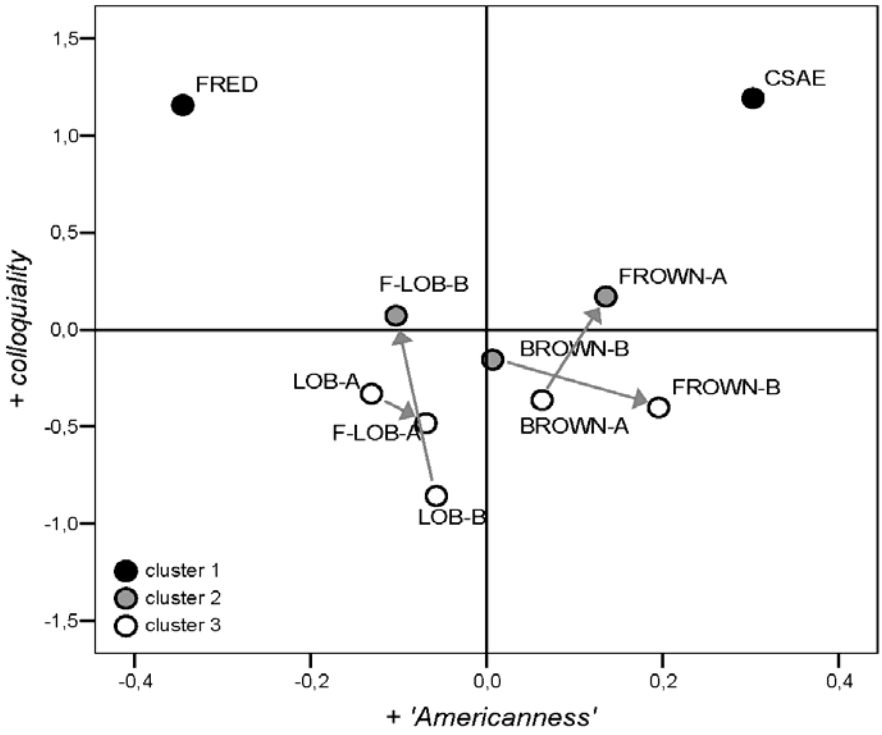

Figure 3. MDS visualization of the *log*-transformed 10 × 9 odds ratio matrix in Table 1. Group memberships derive from hierarchical agglomerative cluster analysis (cf. Figure 2). Arrows indicate drifts in real time.