A Variationist Approach to Cross-Register Language Variation and Change

by

Bridget Lynn Jankowski

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Department of Linguistics
University of Toronto

© Copyright by Bridget Lynn Jankowski 2013
A Variationist Approach to Cross-Register Language Variation and Change

Bridget Lynn Jankowski

Doctor of Philosophy

Department of Linguistics
University of Toronto

2013

Abstract

The comparative method of variationist sociolinguistics has demonstrated that frequency changes are not reliable determinants of whether grammatical change is taking place. Frequency changes can be the result of extra-linguistic register changes, changes within the underlying grammar, or a combination (Szmrecsanyi, 2013; Tagliamonte, 2002a). This work examines two variables known to vary along the written-to-spoken continuum — relative clause pronouns, and the genitive construction — across three registers of English and 100 years, with the goal of furthering our understanding of the relationship between spoken and written language.

The s-genitive (i.e. Canada’s government vs. the government of Canada) is on the rise in the 20th century (Hinrichs & Szmrecsanyi, 2007; Rosenbach, 2007). Statistical modeling confirms the press register leads this increase — a register change. Examination of internal linguistic constraints over time indicates simultaneous grammatical change, with the s-genitive increasing with certain inanimate subtypes.

The WH-forms (who, which) of the relative pronouns have become increasingly restricted to written registers (e.g. Romaine, 1982; Tottie, 1997a), leaving that as the variant used most for subject function in vernacular speech (D’Arcy & Tagliamonte 2010). Although who continues to
be used for animates, which is shown to have lost any grammatical conditioning that it once had and to be undergoing lexical replacement by that for non-human subject antecedents. Unlike the genitives, though, examination of internal linguistic factors reveals no evidence of grammatical change.

The methodology employed here provides a way to tease apart grammatical change from register change, with register-internal change shown to be a motivating factor in change from above. While the vernacular is “the most systematic data for our analysis of linguistic structure” (Labov, 1972a:208), it is not necessarily the most innovative, nor is it always the locus of change. With that in mind, this work provides a model of language change that integrates change across speech and writing.
Acknowledgments

It seems trite to start an acknowledgements page with this work would not have been possible without/I could never have done this without... Sometimes, though, there isn’t a better way. No work like this is done in a vacuum. It is built on the work of others, and fed by encouragement and feedback. There is positively no way this would have happened without the presence and involvement of a large number of people.

Sali Tagliamonte has been my supervisor for nearly ten years. When one is heading into a storm, having a force of nature behind you is a very good thing. When people congratulate me for finishing, I congratulate Sali. She deserves all the wine and a big bouquet of daffodils.

The other members of my supervisory committee, Jack Chambers and Naomi Nagy, have been beyond supportive at every step. Jack was the very first person I talked to in the Linguistics department when I came to U of T, and he has been encouraging me since that day. Jack has a unique ability to provide what seem at first like small, nominally good ideas that eventually become large and result in amazing things. For instance, it was Jack who first suggested Maclean’s magazine as a data source. Naomi Nagy joined the committee after the topic had been decided. She quickly became an important sounding board. She is also an excellent editor. If Jack didn’t catch it, she did. (And if neither of them caught it, that’s on me.)

Rebecca Roeder served on the committee in its initial stages, from 2007-2008. Becky was there when the decision was made that this would be the topic. I might have walked away before I started had it not been for her. The ‘internal-external’ members of the defence committee, Diane Massam and Alana Johns, gamely tackled the topic and brought much-needed theoretical perspective to the defence. The external examiner, Terttu Nevalainen, was spectacular, and provided the fresh eyes and perspective needed to remind me once again why I chose the topic in the first place. I am deeply grateful for her input on the final draft and her presence at the defence.

Portions of the empirical studies have been presented at several conferences over the years, including NWAV 2007 (Philadelphia), NWAV 2009 (Ottawa), AAACL 2009 (Edmonton), ISLE 2011 (Boston), LSA 2013 (Boston), and several meetings of the LVC research group at U of T. Every bit of feedback has helped to refine the work and keep me motivated.
At this point, I think I am going to have to resort to a list:

— My husband, Daniel Jankowski, for laughter, tea, PerlScripts and, well, everything else.
— My parents, Patty and Sandy Marburger, and my in-laws, Lori and Hans Jankowski.
— Many fellow graduate students with whom I’ve shared coffee, hotel rooms, knitting tips, feedback, rides, desks, endless marking parties or just general commiseration. I will surely forget more people than I remember, but here goes: Cathleen Waters, Kevin Heffernan, Eric Smith, Lydia Jarmasz, Leann Brown, Derek Denis, Kenji Oda, Liisa Duncan, Vanessa Hardy, Sarah Clarke, Manami Hirayama, Chiara Frigeni, Alex Motut, Annick Morin, Eugenia Suh, Julia Su, Marina Sherkina-Lieber, Catherine MacDonald, Ulyana Savchenko, Maria Kyriakaki, Monica Irimia, Sandrine Tailleur, Ailis Cournane, Richard Compton, Marisa Brook, Jim Smith, Daniel Currie Hall, Alex D’Arcy, Susana Bejar, Arsalan Kahnemuyipour, and anyone who has ever been one of my henchmen TAs.
— Other current and past professors: Keren Rice, Ron Smyth, Elan Dresher, Elizabeth Cowper, Tej Bhatia, Henry Rogers and Ed Burstynsky. The past and present administration staff of the department, without whom nothing could get done: Bill Forrest, Mary Hsu and Jill Given-King.
— The Mississauga Pops Concert Band, Eoforwic, Vest Yorvik, LSG, Hyde, and a few people who I only refer to by online screen names and have never met in-person. These outlets and people have long served as necessary reminders that there is far, far more to life than academia. Without them, I would have gone crazier years ago.

This work has recently been described as “exploratory.” This is not a term that was ever used while it was in-progress —admitting that a work like this is “exploratory” in nature at any point before completion is a frankly terrifying thought. It’s an accurate assessment, but more than that, I hope this work is a good beginning that can be built upon, critiqued and further refined. Any errors or omissions in interpretation, typos and grammatical infelicities remain completely my own. I look forward to hearing about them, but not for at least a month, if possible.

This is dedicated to authors and writers everywhere. Some day, you just might be part of my next data set. In the meantime, I have some fun reading to catch up on.

Bridget L. Jankowski
Toronto, Ontario/Butler, Pennsylvania
July 2013
# Table of Contents

Acknowledgments......................................................................................................................... iv

Table of Contents.......................................................................................................................... vi

List of Tables ........................................................................................................................................ ix

List of Figures ...................................................................................................................................... xi

List of Appendices ............................................................................................................................ xiii

Chapter 1 Introduction .................................................................................................................. 1

1 Situating the issue: Motivation .................................................................................................. 2

1.1 The place of written data in the study of language change ................................................. 2

1.2 Register variation and change ............................................................................................. 6

1.3 Grammatical (and social) variation and change and the variationist method ................. 13

1.4 The need to differentiate grammatical change from register change ............................... 15

Chapter 2 Diagnostics of grammatical and register change .................................................. 20

2 The variables ............................................................................................................................ 20

2.1 Genitive and relative clause variation ................................................................................. 20

2.2 The registers ......................................................................................................................... 24

2.2.1 Sample design and construction of the Maclean’s and Hansard corpora .................. 27

2.2.2 Limitations of the corpora ............................................................................................... 30

2.2.3 Advantages of the corpora ............................................................................................... 30

2.3 Evidence of grammatical change ......................................................................................... 32

2.4 Evidence of register change ................................................................................................. 35

2.4.1 Part 1: Cross-register comparison of linguistic constraints ......................................... 35

2.4.2 Part 2: register-internal constraints ................................................................................. 38

2.5 Summary ................................................................................................................................. 42

Chapter 3 Relative pronoun variation ......................................................................................... 44
3 Introduction ................................................................................................................................................. 44

3.1 The variable context ............................................................................................................................... 45

3.2 Internal factors: Animacy and grammatical number of antecedent .............................................. 53

3.3 Internal factors: Antecedent definiteness, NP vs. Pronoun, NP modification ............................... 58

3.4 Internal factors: Length and Complexity of Relative Clause, Adjacency .................................. 59

3.5 Register factors: Lexical density (type-token ratio) and Information Status (New vs. Given) .......................................................... 59

3.6 Multivariate analysis of grammatical factors: Subject relatives in Maclean’s and Hansard ............ 63

3.7 Recap and conclusions ......................................................................................................................... 69

Chapter 4 Genitive variation ....................................................................................................................... 72

4 Introduction ................................................................................................................................................... 72

4.1 Variable s– vs. of-genitive ...................................................................................................................... 72

4.2 Internal factors ......................................................................................................................................... 74

4.2.1 Interaction of internal factors .......................................................................................................... 77

4.2.2 Prototypical vs. non-prototypical possession and presence of final sibilant in the possessor ........... 83

4.3 Multivariate Analysis of grammatical factors: animate possessors ................................................. 85

4.4 Multivariate Analysis of grammatical factors: inanimate possessors ............................................. 89

4.5 Type-token ratio (lexical density) and Information status (New vs. Given) ................................. 92

4.6 Multivariate analysis of the register-internal factors ........................................................................ 95

4.7 Genitive variation in vernacular speech ............................................................................................. 99

Chapter 5 Recap, Implications, and Conclusions .................................................................................... 106

5 Recap of major findings ................................................................................................................................ 106

5.1 Lexical replacement vs. grammatical change .................................................................................. 107

5.2 Some implications of the major findings ......................................................................................... 108

5.3 Conclusions ............................................................................................................................................ 110
References...................................................................................................................................................... 113
Appendices.......................................................................................................................................................... 123
List of Tables

Table 1-1: Biber’s (1995:146) Register Dimension 1 (adapted) .......................................................... 11
Table 1-2: Co-occurring features for Biber’s (1995:142) Register Dimension 1 (adapted) ........ 11
Table 1-3: Hypothetical relative pronoun frequency in two registers ................................................. 16
Table 1-4: Hypothetical relative pronoun proportions in two registers .............................................. 17
Table 1-5: Distribution of relative pronoun which in restrictive relative clauses across two time periods in Register B ........................................................................................................... 17
Table 2-1: Sample for Maclean’s and Hansard corpora ................................................................. 28
Table 2-2: Contribution of factors to the probability of variant X of variable (XYZ) across time in speech (hypothetical) ........................................................................................................ 37
Table 2-3: Contribution of factors to the probability of variant X of variable (XYZ) across time in writing (hypothetical) ........................................................................................................... 37
Table 3-1: Overall distribution of relative clause pronouns, Maclean’s and Hansard .................. 46
Table 3-2: Possible relative pronouns by grammatical function ....................................................... 48
Table 3-3: Relative clause pronouns in Maclean’s by grammatical function .................................. 48
Table 3-4: Relative clause pronouns in Hansard by grammatical function ...................................... 49
Table 3-5: Animacy of antecedent in Maclean’s, Subject function ................................................. 55
Table 3-6: Animacy of antecedent in Hansard, Subject function ..................................................... 55
Table 3-7: Animacy of antecedent in Toronto, Subject function (adapted from D’Arcy & Tagliamonte, 2010:392) ......................................................................................................................... 56
Table 3-8: Multivariate analysis of grammatical factors for which with non-human antecedents 64
Table 3-9: Multivariate analysis of grammatical factors for which with non-human antecedents (Hansard).................................................................................................................................. 65

Table 3-10: Multivariate analysis of register factors for which with non-human antecedents (Maclean’s).................................................................................................................................. 66

Table 3-11: Multivariate analysis of grammatical factors for which with non-human antecedents (Hansard).................................................................................................................................. 67

Table 4-1: Multivariate analysis of grammatical factors for s-genitive with animate possessors (Maclean’s).................................................................................................................................. 86

Table 4-2: Multivariate analysis of grammatical factors for s-genitive with animate possessors (Hansard).................................................................................................................................. 87

Table 4-3: Multivariate analysis of grammatical factors for s-genitive with inanimate possessors (Maclean’s).................................................................................................................................. 90

Table 4-4: Multivariate analysis of grammatical factors for s-genitive with inanimate possessors (Hansard).................................................................................................................................. 91

Table 4-5: Multivariate analysis of register factors for s-genitive with animate possessors (Maclean’s).................................................................................................................................. 95

Table 4-6: Multivariate analysis of register factors for s-genitive with animate possessors (Hansard).................................................................................................................................. 96

Table 4-7: Multivariate analysis of register factors for s-genitive with inanimate possessors (Maclean’s).................................................................................................................................. 97

Table 4-8: Multivariate analysis of register factors for s-genitive with inanimate possessors (Hansard).................................................................................................................................. 98
List of Figures

Figure 3-1: Relative pronouns in Maclean’s over time, subject function................................. 49

Figure 3-2: Relative pronouns in Hansard over time, subject function................................. 50

Figure 3-3: Relative pronouns in Maclean’s over time, object function................................. 51

Figure 3-4: Relative pronouns in Hansard over time, object function................................. 51

Figure 3-5: Distribution of relative pronouns by Non-Subject Function over apparent time in Toronto (from D’Arcy & Tagliamonte, 2010:393).................................................................................. 52

Figure 3-6: Lexical Density in Maclean’s within 100-word window of relative which............ 62

Figure 3-7: Lexical Density in Hansard within 100-word window of relative which............ 62

Figure 3-8: Information Status in Maclean’s within previous 100-words of relative which....... 63

Figure 3-9: Information Status in Hansard within previous 100-words of relative which....... 63

Figure 4-1: s-genitive over time in Ontario English........................................................................ 73

Figure 4-2: Proportion of s-genitive by animacy and possessor length (Maclean’s)............... 76

Figure 4-3: Proportion of s-genitive by animacy and possessor length (Hansard).................... 77

Figure 4-4: Place-possessor NPs over time (s- and of-genitive)............................................. 78

Figure 4-5: Organization-possessor NPs over time (s- and of-genitive)................................. 78

Figure 4-6: One-word possessors over time (s-and of-genitive)............................................ 79

Figure 4-7: One-word possessors with s-genitive over time by animacy type.............................. 80

Figure 4-8: Proportion of s-genitive for animate categories by year (Maclean’s).................... 81

Figure 4-9: Proportion of s-genitive for animate categories by year (Hansard)....................... 81

Figure 4-10: Proportion of s-genitive for inanimate categories by year (Maclean’s)................ 82
Figure 4-11: Proportion of s-genitive for inanimate categories by year (Hansard)................. 82

Figure 4-12: Proportion of s-genitive by animacy and semantics of the possessor relationship (Maclean’s).................................................................................................................................. 84

Figure 4-13: Proportion of s-genitive by animacy and semantics of the possessor relationship (Hansard)........................................................................................................................................ 84

Figure 4-14: Proportion of s-genitive by final sibilant of possessor (Maclean’s)................. 85

Figure 4-15: Proportion of s-genitive by final sibilant of possessor (Hansard).................... 85

Figure 4-16: Proportion of s-genitive over time by lexical density of surrounding text (Maclean’s)........................................................................................................................................ 93

Figure 4-17: Proportion of s-genitive over time by lexical density of surrounding text (Hansard)........................................................................................................................................ 93

Figure 4-18: Information status of s-genitive possessor head noun (Maclean’s)................. 94

Figure 4-19: Information status of s-genitive possessor head noun (Hansard).................... 94

Figure 4-20: s-genitive with human + collective possessors over time............................... 100

Figure 4-21: s-genitive with human possessors over time...................................................... 101

Figure 4-22: s-genitive with inanimate possessors over time............................................... 102

Figure 4-23: Proportion of s-genitive by animacy in Ontario English (adapted from Tagliamonte & Jankowski, forthcoming) ......................................................................................................................... 102

Figure 4-24: s-genitive with inanimate possessors over time + Ontario Vernacular collectives (and organizations) (adapted from Figure 4-22 and 4-23)................................................................. 103
List of Appendices

Appendix A: Coding procedure for information status factor group ................................................. 123

Appendix B: Coding procedure for type-token ratio (lexical density) factor group, including sample part-of-speech tagged output. ........................................................................................................ 125

Appendix C: Further details of Maclean’s & Hansard corpora (Speakers, pertinent biographical information and word counts.) ........................................................................................................... 131

Appendix D: Common Coding instructions for Chapter 3 (Relatives) and Chapter 4 (Genitives) .............................................................................................................................................................................................. 135

Appendix E: Coding instructions specific to Relative Pronouns (Chapter 3) .......................... 137

Appendix F: Coding instructions specific to Genitives (Chapter 4) ............................................. 139
Chapter 1
Introduction

The relationship between spoken and written language is an understudied area in sociolinguistics. Guy & Bailey, for instance, in a study of relative pronoun variation using both written and spoken data, point out that “further research comparing formal and informal speech and formal and informal writing is necessary” (1995:150). Pintzuk’s (2003:525) concise summary of the widely accepted view of this topic in the field also directly acknowledges this:

Change originates in the spoken language, and historical linguists generally assume without comment that changes enter the written language in approximately the same order as they appear in speech, after some undetermined time lag. The assumption, therefore, is that the written language reflects the spoken language of some earlier time. This is not necessarily the case; future research comparing written and spoken modern languages may help to determine the chronology of linguistic change.

This work draws together three quantitative approaches to language variation and change. The larger theoretical goal is to better understand the relationship between spoken and written language, and the role of their intersection in the process of language change. Using variationist and historical sociolinguistics methods, alongside insights from corpus linguistics and studies of register variation, the aim is to investigate the effects of register on language change in a way that has not previously been done on a large scale, and to provide a baseline for future research tracking language change across spoken and written registers.

The major contribution of this work will be to properly identify potentially independent processes of variation and change at both the grammatical level and the register level, and the formalization and demonstration of a methodology to do so. This methodology relies on the simultaneous comparison between spoken and written language in order to account for and differentiate between grammatical change and register change. These two types of change can happen at the same time and so mask the effects of one another.

For the remainder of this chapter, I situate these issues within a discussion of the motivations for examining written language alongside spoken language. The terms grammatical change, register and register change are introduced in the context of a literature review of quantitative variationist sociolinguistics, corpus linguistics and register variation, and are further defined as they will be used in this work. Finally, a methodology that can test for both grammatical change
and register change within the same analytical framework is outlined, to be expanded upon and put into practice in later chapters.

The issue of needing to adequately differentiate between grammatical change and register change can only be fully explained after a discussion of the place of written data in the study of language variation and change, including the gap in the literature between studies using written data and those using spoken data. From this will follow a brief survey of the three quantitative approaches to this topic that I will draw on in the thesis: variationist and historical sociolinguistics, register variation, and corpus linguistics.

1 Situating the issue: Motivation

1.1 The place of written data in the study of language change

The comparative method of quantitative sociolinguistics (Poplack & Tagliamonte, 2001:88–102; Tagliamonte, 2002a), following the variationist tradition (Labov, 1966; inter alia), draws on comparative-historical linguistics in combination with quantitative analysis backed up by statistical modeling that tests hypotheses about language change. Most studies in this tradition deal with language variation and change in either written or spoken data, with the choice usually necessitated by the data available—if spoken data is available, it is preferred by far. As Labov (1972b:109) cautions, “Either our theories are about the language that ordinary people use on the street, […] or they are about nothing at all.”

In this tradition, the vernacular, or the style where conscious monitoring of speech is lowest, is considered to be “the most systematic data for our analysis of linguistic structure” and “where fundamental relations which determine the course of linguistic evolution can be seen most clearly” (Labov, 1972a:208). As this focus on vernacular speech is considered a core methodological tenet in modern sociolinguistic research, it is understandable why comparatively little quantitative sociolinguistic work has been done that attempts to elucidate the relationship between formal spoken language in which speech is more consciously monitored (to varying degrees dependent on the social context), various forms of written language, and the vernacular. However, as Romaine (1982:122) says:
If we accept Labov’s views [...] that texts[^1] can be understood only in terms of their relation to
the spoken language that ordinary people use on the street, then we must content ourselves with a
sociolinguistic theory which is very restricted in scope and application. And I would add that such
a sociolinguistic theory could not make any serious claims about being a theory of ‘language.’

Despite the assumption that the vernacular will always lead language change (Pintzuk, 2003,
above; see also Milroy & Milroy, 1999:47), there is evidence that a small number of changes
have entered the language from writing and subsequently spread to speech:

…the standard view—with which I basically agree—is that the primary locus of language change
is the spoken language, and that changes in the spoken language then seep into the written
language. Most language change is thus change from below, in Labov’s terminology, whereas
changes from above, induced by literacy or prescriptive teaching, are relatively infrequent.
However, they do happen, and are sometimes of paramount importance. (Tottie, 1997a:84)

One example is the wh-forms of the relative pronouns (who, whom, which…), which have been
shown to have entered the language through the influence of Latin grammar—something that
would have initially only been available to the literate speakers in the late middle ages—and
which then is thought to have spread to the spoken language of all speakers (Tottie, 1997a:84). In
fact, this prestigious “change from above” (Labov, 1994:78) never fully took hold in the
vernacular, remaining in competition with the older relative pronoun that and the Ø variant to
this day (Romaine, 1982:205–9, 212–14).[^2] The relative pronouns therefore represent an ideal
variable with which to study language variation and change across speech and writing, and will
be discussed further in this context in section 2.1 below.

Milroy & Milroy (1999:26–30) note that modern written language is inextricably linked to
prescriptivism; the rise of writing results in a need for standardization, which is achieved by
enforcing prescriptivism. What, then, can be gained by the intentional investigation of written
language and more closely monitored forms of speech when vernacular speech data is available
for the variety under investigation? Tottie (1997a:84) points out:

---

[^1]: Labov (1972b:100) writes that “The fundamental methodological fact that historical linguists have to
face is that they have no control over their data; texts are produced by a series of historical accidents…
the great art of the historical linguist is to make the best of this bad data—‘bad’ in the sense that it may be
fragmentary, corrupted or many times removed from the actual productions of native speakers.”

[^2]: Interestingly, Tottie (1997a:84–5) finds that despite the common modern prescriptive practice of editing
which to that in written documents—a 20th century prescriptive rule that interestingly reverses the
previously prestigious use of which—more literate speakers in more formal situations continue to lead the
use of who and which in restrictive subject relative clauses (1997a:91).
Most linguistic change and evolution took place before the advent of writing, and literacy cannot therefore have been a factor. It is also a fact that scholars who have written on language change have rarely dealt with the dichotomy of spoken and written language: the assumption has been that the spoken language is always the primary phenomenon […]. For quite a while, prescriptivism in language has been anathema to the serious linguist. Serious linguistics meant and means description, not prescription. But […] this does not mean that prescriptivism should not be studied as a factor in language use and change.

Further, changes such as the introduction of the WH-relative pronouns mentioned above, or the introduction of the Norman French of-genitive in Middle English as a competing construction to the older s-genitive (Barber, 1964:132–133; Jespersen, 1909-49:327–328) long predate the explosion of mass literacy and increased prescriptivism and standardization of the twentieth century. Biber also argues for speech and writing to be studied on the same footing, with both playing a part in our overall “communicative competence” of language (i.e. Hymes, 1972, 1974:75). Such communicative competence:

…includes formal knowledge of the range of speech-act variation, dialect variation, and register variation, as well as knowledge of when these different linguistic forms are appropriate; […] communicative competence is concerned with the form and use of all language—both speech and writing. Within this framework, neither speech nor writing needs to be considered primary to the exclusion of the other. Rather, both require analysis, and the linguistic comparison of the two modes becomes an important question. (Biber, 1988:8)

The practical reality is that the use of written language data in linguistic research is often unavoidable. Historical linguists, by necessity, use written data. Often the available texts are decidedly literary and not considered to be faithful representations of the vernacular. There are times when the amount of data needed in order to study a feature can make it difficult to obtain enough to study it in the vernacular — this is an issue with morphosyntactic variables in particular (Labov, 1972a:190–91).³ Romaine (1982:122) further argues:

I do not share Labov’s views about the relationship between the written and spoken language. I also do not accept that historical data can be bad in the last sense mentioned by Labov.⁴ […] the only way in which they can be bad [in that sense] is by invidious or inappropriate comparison with the spoken language.

³ For example, in Tagliamonte’s 2004 study of deontic modality (must, have to, have got to, got to), she found a total of 458 tokens out of 1.2 million words of vernacular speech from her York Corpus. In the 14 hours of transcribed recordings of natural speech from part 1 of the Santa Barbara Corpus of Spoken American English (Dubois, Chafe, Meyer & Thompson, 2000), only 71 tokens are found. In contrast, Jankowski (2004) found <1500 viable tokens of this variable across a corpus of 35 American English and 36 British English plays.

⁴ See note 1 above.
To this, I would add that if the vernacular “is the style from which every other style must be calibrated” (Tagliamonte, 2006:8, citing Labov, 1984:29), and the use of data representing styles other than the vernacular is often an unavoidable methodological necessity, then surely we should undertake investigation of non-vernacular language—including that from more formal speech (i.e. more closely monitored\(^5\)) as well as written sources—as rigorously as we do vernacular speech data, in order to provide an appropriate comparison, in Romaine’s terms, to vernacular speech.

In variationist theory, this argument speaks to Weinreich, Labov & Herzog’s constraints problem (1968:183–4), or “to determine the set of possible changes and possible conditions for change.” Variation between spoken and written forms of a single language is, by definition, stylistic variation. Further, “…not every combination of linguistic and social factors has been observed in studies to date […] In no case, for example, have we found a variable which originated as a social stereotype with stylistic stratification and at a later stage emerged as a social variable without stylistic shift” (1968:184). We should expect, then, that any change that is visible through studying variation in written language will have stylistic, or register correlates, that must be taken into consideration. As Romaine (1982:123–4) says:

> If linguistic variables are assumed to be embedded in both social and stylistic continua in a predictable way, then a framework which explicates this imbrication [overlapping] represents a means of uncovering social context in historical records; assuming, of course, that we can reconstruct a fully elaborated stylistic continuum, then we can speculate about its likely connections with the social continuum, and thus ‘reconstruct social context’ by a process of extrapolation.

Put very simply, we know that language changes. What we don’t know is why. In sociolinguistic theory, this is the actuation problem, the question of why “changes in a structural feature take place in a particular language at a given time, but not in other languages with the same feature, or in the same language at other times” (Weinreich et al., 1968:102).

J. Milroy refers to the actuation problem as “insoluble”, since “a solution to it implies the capacity to predict, not only what particular change will happen, but also when and where it will

\(^5\) Early variationist studies such as Labov (1972a: 208) and Trudgill (1974) used a model of style variation now described at the “Attention to Speech” model (Schilling-Estes 2002:378–9). Under this model, speakers are described as style-shifting in response the to level of attention they are paying to their own speech, on a continuum of casual (i.e. more vernacular) to careful (i.e. more formal) speech.
happen.” Yet he also notes that it is a worthwhile endeavour to “try to improve the accuracy of our predictions, and of course this greater accuracy includes the ability to specify the conditions under which something will not happen as well as the conditions under which it will happen” (1992:20–1).

Yet there exists a large gap in the literature; we see studies of language change using modern spoken data (e.g. Labov, 1963, 1966/1982; L. Milroy, 1987; Trudgill, 1974), studies that use much older written data (e.g. Nevalainen & Raumolin-Brunberg, 2003; Romaine, 1982), and longitudinal studies of a single written register (Biber & Finegan, 1989, 1992), but rarely are spoken and written data brought together unless the study is focused on register differences and not grammatical differences. This work aims to bridge that gap. While this does not solve the problem of trying to connect language change research where only written sources exist to current research where speech data is available, it will provide a framework for bridging this gap between spoken and written data in studies of ongoing language change in the 20th and 21st centuries.

1.2 Register variation and change

In light of these arguments for the need to examine both written and spoken data in studies of language variation, it is logical that, as Mair (2006a:203) suggests, comparison between written and spoken data would reveal contrastive stylistic and social conventions:

Shifting frequencies of grammatical constructions in corpus data need not necessarily point to changes in the underlying systems of grammatical choices, but may be symptoms of changes in genre conventions or communicative styles.

The term register is often used to refer to language as it is used in particular situational contexts (Crystal, 1991:295; Schilling-Estes, 2002:375), such as a speaker talking “more formally” to a co-worker in a job–related setting than they would to a friend in a pub. In historical and corpus

---

6 A rare exception to this is the work of Szmrecsanyi and Hinrichs (2008) and Hinrichs & Szmrecsanyi (2007), which uses corpora of written and spoken data in their analysis of both register and grammatical variation in the English genitive. Those studies will be discussed in further detail in Chapter 4, but it is worth noting here that the corpora used pre-existed the studies and were not designed to be specifically comparable in a sociolinguistically accountable manner.
linguistics, where written data is common due primarily to the availability of relatively permanent sources, as well as the lack of recorded and preserved speech until the last century, the terms register, genre and text type are often used interchangeably to refer to texts that serve different situational purposes and functions. For a brief summary of various researchers’ use of these terms, the reader is referred to Biber (1995:8–10).

I assume a definition of register as “any variety associated with particular situational contexts or purposes” (Biber, 1995:1). This includes any situational context in which language is used — spoken, signed or written. The speech of an informal sociolinguistic interview, for example, can be considered a single register. In distinguishing different text registers, which can vary dramatically in levels of generality (i.e. “journalistic prose” is a very general text register, while “obituaries” and “sports reportage” are more specific), I use the following guideline from Biber (1995:9–10):

‘…text categorizations made on the basis of external criteria relating to author/speaker purpose’ and ‘the text categories readily distinguished by mature speakers of a language; for example… novels, newspaper articles, editorials, academic articles, public speeches, radio broadcasts, and everyday conversations. These categories are defined primarily on the basis of external format’ (Biber, 1989:5–6). In practical terms, these categories are adopted because of their widespread use in computerized language corpora.

Such classifications are the basis of the register categories used in much corpus construction, and will help to form the basis of the data selection and corpus construction of the present work, to be discussed in Chapter 2. The BROWN family of corpora remain a standard for parallel, comparable text corpora, and are still regularly used and cited today by corpus linguists studying language change (e.g. Hinrichs & Szmrecsanyi, 2007; Hundt, 2007; Krug, 2000; Leech, 2003; Leech & Smith, 2005; Mair, 2006a; Szmrecsanyi & Hinrichs, 2008;). The BROWN and LOB corpora comprise data from 1961 American (Francis & Kucera, 1979) and British (Johansson, Leech & Goodluck, 1978) written English, respectively. FROWN and FLOB are comparable, matching corpora compiled from 1991 (Hundt, Sand & Skandera, 1999; Mair, 1997). All four

---

7 It is true that speech style can differ dramatically across a single interview. Taking this into account, multiple sub-registers could be defined to distinguish differences in style or formality. This does not negate the larger register category as described here, as long as the interview as a whole can be considered the product of a single “situational context or purpose” (Biber, 1995:1).

8 The BROWN corpora family, as well as the London-Lund corpus and many others, are distributed electronically by ICAME: http://khnt.hit.uib.no/icame/.
corpora follow the same format: 500 text excerpts of 2000 words each from fifteen registers, for a total of one million words each.

Register categories in the BROWN-family corpora include Press (reportage, editorial and theatre/book/music reviews), Religion, Skills and hobbies, Popular lore, Biographies and memoirs (including a set of letters), Miscellaneous (including government documents and industry reports), Academic or “learned” documents, and several separate categories of “imaginative prose” or Fiction: general, mystery and detective, science fiction, adventure and western, romance, and humour (Francis & Kucera, 1979). It should be noted, though, that the texts included in the BROWN family of corpora were chosen by random sampling. No attempt was made to create a balanced sample design in terms of the social characteristics of the authors, such as gender, age or birthplace, as the original BROWN corpus was not created for sociolinguistic research with the intent to take such features into consideration. However, the register categories of these corpora will be shown to be relevant as a starting point for pinpointing registers for use in the current work (Section 2.2, below).

Spoken language can also be considered a further set of available registers, with the communicative function of the register determined by the situational context of the data. Popularly-cited corpora that include both spoken and written components include the International Corpus of English (ICE) (see Greenbaum & Nelson, 2009; Nelson, Wallis & Aarts, 2002, as well as studies by Krug, 2000; Rosenbach, 2002, 2003, 2005 among others), and the million word London-Lund corpus (Greenbaum & Svartvik, 1990; see studies by Krug, 2000; Mair, 2006a; Tottie, 1997a; among many others). The spoken categories of the ICE corpora

---

9. Text corpora designed to take sociolinguistic factors into account do exist, however they are more the exception than the norm. The Corpus of Early English Correspondence (Nevalainen & Raumolin-Brunberg, 2003) is one such example. This will be discussed further in Chapter 2.

10. It would be remiss to not also mention the enormous British National Corpus (BNC Consortium 2007; Burnard, 2007), and the American National Corpus (ANC; Ide & Suderman, 2004). Each will ultimately contain 1,000,000 words of digitized, searchable written and spoken English. The ANC is still in-progress, and is not yet considered “balanced.” Though early releases have begun to be used for research, the primary use of both the BNC and ANC in corpus linguistic research so far have been to retrieve raw frequency data for lexical items and grammatical constructions. Canadian English is represented by the newly constituted Canadian English ICE, which can be used for parallel analysis of written registers and speech, as well as by part-of-speech/lexically tagged or parsed/syntactically tagged corpora. (http://ice-corpora.net/ice/icecan.htm). These remain a largely untapped resource in variationist sociolinguistics.
include Face-to-face conversations, Phone calls, Classroom Lessons, Broadcast Discussions, Broadcast Interviews, Parliamentary Debates, Legal cross-examinations, Business Transactions, Spontaneous commentaries, Unscripted Speeches, Demonstrations, Legal Presentations, Broadcast News, Broadcast Talks and Non-broadcast Talks.

Scholars have continued to refine how best to classify texts to achieve a representative sample of registers when compiling large corpora for use in linguistic research. Görlach (2001, 2004) has compiled lists of well over 2000 text types, analyzing them in terms of semantic features that categorize and describe their communicative functions, including such factors as the topic (i.e. legal, literature, religion, politics, declarations, directives) or intention (i.e. amuse, affirm, codify, inform, publicize, remind, teach) of the text (Görlach, 2002:17). He makes a further case (Görlach, 2004:104) for revisiting these commonly occurring categories used in corpus construction:

Some scholars have argued that the ‘established’ categories and their proportional representation should be retained for easier comparison; but others have insisted that we need a less impressionistic basis to achieve a truly representative corpus.

Text functions “develop with changes in communicational intentions, the media, educational facilities and other social factors” (Görlach, 2002:18). How, then, does one quantify the notions of more (or less) formal, or “changes in communicational intentions” (Görlach, 2002:18)? What is also needed is a way to define register change in quantifiable, internal, linguistic terms, in order to pinpoint register that will become appropriate diagnostics of language change at several points along a continuum of formality. Görlach (2004:104) suggests that “the most promising linguistic approach to the problem is probably Biber’s (1988).”

However, due to the enormity of their sizes, the time needed to mine these corpora — given the requirement in variationist studies for balanced, parallel samples of several registers where sociolinguistic information for speakers and authors of all texts is available — is comparable to the time required to create a smaller corpus, where such factors can be controlled for as data is gathered.

Görlach’s (2001, 2004) corpora come from Old, Middle and Modern English, using data from a large variety of sources ranging from bibles to grammar books to scholarly prose to recipes, letters and anything else available across the history of English. For scholars who compile corpora, Görlach (2001, 2002, 2004) remain the citations of choice for text classification, since the breadth of his research into the history of English, English text classification and stylistics is so large.
Biber (1988, 1995) and Biber & Finegan’s (1989, 1992) *Multi-Dimensional* (short: MD) approach has provided a more linguistic-internal approach to register variation and change. As will be shown shortly, this model is not an appropriate methodology for getting at the independent processes of variation and change at both the *grammatical* level and the *register* level. However, it will provide a necessary guide (see Chapter 2) for choosing appropriate diagnostics for the empirical studies in Chapter 3 and Chapter 4, and so a brief synopsis is provided here.

*Register* is defined by Biber (1995:1) as “any variety associated with particular situational contexts or purposes. Registers are demonstrated to be linked to the co-occurrence of specific lexical and grammatical features, in which “different co-occurrence patterns are analyzed as underlying *dimensions* of variation. [...] the co-occurrence patterns comprising each dimension are identified quantitatively, rather than on an *a priori* functional basis” (Biber, 1995:30).

Biber (1995, ch. 6 141–69, for detailed methodology see Biber, 1988 ch. 4–7) provides several “dimensions” along which language use can be distinguished according to register. The first of these, *Involved versus Informational Production*, relates to the amount of “informational focus and a careful integration of information in texts” (Biber, 1995:141). Linguistic features associated with the “involved” end of the cline are indicative of language production under real-time processing constraints, while linguistic features associated with the “informational” end of the cline include those which allow for more careful production with possibly time allowed for editing (Biber, 1995:117). The cline representing Biber’s Dimension 1 is shown below in Table 1-1. The list of the co-occurring features associated with this dimension is shown in Table 1-2.

The co-occurrence of the following set of linguistic features in Table 1-2 is considered to be indicative of the “informational” end of the cline shown on Table 1-1: nouns, (longer) word length, prepositions, (higher) type-token ratio,\(^\text{12}\) attributive adjectives, place adverbials, agentless passives and past participial pronominal clauses. The register of press reportage falls at this end of the cline. The register of face-to-face conversations has a low probability of these features co-occurring, and so falls at the “involved” end of the cline (Biber, 1995:142–46). In other words, while face-to-face conversations certainly make use of nouns, prepositions, attributive adjectives,\[\]

\(^{12}\) This feature will be discussed in section 2.4.2.
Table 1-1: Biber’s (1995:146) Register Dimension 1 (adapted)

<table>
<thead>
<tr>
<th>INVOLVED</th>
<th>TELEPHONE CONVERSATIONS</th>
<th>35</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FACE-TO-FACE CONVERSATIONS</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Personal Letters</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>SPONTANEOUS SPEECHES</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>PUBLIC CONVERSATIONS</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Romance fiction</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PREPARED SPEECHES</td>
<td>-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-10</td>
</tr>
<tr>
<td></td>
<td>BROADCASTS</td>
<td>-15</td>
</tr>
<tr>
<td></td>
<td>Religion</td>
<td>-20</td>
</tr>
<tr>
<td></td>
<td>Humor</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Popular lore; Editorials; Hobbies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Biographies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Press reviews</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Academic prose; Press reportage</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Official documents</td>
<td></td>
</tr>
</tbody>
</table>

Table 1-2: Co-occurring features for Biber’s (1995:142) Register Dimension 1 (adapted)

Co-occurring linguistic features on English dimension 1: ‘Involved vs. Informational Production.’ (Features in parentheses have lower weights and are not used in the computation of dimension scores)

<table>
<thead>
<tr>
<th>Dimension 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Involved Production’</td>
</tr>
<tr>
<td>Positive features:</td>
</tr>
<tr>
<td>Private verbs</td>
</tr>
<tr>
<td>THAT deletion</td>
</tr>
<tr>
<td>Contractions</td>
</tr>
<tr>
<td>Present tense verbs</td>
</tr>
<tr>
<td>Second person pronouns</td>
</tr>
<tr>
<td>DO as a pro-verb</td>
</tr>
<tr>
<td>Analytic negation</td>
</tr>
<tr>
<td>Demonstrative pronouns</td>
</tr>
<tr>
<td>General emphatics</td>
</tr>
<tr>
<td>First-person pronouns</td>
</tr>
<tr>
<td>Pronoun IT</td>
</tr>
<tr>
<td>BE as main verb</td>
</tr>
<tr>
<td>Causative subordination</td>
</tr>
<tr>
<td>Discourse particles</td>
</tr>
<tr>
<td>Indefinite pronouns</td>
</tr>
<tr>
<td>General hedges</td>
</tr>
<tr>
<td>Amplifiers</td>
</tr>
<tr>
<td>Sentence relatives</td>
</tr>
<tr>
<td>WH questions</td>
</tr>
<tr>
<td>Possibility modals</td>
</tr>
<tr>
<td>Non-phrasal co-ordination</td>
</tr>
<tr>
<td>WH clauses</td>
</tr>
<tr>
<td>Final prepositions</td>
</tr>
<tr>
<td>(Adverbs)</td>
</tr>
</tbody>
</table>

| ‘Informational Production’ |
| Negative features: |
| Nouns | -0.80 |
| Word length | -0.58 |
| Prepositions | -0.54 |
| Type–token ratio | -0.54 |
| Attribute adjectives | -0.47 |
| (Place adverbials | -0.42 |
| (Agentless passives | -0.39 |
| (Past participial post nominal clauses | -0.38 |

etc., and can include long words and an occasionally high type-token ratio, it is the higher probability of co-occurrence and shared variance for the set of these features together that is indicative of the “informational” end of Dimension 1. There is a higher probability of co-occurrence for these features in press-reportage than in face-to-face conversation.
Using such sets of co-occurring features, Biber (1994:35–6) was able to describe a number of register “dimensions” — continua along which various registers, both written and spoken, align:

The linguistic content of a dimension comprises a group of linguistic features (e.g., nominalizations, prepositional phrases, attributive adjectives) that co-occur with a markedly high frequency in texts. Based on the assumption that co-occurrence reflects shared function, these co-occurrence patterns are interpreted in terms of the situational, social, and cognitive functions most widely shared by the linguistic features.”

A register change within this model is when the co-occurring groups of linguistic features that are used in a particular situational context alter, either by the increase or decrease in frequency of one or more of those features (Biber & Finegan, 1989, 1992). Referring to Table 1-2 above, for example, a register change toward a “more informational” production could involve, in part, a decrease in frequency of longer NPs in the Press Reportage register, causing it to fall at a different point on the cline.

However, while Biber’s model contains a quantifiably testable definition of register change, this model does not take into account that the grammatical function of individual linguistic forms is not static, nor is there always a one-to-one relationship between linguistic form and function:

The primary analytical goal of the MD approach is to identify the basic parameters of variation [n.b. across register] in a language, not to provide detailed functional analyses of the individual features comprising a dimension. As a result, individual features can have particular functions [14] not incorporated into the shared basis of a dimension. (Biber, 1995:135–6)

For example, in Table 1-2 above, some of the co-occurring linguistic features indicative of the “involved” end of Dimension 1 are discourse particles. These are categories with a large number of lexical items serving a variety of grammatical functions, including but not limited to hedges (e.g. kind of, sort of, something like, also called ‘general extenders’), downtoners (e.g. partially, barely, hardly) and emphatics (e.g. really, very, so, also called as ‘intensifiers’) (Biber, 1988: 240–1; Biber, 1995:96, 136–7). Items within these categories could further be serving multiple

---

13 Biber (1988, 1995) calls these sets, or groups, of co-occurring features “factors”, and refers to his multi-dimensional style of analysis as a “factor analysis”. Since the variationist tradition uses the term “factor” in a different way from this, I have not used the term here.

14 Biber (1995:136) distinguishes four separate notions of function. These include the discourse or grammatical function; the “situational” function (if a form carries information about the time and place of the discourse, or speaker and interlocutor presence and participation); the effect of processing constraints during production; and other sociolinguistic functions.
sociolinguistic functions. For instance, both intensifiers and general extenders are known to have numerous sociolinguistic correlates, such as being indicative of vernacular speech as opposed to formal speech or writing, and with particular lexical items marking speaker age and gender (Tagliamonte, 2008; Tagliamonte & Denis, 2010).

Additionally, and most importantly, individual linguistic features can **take on new functions over time**:

...linguistic forms gradually shift from one function to another. This trajectory can be viewed in the varying strength and distribution of independent linguistic features associated with one of the evolving morphemes. ... Practically speaking, such changes should be visible in an ordered series of shifts in factor weights. (Tagliamonte, 2002a:747–48, citing Labov, 1982:76)

This is a different type of change, namely *grammatical change*. In order for grammatical change to be identified within the MD approach, the “dimensions” would have to be recalculated for each time period under investigation, and compared to modern register dimensions. Differences from one time period to another in the sets of features that comprise a particular dimension would be an indication that one or more features within the set could have changed their grammatical function in some way. Biber’s MD approach was never intended to provide an analysis of how *individual* linguistic features change over time with respect to their grammatical or social functions. The variationist method, introduced in the following section, is more suitable to such an analysis. We will return to Biber’s register dimension 1 in Chapter 2, where it will be used to identify diagnostic variables and registers for study within variationist sociolinguistic methodology.

### 1.3 Grammatical (and social) variation and change and the variationist method

The variationist method of studying language variation and change requires the identification of a linguistic variable of which there is more than one form, or **variant**. Variants are, in short, “two or more ways of saying the same thing” (Labov, 1972a: xiv; G. Sankoff, 1980:55), or forms [that] perform the same function in the grammar. Further:

In theory, no two forms can have identical meaning, but in practice two different forms can be used interchangeably in some contexts even though they may have distinct referential meanings in other contexts. In fact, you are dealing with at least two different levels of meaning: 1) comprehensive meaning, which takes into consideration every possible inference; and 2) meaning as it is used in the speech community. While the first is subject to idiosyncratic interpretation and an infinite range of potential meanings, the second is by definition a consensus that is shared and relatively constant. (Tagliamonte, 2006:73–4)
All instances of the variable are identified, and each instance is counted and analyzed according to the internal linguistic environments and external social conditions that could potentially contribute to the choice of which variant is used. The way in which the linguistic and social environments condition the use of the variable can be viewed as a “constraint hierarchy” (Poplack & Tagliamonte, 2001:94):

The classic examples are constraint hierarchies for the expression of certain allophones (or the application of optional phonological and morphophonological rules), but it is also true of syntax, in the study of variable rule order, optional movement or deletion rules, and in preferences among semantically or functionally equivalent phrase structures. Moreover, it is these variable aspects of grammar which are always the locus of linguistic change. (D. Sankoff, 1988a:146–147)

Labov (1982) and D. Sankoff (1988a) lay down methodological considerations for the study of linguistic change in the variationist framework, wherein change is reflected in an “ordered series of shifts in underlying probabilities associated with each environmental factor” that condition linguistic variation (Labov, 1982:75). Internal linguistic factors that condition a variable in a probabilistic manner are representative of the function of that variable in the grammatical system. Changing probabilistic effects of internal factors on a variable can be taken to represent diagnostics of grammatical change. In the comparative sociolinguistic method, grammatical change is indicated by the changing significance, magnitude of effect, and constraint hierarchies of internal linguistic factors in a multivariate analysis (Tagliamonte, 2002a:731).

The primary analytic tool of the variationist method is statistical analysis, typically log-linear regression using the variable rule program (Rand & Sankoff, 1991; D. Sankoff, 1978, 1985, 1988b; Sankoff, Tagliamonte & Smith, 2005). Using this analytic method, changing constraints (i.e. the direction of the effects of internal and external factors that condition the observed variation) compared across time in a single language variety can be used to infer language change. When compared across related varieties of a language, they can provide evidence of underlying similarities and differences in the grammar that are not evident simply by comparing overall rates of use (Poplack, 2000).

The variationist method, originally used to study sound change (Labov, 1963, 1966) is equally effective for the study of syntactic, morphosyntactic and grammatical change (see Bayley, 2002; Green, 2007; Labov, 1982, 1994, 2001; Pintzuk, 1999; Poplack & Tagliamonte, 2001; D. Sankoff, 1988a; Sankoff & Thibault, 1981; Tagliamonte, 2002a, 2006, 2012; Young & Bayley, 1996). This methodology will be further elaborated in Chapter 2, with reference to and
predictions for the variables that will serve as diagnostics in the present study (see Chapter 2.1, below).

1.4 The need to differentiate grammatical change from register change

Grammatical change is something that we expect to happen to language no matter the register. By using Biber’s linguistically-defined register dimensions to delineate the focus of a study of register change, we must acknowledge that some of the individual linguistic features that comprise a co-occurrence set indicative of a register dimension could themselves be implicated in grammatical change. Some of these linguistic features are variants of known sociolinguistic variables: for example, the wh-forms of the relative pronouns are counted as a separate linguistic feature from the non-wh-forms in the MD approach, because that-relatives and wh-relatives occur at significantly different frequencies in speech and writing. That means their individual patterns of co-variance with other linguistic features is a more appropriate diagnostic of different points along a register dimension such as that shown in (1a) (Biber, 1988:234) than if “relative pronouns” were considered as a whole.

This is very different from variationist methodology, as described in 1.3 above, where the study of a single linguistic variable as an indicator of grammatical change must include all variants of that variable in every case where it occurs as well as those where it could have occurred and did not (Labov, 1972a:72). Biber acknowledges the importance of the variationist method for elucidating the “chronology of linguistic change” mentioned by Pintzuk (2003:525, see above):

Romaine’s (1980, 1982) ‘sociohistorical’ approach shows how structural changes enter a language in particular registers and subsequently evolve at different rates in different registers […] based on the relative frequency of forms across registers from different historical periods. […] these empirical studies are important for their analysis of the interplay between language change and register variation… (Biber, 1995:13–14)

However, we must further take into account the internal linguistic conditions, or constraints, that are in operation on those variables, and which serve as diagnostics of grammatical change. Bauer (2002:103) describes another relevant example for the English progressive:
Mair (1998:148) reports that an increase in progressives between Brown and Frown (and also between LOB and FLOB) is due to the more frequent choice of an informal option rather than a non-progressive formal option in places where either is possible. While this might be a language change, it might equally be viewed as a societal change in perception of formality or as no change at all, just a different exploitation of precisely the same system.

The crux of the issue is that variationist sociolinguistics, corpus linguistics and MD studies of register variation do not use “frequency” measurements in the same way. Studies such as Mair 1998 count raw frequencies of a variable (i.e. number of occurrences per 1000 words), a long-held methodology in historical corpus linguistic research. Szmrecsanyi points out that “…in the corpus-based historical linguistics community, we seem to be dealing with a deeply entrenched reliance on the diagnostic power of corpus frequencies” (2013:2), and that “…text frequencies are a regrettably unreliable and inconclusive diagnostic of grammar change: they are at best inconclusive, and at worst misleading” (2013:14).

As a simple hypothetical example, the variants of the relative clause pronouns can include that, who, which and Ø. If only the frequency of that is counted as in Table 1-3, and it appears to differ dramatically between two registers, we might say at first glance that Register B uses that nearly twice as often as a relative clause pronoun:

<table>
<thead>
<tr>
<th>Relative Pronoun</th>
<th>Register A</th>
<th>Register B</th>
</tr>
</thead>
<tbody>
<tr>
<td>that</td>
<td>134</td>
<td>259</td>
</tr>
<tr>
<td>who</td>
<td>109</td>
<td>182</td>
</tr>
<tr>
<td>Ø</td>
<td>54</td>
<td>160</td>
</tr>
<tr>
<td>which</td>
<td>32</td>
<td>240</td>
</tr>
<tr>
<td><strong>Total N</strong></td>
<td><strong>347</strong></td>
<td><strong>853</strong></td>
</tr>
</tbody>
</table>

This is only part of the explanation. If it is the case that Register B simply has more relative pronoun contexts overall than the other registers, then it is possible that when the proportion of that in all relative pronoun contexts is compared between them, the picture will be very different.

---

15 Variation in restrictive relative clauses is the subject of Chapter 4. As such, I will not discuss the details of circumscribing the variable context, or the relevance of the internal factor of antecedent adjacency here, since these examples are provided here only to illustrate the methodological underpinnings of the comparative method.
In Table 1-4, we see that use of *that* in Register A is actually nearly the same rate across both registers: 38.5 percent in Register A vs. 30 percent in Register B. In this example, then, the rate of use of *that* is in fact slightly higher in Register A. In order to compare across varieties (in this case, two text registers) in an accountable way, we must compare frequencies proportionally (Poplack, 2000:14–15).

**Table 1-4: Hypothetical relative pronoun proportions in two registers**

<table>
<thead>
<tr>
<th>Relative marker</th>
<th>Register A</th>
<th></th>
<th>Register B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td><em>that</em></td>
<td>38.5</td>
<td>134</td>
<td>30</td>
<td>259</td>
</tr>
<tr>
<td><em>who</em></td>
<td>31</td>
<td>109</td>
<td>21</td>
<td>182</td>
</tr>
<tr>
<td>Ø</td>
<td>15.5</td>
<td>54</td>
<td>19</td>
<td>160</td>
</tr>
<tr>
<td><em>which</em></td>
<td>9</td>
<td>32</td>
<td>28</td>
<td>240</td>
</tr>
<tr>
<td><strong>Total N</strong></td>
<td>29</td>
<td>347</td>
<td>71</td>
<td>853</td>
</tr>
</tbody>
</table>

Further, as shown in Table 1-5, the internal linguistic constraints conditioning the use of a variable can remain constant despite dramatic differences in overall frequency or proportion. Despite the widely different frequency counts between the two times (16 instances of adjacent restrictive relative *which* in Time 1 vs. 79 instances in Time 2), and the difference in the proportions (33 percent adjacent and 62 percent non-adjacent in Time 1 vs. 93 percent adjacent and 100 percent non-adjacent in Time 2), the *direction of the effect* — namely, that restrictive relativizer *which* is more likely to occur with a non-adjacent antecedent — is the same in each time period. This is the diagnostic of the *constraint hierarchy* described in section 1.3 above, or the probabilistic effect of internal constraints on a variable.

**Table 1-5: Distribution of relative pronoun *which* in restrictive relative clauses across two time periods in Register B.**

<table>
<thead>
<tr>
<th>Adjacency of antecedent to relative pronoun</th>
<th>Time 1</th>
<th></th>
<th>Time 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td><em>Adjacent</em></td>
<td>33</td>
<td>16</td>
<td>93</td>
<td>79</td>
</tr>
<tr>
<td><em>Not adjacent</em></td>
<td>62</td>
<td>5</td>
<td>100</td>
<td>8</td>
</tr>
<tr>
<td><strong>Total % which</strong></td>
<td>38%</td>
<td>56</td>
<td>94%</td>
<td>93</td>
</tr>
<tr>
<td><strong>Total N (that + which)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Frequency differences like those shown in Tables 1-3–1-5 are often evident in cross-register variation or change over time. But it is the probabilistic effect of the constraints on the grammar that provide a way to distinguish underlying grammatical differences between varieties.

What is needed is a method for determining whether an increase or decrease in the use of a particular variant of a variable within a single register is due to an ongoing grammatical change, or a change in the way that register is used. In short, we need a way to determine if the register is changing, or if the underlying grammar is changing with respect to a particular variant. I propose that by making use of the linguistically-defined register dimensions of Biber 1988/1995 to pinpoint appropriate registers along a single dimension, as well as using previous corpus linguistic and variationist research to identify appropriately diagnostic variables within those chosen registers — variables that are known to vary by register and are possibly also implicated in grammatical change — it is possible to differentiate grammatical change and register change within a single analytical framework, namely the comparative method of variationist sociolinguistics described above in section 1.3. Using this methodology, internal linguistic constraints, which can be used as diagnostics of grammatical change, can be compared for a single variable across multiple registers.¹⁶ This is very much in line with the suggestion of Szmrecsanyi (2013:2), though he uses the term environment(al) rather than register:

…fluctuating frequencies of grammatical variants are a function not only of changing grammars, but are also conditioned by environmental changes in the textual habitat. So the crucial problem is that diachronically variable text frequencies often entangle environmental differences and grammatical changes. To disentangle the two types of change, we will argue that instead of focussing solely on text frequencies (how often do language users use some linguistic variant?), analysts need to explore the possibly historically evolving probabilistic conditioning of variants (why do language users use the variants that they use?). This approach yields a more reliable diagnostic of grammar change.

Writing from the historical corpus linguistic perspective, Szmrecsanyi (2013:2) continues:

…variationist sociolinguists, for example, have been long aware that language change may manifest itself in extremely subtle shifts in the stochastic effects of conditioning factors, and that mere text frequencies (or variant rates, for that matter) may be as much about culture as they are about language.

¹⁶ This methodology will be fleshed out further, with practical examples, in Chapter 2.
While this is true, I would argue that variationist sociolinguistics also currently lacks something crucial; here I return once more to the “gap in the literature” referred to above in Section 1.1, namely the hesitancy to use written data unless it is a last resort. Szmrecsanyi (2013:5) sums this up well:

…the long-term historical record that we have in historical linguistics at the present time does not document speech. In the absence of historical records of face-to-face interaction, however, the dynamics of written registers (read: environmental changes affecting the textual habitat) are a serious confounding […] Again, we stress that this problem is in principle well-known (see, e.g., Biber and Finegan 1989; Biber and Conrad 2001; Hundt and Mair 1999). Yet, there has been a tendency to shrug off the problem, and/or to concede defeat by accepting that trying to disentangle grammar change from environmental change (such as dynamically evolving written text types) is a hopeless endeavor.

By comparing changing grammatical constraints on a variable across multiple points in time as well as in multiple registers, using both written and spoken data in the same analysis, we increase our ability to predict the effect of stylistic variation on language variation and change. This is particularly crucial in establishing whether a change is happening above or below the conscious level for speakers of the language. Labov (1994:79) notes that changes from below are generally “systematic changes that appear first in the vernacular and represent the operation of internal, linguistic factors.” Changes found first in written registers (or registers that require less on-line processing time and allow for some editing during production) are, as noted previously, likely the result of prescriptivism or literacy, and would represent a change from above. Further, by holding all the data used, written and spoken, to sociolinguistically accountable standards, our cross-comparisons can be considered more sound.

Finally, we should note that variationist analyses traditionally do not incorporate the use of text frequency measures in the same analysis as internal linguistic constraints — text frequency remains under the purview of historical corpus linguists. As we will see, not only is it possible to combine such measures into the variationist framework, it is a useful tool in disentangling grammatical and register change.

---

17 This issue will be elaborated on further in Chapter 2, but see the discussion above in section 1.2 about the lack of sociolinguistic accountability in the BROWN family of corpora, and other widely-used historical text corpora.
Chapter 2
Diagnostics of grammatical and register change

In this chapter, I motivate the use of three language registers (two text and one spoken) and two linguistic variables that will serve as diagnostics for investigating the relationship between speech and writing in the course of language change. I will employ comparative, quantitative variationist methods, and demonstrate the need to tease apart grammatical and register change, and the difficulties with doing this. The methodology is laid out in detail here, including hypotheses presented about the diagnostic variables as they relate to this issue. The corpora created for the study are described, as is the sample design of data from those corpora.

2 The variables

It perhaps goes without saying; a crucial starting point is that the variables chosen for study must be appropriately diagnostic of the issues under investigation. Since we are looking to distinguish grammatical change and register change, chosen variables should be 1) known to vary by register and 2) implicated in grammatical change. Two variables will suffice to demonstrate the issues: one that can be shown to be undergoing either grammatical or register change, but not both, and one that can be shown to be undergoing both grammatical and register change.

2.1 Genitive and relative clause variation

Two variables were introduced in section 1.1. The first is the older synthetic “Saxon” s-genitive and the analytic “Norman” of-genitive in Modern English, as in (1) and (2):

(1)  a. …the continued callous indifference of the federal government. [Hansard 1956/u]
    b. The federal government’s environmental plan… [Hansard 1956/u]
(2)  a. …use it to house Canada’s first responsible government… [Maclean’s 1956/h]
    b. …doing a work of national importance and polishing the treasures of Canada.
        [Maclean’s 1956/h]

The variation between these forms is well-documented (Barber, 1964:132–133; Jespersen, 1909-49:327–328; Mair, 2006a:144–48). Moreover, the continued use of the s-genitive in the 20th century has been widely studied (see, for example, Hinrichs & Szmrecsanyi, 2007; Ljung, 1997;
Rosenbach, 2002, 2003, 2005; Szmrecsanyi & Hinrichs, 2008). Variation between these forms is also known to be affected by speech style or register, an observation made in the early 20th century by such grammarians as Jespersen and Poutsma (Ljung, 1997:21). Notably for this investigation, journalistic prose appears to be a locus of the continuing increase of the s-genitive in standard English text registers, and further its use is claimed to be spreading to inanimate subjects (Hinrichs & Szmrecsanyi, 2007; Mair, 2006a).

A second promising variable is variation among the restrictive relative clause pronouns. The productive variants in modern English are that, who, which, and Ø, as in (3):

(3) a. …if there's one man on earth that ought to be paid on a preference, it's the man who prints a newspaper. [Maclean’s 1906/a]
   b. May I ask if this $3,000 that has been voted year after year is to be given to Dr. Bell or whether the $3,000 which is asked to be voted on at the present time is to go Dr. Bell? [Hansard 1906/A]
   c. Sir Wilfrid is not above adding to it the red necktie and white vest which Sir John Macdonald so often favored [Maclean’s 1906/b]
   …his approaches to solving the real political problems they raise remain vague. [Maclean’s 2006/Ø]
   d. to give some attention to the remarks that have been made […] I should like to add something to the remarks Ø I have already made [Hansard 1956/N]

Variationist studies (for example, D’Arcy & Tagliamonte, 2010 for Canadian English, Tagliamonte, Smith & Lawrence, 2005 for British English) have shown that the wh-forms of the pronoun, which were late additions to the relativizer system from Norman French, were never fully incorporated into vernacular speech, where the variants used to this day are still primarily that and Ø. At the same time, written registers make ample use of the WH-forms (Ball, 1996; Montgomery, 1989; Romaine, 1982; Sigley, 1997; Tottie 1997a). Results from 20th century Canadian English (D’Arcy & Tagliamonte, 2010; Jankowski 2009) suggest that which—almost non-existent in 20th century vernacular speech, but still used in written registers—was once favoured over that for non-human antecedents in subject relatives. It has begun to lose ground to that for this function in the 20th century, while who continues to be used for animate subject relatives.

With the target variables identified, we must establish whether each one is undergoing change, and what type. For both of these variables, we can posit at least two scenarios for any change that
might be taking place. For example, there could be lexical replacement happening. For the relative pronouns, this would mean that which is simply disappearing in all linguistic contexts, though possibly at different rates in different registers. For the genitives, the s-genitive could be on the way to completely replacing the of-genitive. It is also possible that deeper grammatical changes are in progress, whereby which (or any of the relative pronouns, or the s- or of-genitive), are undergoing grammaticalization and becoming restricted to particular linguistic contexts, or taking on new functions in the grammar.

At this point, I must further clarify the notion of grammatical change as it relates to the topic of grammaticalization. Grammatical conditioning and grammatical change were discussed previously in 1.2 and 1.3. Grammatical conditioning is “the configuration of factors affecting the occurrence of the variant forms” (Poplack, 2000:14), while grammatical change was defined as a process in which individual linguistic features and forms take on new functions over time (Tagliamonte, 2002a:747–8), as indicated by the changing significance, magnitude of effect, and constraint hierarchies of internal linguistic factors in a multivariate analysis (2002a:731).

Grammaticalization is a particular type of grammatical change that involves the slow and gradual process of new grammatical, functional elements of language evolving out of what were once lexical, or content elements (Tagliamonte, 2012:87–88). This is “a collection of interrelated language changes within a construction, that includes phonetic reduction, decategorisation, semantic change and pragmatic shift” (Cheshire, 2007:166, citing Bybee, 2003).

However, grammaticalization is not the only type of language change that can involve the operation of, and changes to, grammatical conditioning of a form. Lexical replacement can also involve changes to grammatical conditioning, such as the weakening of previously significant effects of grammatical constraints on obsolescing forms to the point that a grammatical constraint is no longer significant, until the weakening form disappears completely.  

---

1 These four “mechanisms” have alternately been referred to as erosion, decategorialization, desemantization and extension. (Tagliamonte, 2012:88). See Tagliamonte (2012:87–93) for examples of these processes in the variation and change literature.

The term *grammatical change*, then, refers here to a particular type of linguistic change (i.e. grammaticalization), and has been differentiated from *register change.* Changes in grammatical conditioning over time, while they can be used as a **diagnostic** of grammaticalization, do not automatically imply it. There are several other diagnostics of grammaticalization, such as *layering* (new functions for forms existing side-by-side with older functions) *specialization* (of a form to more general meanings as it continues to grammaticalize) and *persistence* (of grammatical constraints from older functions of a form continuing in the current form) (Hopper & Traugott, 1993:22–31).

Returning now to the diagnostic variables to be used in the present study; if changes are in progress for either variable (genitives or relative pronouns), we can hypothesize, based on the common assumption stated earlier that “changes enter the written language in approximately the same order as they appear in speech, after some undetermined time lag” (Pintzuk, 2003:525), that a change will take place first in vernacular speech and then spread to other registers. While usually accurate, it was pointed out that this assumption has not always held true. In fact, it bears repeating that the introduction of the *wh*-relative pronouns in Middle English is one such instance where the opposite is believed to have happened (Romaine, 1982; Tottie, 1997a:84). Cross-register comparison is therefore all the more essential in establishing the timeline and nature of change for this variable.

It is also possible that a change in frequency of certain forms (such as an increase of relative pronoun *that* or s-genitive) in the written registers is not due to any linguistic or grammatical

---

3 *Register change* was defined in 1.2 to be “when co-occurring groups of linguistic features that are used in a particular situational context alters, either by the increase or decrease in frequency of one or more of those features” (Biber & Finegan, 1989, 1992).
change such as those scenarios described above, but is because the pattern of co-occurrence of these features with other linguistic features, such as those described for the “informational” end of Biber’s register Dimension 1 described in Chapter 1, is changing within a register. If the co-occurrence pattern of features in a register is changing such that the register would fall at a different point on a register dimension (perhaps closer to vernacular speech than it was, i.e. the register is becoming more “informal”), then this is register change. Finally, it could be a combination of both register change and grammatical change. If so, we must determine the degree to which each of those processes is affecting the variable in question in order to gain a full perspective on the change. The evidence needed to answer these questions, and discussion of the indicators of grammatical and register change, including hypothetical data examples illustrating the three scenarios, will be discussed in sections 2.3–2.5. The relative pronoun study will be the focus of chapter 3, and the genitive study the focus of chapter 4.

2.2 The registers

Recall that the assumption of the “chronology of change” from speech into writing is that linguistic change will appear first in speech, then proceed into the written language after a lag of some (unknown) time (Pintzuk, 2003:525). In order to test this, two registers should be chosen that will capture the relative chronology of vernacular changes coming into the written language. Using Biber’s register dimensions, we can pinpoint appropriate registers along a continuum of written to spoken language. If a register can be found that straddles the boundary of speech and writing, such as a written-to-be-spoken register (Hundt, 2008:169), this would be a likely candidate for the point at which changes enter writing from speech. We now have a starting point for choosing the registers that will serve as appropriate diagnostics for examining cross-register variation and change.

Returning to Biber’s (1988, 1995) Dimension 1, Involved versus Information Production, recall from section 1.2 that the co-occurring linguistic features associated with the “informational” end of this dimension include those that allow for more careful production due to the possibility of

---

4 The use of two specific registers as appropriate diagnostics will be motivated in section 2.2.
editing time beforehand (Biber, 1995:117). The cline representing Biber’s Dimension 1 and its associated list of co-occurring features were shown in Table 1-1 and 1-2.

There is a high degree of association between the registers on the “informational” end of this dimension and the co-occurrence of features at that end, including, notably, more nouns and a higher type-token ratio. The registers closer to the “involved” end of the cline have the opposite feature-association characteristics, including a lower type-token ratio and fewer nouns. This would indicate that both of the variables discussed above in section 2.1, the s-/of-genitives and the restrictive relative clause pronouns, can be considered relevant within this dimension, since both relative clauses and s-/of-genitives are nominal constructions — genitives require a noun as their head, and relative clauses require a nominal referent for the relativizer. Also of note is the linguistic feature of “sentence relatives” (i.e. non-restrictive relative clauses), which are associated with the “involved” end of this dimension. This is relevant because the variable context for the relative pronouns is restrictive relative clauses, which are thus associated with the “informational” end of the dimension.

We can therefore use the distribution of registers along this register dimension as a guide for choosing the appropriate data sets to investigate variation and change across registers, using these variables. The best results for highlighting cross-register variation would be achieved by using registers from the top, middle and bottom of the dimension, such as face-to-face conversations, spontaneous and/or prepared speeches, and press reportage.

The face-to-face conversation data will come from two corpora housed in the sociolinguistics lab at the University of Toronto: the Toronto Corpus and Directions of Change Corpus. These two corpora, collected from 2002–2008, consist of 3.3 million words combined of vernacular speech in the form of sociolinguistic interviews from 282 speakers of Ontario English spanning the ages of 9–92, from highly urban Toronto, to several smaller cities and towns across Ontario (Tagliamonte, 2003; Tagliamonte & D’Arcy 2004; Tagliamonte & Jankowski, forthcoming). The corpora have been widely used for sociolinguistic analysis of contemporary Canadian speech in apparent time (see, e.g., D’Arcy, 2005; Tagliamonte & D’Arcy, 2004, 2007, 2009; Tagliamonte & Denis, 2010). Since the intent is to track language change as it proceeds, hypothetically, from spoken vernacular into written registers, the text registers must cover the same period of time and
tap into, if not the exact same speech community, then certainly the same general variety of English.

What is needed is a large, accessible source of both spontaneous and prepared speeches, that can be used to represent the registers at the mid-point of the dimension. For this, a corpus has been constructed of data from the Canadian Federal Government *Hansard* records. These are the transcripts of the debates on the floor of the Canadian House of Commons. *Hansard* is a real-time data set, and as it spans the entire 20th century, it contains sufficient time-depth to compare to the apparent-time Toronto Corpus. For reasons that will be explained in detail below (section 2.2.1), the *Hansard* corpus should be considered a register that straddles the boundary between speech and writing—a strangely but fortuitously “in-between” register. It is comprised partly of carefully monitored but unprepared speech as well as edited speeches that have either been reduced from their spoken form to writing, or were written to be read aloud, not unlike sermons, classroom lectures or broadcast new reports where the speaker is following their printed notes.

This would seem to be an ideal testing ground for the hypotheses about the chronology of linguistic change between speech and writing. Since *Hansard* has this seemingly in-between written/spoken status, we might hypothesize that it would provide a point of entry for changes entering the written language from speech, with incoming forms appearing in *Hansard* before they are found in purely written registers. However, *Hansard* is also an extremely formal register, meaning that it has the potential to be considerably more resistant to incoming innovations than the purely written press register. The implications and possible reasons behind this will be discussed as part of the empirical chapters (3 and 4), and chapter 5.

Finally, press reportage is a register that falls at the opposite end of Biber’s *Involved–Informational* dimension from face-to-face conversation, making it an ideal choice for the third comparison register. Press reportage has been described as an “agile” register, as it appears to exhibit more innovative linguistic forms than other literary written registers (Hundt & Mair 1999:235). Biber attributes this to the factor of *economy*, or the need for as much information as possible to be packed into a finite amount of available text space. This factor is described as increasing over time, combined with a continuing need for the press register to appeal to a wider audience (2003:180).
Press reportage has also been specifically implicated in ongoing change in the s- and of- genitive formation in English by Hinrichs & Szmrecsanayi (2007:441), who write that:

Read together, Biber (2003) and Hundt & Mair (1999) demonstrate that the linguistic responses to the demands of both popularization and economy are defining developments which need to be considered in a study of English newspaper language. Also, newspaper prose seems to be the most promising genre to analyze in any study of language change in progress, given its openness to innovation. However, as Hundt & Mair (1999:236) point out, one should be careful not to generalize the findings from a newspaper corpus to other genres, considering this special place of press language in the spectrum of genres.

We can therefore hypothesize that press reportage would facilitate changes that are entering written registers (see also Westin, 2002), such that forms entering the written registers from speech will appear in press reportage before other written registers. The press reportage register will be represented by Maclean’s magazine, again drawing a sample spanning the entire 20th century (from approximately the years 1906, 1956, 2006). The overall hypothesis for the “chronology of change” across the chosen registers is that new forms will first appear in vernacular speech (face-to-face conversations) and then in the written-to-be-spoken Hansard data, and then in the press reportage of Maclean’s.

2.2.1 Sample design and construction of the Maclean’s and Hansard corpora

Two corpora of twentieth century Canadian English were compiled from the two text registers discussed above in section 2.1. The press register is represented by Maclean’s magazine, a national Canadian news magazine based in Toronto that has been published at least monthly on a continuous basis since late 1905, and is currently published weekly. As shown in Table 2-1, articles were sampled from three time periods: 1906–1912, 1956–59 and 2006–7, with approximately 100,000 words in total of journalistic prose. Authors were carefully chosen and were included only if it could be reasonably established that they were born and raised in the

---

5 The uneven year spans for the three time periods in Maclean’s were due to the need to identify authors who fit the corpus inclusion criteria. This was more difficult for Maclean’s authors than Hansard speakers. Biographical information for current and past Canadian Federal MPs is publically available. In Maclean’s — and indeed in the news genre in general — it is not uncommon for authors to be identified only by their first initial and last name, and sometimes only by pseudonyms. Further, the citizenship requirements to become a Canadian elected official mean that of the two candidate pools, a larger number of possible candidates can be found per year in Hansard than in Maclean’s.
province of Ontario, and were between 30–65 years of age in the year(s) from which their data was sampled.6

<table>
<thead>
<tr>
<th>Year</th>
<th># authors*</th>
<th># speakers*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1906–12</td>
<td>11 (34,000 words)</td>
<td>1906 8 (67,000 words)</td>
</tr>
<tr>
<td>1956–59</td>
<td>7 (34,000 words)</td>
<td>1956 7 (67,000 words)</td>
</tr>
<tr>
<td>2006–07</td>
<td>7 (34,000 words)</td>
<td>2006 9 (67,000 words)</td>
</tr>
</tbody>
</table>

*Includes men only, all between 30–65 at the time of sample, born and raised in Ontario, see appendix C for more details of the chosen authors/speakers.

The Hansard data set, as also shown in Table 2-1, consists of between seven and nine speakers in each of 1906, 1956 and 2006, with a total of approximately 200,000 words between them. Hansard can include anything from prepared speeches and drafted legislation that is being read into the official record, to transcripts of the mostly unscripted, open-floor parliamentary “question period” where ministers are publicly questioned by other members of parliament (MPs) about matters of policy. Questions and responses can sometimes be quite animated.7 Both prepared and spontaneous speeches are included, but obviously written-to-be-printed materials, such as bill proposals that are being officially read into the parliamentary record, have been excluded. MPs often, but not always, read prepared statements even during the sometimes raucous and open-floor question period.8 It should be noted, though, that Hansard transcripts are edited prior to publication. From the current Hansard Association Style Guide:

---

6 Exceptions were made for two authors in the earliest time period, where the pool of possible candidates was considerably smaller than in the mid and late periods due to lack of available biographical information. Both authors (Augustus Bridle and Harry Anderson) were born in England but moved to Canada before their teens (at 10 and 11 years old, respectively). Both received most, if not all of their schooling in Canada, and lived in Canada for the rest of their lives.

7 However, all interaction must, in theory, still follow parliamentary protocol and the format of a formal debate. Members are referred to by title (e.g. “The Honourable Member from St. George” or “my honourable friend”), and all participants must address comments to the Speaker of the House rather than to one another.

8 There is no clear way in Hansard transcripts to tell when a speaker is reading from a prepared speech or perhaps from notes, as an instructor might do in a classroom lecture, or if they simply have strong rhetorical skills and are speaking freely. It is clear, however, that the reading of speeches from prepared text was strongly discouraged until at least the later 20th century. In 1906 and 1956 Hansard, individuals with exceptional oratorical skills are often openly praised by other members on the record, while those who appear to be reading too much from their prepared notes are occasionally chastised and called out on
Hansard, or the official report of debates, according to Beauchesne’s sixth edition, citation 55: “is not a verbatim transcript of the debates of the House. It is a transcript in extenso. In the case of repetition or for a number of other reasons, such as more specific identification, it is acceptable to make changes so that anyone reading Hansard will get the meaning of what was said. Those who edit Hansard have an obligation to make a sentence more readable since there is a difference between the spoken and the written word.” (Hansard, n.d.:48)

The Hansard style guide includes prescriptive grammar rules for a number of potential sociolinguistic variables, including subject-verb agreement, that-comp and proper use of who, whom, which and that as relative pronouns. The guideline for which states that it “introduces a nonrestrictive clause” and “is always set off by commas,” while that “is the relative pronoun used to introduce a clause that restricts or defines its antecedent” and “should never be set off by commas” (Hansard, n.d.:32). However, as will be shown (particularly in chapter 3), these prescriptions were and are not consistently enforced. Keeping all of the above in mind, I take the point of view that the Hansard record represents a written-to-be-spoken register (Hundt, 2008:169). This means that it falls between a verbatim spoken-word transcript, like those in the sociolinguistic-interview Toronto corpora, and something intended from the outset to be read as written, like Maclean’s.  

the record by the Speaker or other members. The only times when it is obvious that material is being read directly are when drafts of legislation are being read into the official record. Written-to-be-printed is a wholly different register from written-to-be-spoken material (Biber, 1995:177; see also Chapter 1, Table 1a, “prepared speeches” vs. “spontaneous speeches”), and so such portions were excluded from the data collection.

9 It is unclear from the available documentation exactly how long the current version of the Hansard style guide has been in place and how often it is updated. However, it does contain examples that refer to reasonably recent events, so my assumption is that it reflects current editorial policy.

10 In addition, prior to the advent of more widely available audio recording technology in the 1960s, Hansard reporters took their initial notes in Pitman shorthand, a phonetic writing system. Early reporters took shorthand by hand (in fountain pen or pencil). Shorthand transcription machines were eventually introduced as the technology became available. These shorthand notes were then typed out, sometimes by a third party. (Hansard, 2005:25–8)

11 This notion has been independently confirmed by conversation with a former Ontario Legislature Hansard transcriptionist, who explained that transcripts are edited for readability to the extent that entire clauses might be moved around (R.G., p.c., October 24, 2009). This has been further confirmed by comparison of recent transcripts (available next day at http://www.parl.gc.ca/Default.aspx?Language=E) with recorded webcasts of the debates (available same day at http://parlvu.parl.gc.ca/parlvu/).
2.2.2 Limitations of the corpora

Another major limitation of the corpora is in the gender dimension. While data has been collected from women, the data used for these registers in the empirical chapters 3 and 4 comes only from men. This is due to the very limited presence of female journalists in the earliest time period, and the complete absence of female politicians in the House of Commons in the early 20th century. Further, while there was no shortage of female journalists writing for *Maclean’s* in 1956, the topics they report on are highly circumscribed—so much so as to constitute a sub-genre unto themselves within the press register. Unlike the longer “hard news” articles even somewhat lighter human interest stories written by their male co-workers, female journalists wrote about distinctly superficial issues. Due to these limitations and possible confounding factors, only data from men has been included in the analyses in chapter 3 and 4.

2.2.3 Advantages of the corpora

The corpora are comparable in size to the 500,000-word London-Lund Corpus, and are in fact larger than their corresponding genre sections (such as press reportage and formal speeches) of the ARCHER corpus, or the BROWN family of corpora (BROWN, LOB, FROWN and

---

12 Agnes Macphail, 1890–1954, was the first female Member of Parliament in Canada, elected in 1921. In the sample year of 1956, there were four female MPs sitting in the House. Of them, only three were born in Ontario, and data is sparse. Sybil Bennett died in November of that year while in office.

13 June Callwood, well-known social justice activist, reporter and broadcaster, wrote articles for *Maclean’s* in 1956 about theatre. Kate Aiken, a well-traveled and respected reporter/broadcaster who interviewed Mussolini and reported from behind the front line during the Spanish Civil War, contributed practical travel advice on how a woman should pack light, but always have a spare set of pantyhose, how to avoid indigestion during overseas flights (do not eat the food) and an anecdote about changing from trousers—necessary to navigate a motorbike across London—into a ball gown in the coal cellar of Westminster Abbey in order to cover Queen Elizabeth’s coronation (“Kate Aiken lists ten ways to enjoy your travels,” *Maclean’s Magazine*, 31 March 1956). Agnes Macphail, in between stints as a federal MP and an Ontario MPP from 1940–43, wrote agricultural columns for the *Globe and Mail* (see mini-biography at: http://collectionscanada.ca/women/002026-826-e.html).

14 http://khnt.hit.uib.no/icame/manuals/LONDLUND/ll.htm

15 http://www.llc.manchester.ac.uk/research/projects/archer/archer3_1/
FLOB). The Maclean’s and Hansard corpora represent a larger time span than the BROWN family corpora, which are “single year” corpora.  

The Maclean’s and Hansard corpora are sociolinguistically cohesive, taking into account each author or speaker’s place and year of birth, their gender, and containing only those people born in the province of Ontario and between the ages of 30–65 in the sample year. This makes for a real-time corpus that is as sociolinguistically accountable as possible. Most, if not all, speakers can be considered educated and middle-class, since all are literate and employed, though it is possible there are some upwardly-mobile working class and upper class speakers.

Like the LLC, ARCHER and BROWN corpora, the Maclean’s and Hansard corpora have been digitized as searchable text files. The corpora have also been tagged for part-of-speech (PoS) in order to count lexical categories. This process was greatly aided by using automated part-of-speech tagging software. The pros and cons of this are discussed in Appendix B, along with a sample of data output from the “Tree-tagger” application. The results are put to the test and presented in the studies in chapter 3 (relatives) and chapter 4 (genitives).

---

16 The BROWN family corpora are 1 million words each IN TOTAL, across all genres: http://khnt.hit.uib.no/icame/manuals/. The press reportage section of each is 44 sections containing 2000 words each for 88000 words. BROWN and LOB were sampled in 1961, FROWN and FLOB in 1991 (Francis & Kucera, 1979; Johansson et al., 1978; Mair 1997; Hundt et al., 1999, see further description above in section 1.2.2).

17 Methodologically, this approach to corpus construction owes a great deal to the model of the Corpus of Early English Correspondence, or CEEC (Nevalainen & Raumolin-Brunberg, 1996, 2003). That real-time, historical corpus consists of 2.7 million words from over 6000 letters written by 778 individuals from the time period c. 1410–1681. Careful research into each contributing text has resulted in a corpus that is stratified by gender, level of literacy and socio-economic status (2003:43–6). The focus of the Maclean’s and Hansard corpora is on representing written language at particular levels of standardization, so the corpora were never intended to be “balanced” in terms of representing a wide range of socio-economic strata. Lack of gender stratification in the corpora was discussed above in 2.2.2.

18 This was necessary in order to test for the effect of type-token ratio. This factor group will be discussed in section 2.4.2, and again the chapters 3 and 4 as it relates to the analysis of the two linguistic features.
2.3 Evidence of grammatical change

In testing for grammatical change, I follow the quantitative, variationist methodology of comparative sociolinguistics. Each instance of a variable is coded for a variety of internal and external factor groups that could potentially condition the variation. Each of these factor groups represents a hypothesis to be tested. For purposes of demonstrating the methodology, I will restrict discussion to the variable $s$-/of-genitives for the remainder of this section, in advance of the full study in chapter 3.

A number of internal factor groups have been suggested to condition the variant choice in $s$-/of-genitives. These include animacy of the possessor NP, length of the possessor and possessum phrases, and the phonological factor of “final sibilant in the possessor”. The $s$-genitive has been claimed to be favoured by possessors that are human, and following that, animate in some way, such as animals and collective nouns, and disfavoured by inanimates (i.e. the dog’s leg vs. the leg of the table) (Ljung, 1997:25). The likelihood of a possessor NP using an $s$-genitive may therefore fall along an animacy hierarchy such as the following: animate human > animate non-human > inanimate.

The length of the possessor and possessum NPs have been claimed to affect variant choice due to the principle of “end weight,” which states that longer and more complicated constituents will come after shorter ones (Rosenbach, 2005:616–17). This means that the hypothesis for genitives is that shorter possessors are more likely to take the $s$-genitive, while shorter possessums are more likely to take the of-genitive, since the possessor appears before the possessum with the $s$-genitive, while with the of-genitive, the possessum appears first, as in the cat’s long orange tail vs. the tail of the large orange cat (Hinrichs & Szmrecsanyi, 2007:453-6; Szmrecsanyi & Hinrichs, 2008:299–300). Finally, the presence of the sibilant would be expected to discourage

19 The reader is referred to the full empirical study of genitive variation in chapter 4 for more in-depth discussion of these and other factor groups relevant to the variable genitive. The factors briefly discussed here are intended only to introduce genitive variation for the purpose of explicating the methodology to be used in the empirical studies (Ch. 3 and Ch. 4), with concrete examples (but hypothetical data).

20 For genitives, we will see that inanimates should also be broken down into further categories, including places, objects, activities, body parts and units of time, though this is not relevant to the current discussion.
an immediately adjacent sibilant, and so discourage use of the s-genitive (Hinrichs & Szmrecsanyi, 2007:452-3; Rohdenburg, 2003).21

By categorizing all instances of the variable according to each of the factors (hypotheses), and then testing these claims in a multivariate analysis, we can determine if the hypothesis for each factor group is supported. In a multivariate analysis, a factor weight, or probability between 0 and 1, is assigned to each factor by a statistical regression using GoldVarb X (Sankoff et al., 2005; Rand & Sankoff, 1991). The listing of factor weights from most to least favouring environments is the constraint hierarchy. For the factor of animacy described above, we predict that the corresponding factor weights assigned for each animacy category are ranked to support this hierarchy and a factor group as a whole is found to be statistically significant at the $p < .05$ level, this means that there is less than 5 percent probability that the results are due to chance. In that case, the factor group can be considered to be playing a role in conditioning the variation.

By comparing the constraint hierarchy of factors using comparable data from different points in time, we determine if this constraint hierarchy has remained consistent over time, or if it is changing. A shift in the constraint hierarchy of an internal, grammatical factor over time is an indication of grammatical change (Labov, 1982:75; Poplack & Tagliamonte, 2001:225–34; D. Sankoff, 1988a:146–147).

Returning to the example of the animacy constraint with the s-genitive, if the s-genitive is found to be favoured across time with non-human animate possessors, and this constraint hierarchy shifts (contrary to previous and expected findings) so that inanimate possessors start to be used with the s-genitive,22 then we can say that the grammatical function of the form is shifting and a grammatical change is taking place. The significant factor groups of those hypothesized to condition the variant choice23 can further be ranked in descending order according to which ones

---

21 Similar factor groups have been indicated as being relevant for the variable relative pronouns in restrictive relative clauses, including the animacy or humanness of the antecedent, as well as the length or complexity of the relative clause. These two internal factor groups, along with the grammatical function of the pronoun in the relative clause and others, have been suggested to condition the variant choice. This variable, as well as hypotheses for each factor group, will be further discussed in chapter 3.

22 This would be shown by factor weights closer to 1.

23 e.g. length of the possessor and possessum NP and presence of a final sibilant, in addition to animacy.
exert the strongest to least influence on the choice of variant. This measurement is the *magnitude of the effect*.

The differences between constraint hierarchies of individual factors, overall constraint ranking of factor groups by magnitude of effect, and the level of significance of the internal linguistic factors (at the cutoff of $p < .05$ for statistical significance) provide the key evidence. These “three lines of evidence” (Poplack & Tagliamonte, 2001:92–5) are provided by the analysis embodied in the statistical model, in this case GoldVarb. Changes with respect to these three types of evidence (shifting significance levels, constraint hierarchies and magnitudes of effect of significant factors) would suggest that a variant is undergoing a change in its grammatical function; in short, grammatical change (Poplack & Tagliamonte, 2001:225–34).

The method detailed here of testing for grammatical change will also provide further evidence toward the “chronology of change” from speech into writing (Pintzuk, 2003:525). If there are grammatical changes taking place, we are interested not only in whether they take place first in speech or in written data, but also whether they proceed through the grammar of the language in the same way in speech and writing, with the same rankings of internal linguistic constraints over time. The evidence provided by parallel multivariate analyses at different points in time allows us to view what the variable grammar looks like at any single point in time, and then compare those point in order to determine the path by which a change progresses through the grammar.

If a variable is changing so that a variant is coming to be used in different (or more, or fewer) contexts, such as the *-genitive coming to be used with inanimate NPs, then this will appear in the constraint hierarchy as a shift in a factor weight for that context. We can then determine whether different aspects of a single grammatical change (i.e. increased use of *-genitive in inanimate contexts vs. animate contexts) progress through different registers at the same time and speed, in the same manner. If the time-depth of the data is such that we manage to capture the starting point of a change, we can determine the register in which changes to different contexts (such as animate vs. inanimate contexts) of a single grammatical feature (i.e. the genitive construction as a whole) originate.
2.4 Evidence of register change

What is needed is no more and no less than a model of how changing stylistic conventions and changing discourse traditions ultimately lead to changes in the underlying system of grammatical choices. (Mair, 2002:186)

I propose two types of evidence for testing for register change in conjunction with the factors that test for grammatical change.

2.4.1 Part 1: Cross-register comparison of linguistic constraints

The first, and most crucial piece of evidence is the comparison of a single linguistic variable across several registers that occupy different points along a dimension for which that variable is a linguistic diagnostic, i.e. the feature co-occurs with other features in a way that is indicative of one of Biber’s (1988, 1995) register dimensions. If a variable that is part of a set of linguistic features diagnostic of registers at one end of a dimension (i.e. press reportage, found at the informational end of Biber’s Dimension, see Table 1-1) comes over time to be used in registers that fall along other points along the dimension (i.e. spontaneous speeches or face-to-face conversation), then this can be taken as evidence of register change.

This type of comparison has its roots in the comparative method of variationist sociolinguistics, which can be used to trace genetic relationships between historically related dialects and track the paths of grammatical change through them (Tagliamonte, 2002a:747–56). Using the statistical tool of multivariate analysis (Poplack & Tagliamonte, 2001:92–5) and the comparative method, we can determine if there is cross-register variation by comparing the significance, constraint rankings and constraint hierarchies of a variable. Shifts in the constraints for the variable at different points in time within a single register could indicate 1) a change in the underlying grammatical system of the language (grammatical change), or 2) a change in how the variable is being used within that single register, such as when a written register begins to use variants that were previously only found in vernacular speech (register change).

For example, the variable s- vs. of- genitive was previously noted to be relevant to the informational end of Biber’s Dimension 1 (see Tables 1-1 and 1-2), as it is (in either form) a nominal construction, and frequency of the genitive construction overall can be directly correlated to the frequency of nouns on a register (Biber, Johansson, Leech, Conrad & Finegan,
The register of press reportage appears at this end of the dimension, since it included a high collocation of the linguistic features at that end. This variable was also previously noted to be dependent on speech style or register (Ljung, 1997:21). The periphrastic of-genitive considered to be the more formal, used to a higher degree in registers such as academic prose (Biber et al., 1999:302), but with the synthetic, non-periphrastic s-genitive on the increase in journalistic prose and other standard English text registers (Hinrichs & Szmrecsanyi, 2007; Mair, 2006a).

In a synchronic analysis at a single point in time, we would therefore expect to find an adequate frequency of this variable for analysis in such a register. If we find one of the variants, such as the of-genitive, is decreasing in frequency of use over time in press reportage, then this could be either grammatical change or register change. It is possible that another linguistic feature is undergoing grammatical change, and taking on the work of the feature that is decreasing in frequency. However, if fewer inanimate possessors are present in the data (something that can be directly affected by the topic of discourse, or by changes to the communicative function of a register), then there would be fewer contexts present overall in which the of–genitive would more likely be used, but this would be register change and not grammatical change.

By comparing a single diagnostic variable at different points in time in a single register I will be able to differentiate between grammatical and register change. The significance, constraint rankings and constraint hierarchies of internal factors will be taken to indicate two possible types of change when found in the following patterns:24

1) If there are different significance levels, constraint rankings and constraint hierarchies of internal factors across time within a single register for an incoming or outgoing variant, then this indicates grammatical change.

2) If there are changing proportions of occurrence for a(n incoming or outgoing) variant across time in registers that occupy different points on a register dimension, such that the proportion of a variant indicative of one register comes to be indicative of a register at the other end of the spectrum, but there is no evidence of differing levels of significance, constraint rankings and constraint hierarchies of internal factors for the variant across time in that same register, this indicates register change without grammatical change.

This differs considerably from the sociolinguistic tradition (Labov, 1966/1982) of including speech or text “style” as an independent factor in a single analysis. The underlying difference of assumption to the present approach is that the style, or register, is not independent of the grammatical features under investigation.

---

24 This differs considerably from the sociolinguistic tradition (Labov, 1966/1982) of including speech or text “style” as an independent factor in a single analysis. The underlying difference of assumption to the present approach is that the style, or register, is not independent of the grammatical features under investigation.
3) If there are changing proportions of occurrence for a (incoming or outgoing) variant across time within registers that occupy different points on a register dimension, as well as evidence of differing levels of significance, constraint rankings and constraint hierarchies of internal factors for that variant across time within one (or more) register(s), then this indicates both grammatical and register change.

Table 2-2 and Table 2-3, using fictional numbers for a hypothetical variable across two time periods and in two registers (see section 2.2), illustrate these three scenarios. Looking only at Table 2-2, “speech”, there is evidence of changing constraint hierarchies for factor group 1 (factor B disfavours variant X in 1956, but comes to favour it in 2006), and changing significance (FG 2 is not initially significant, then becomes so), although the magnitude of effects stays stable (the factor group is always the strongest effect in its time period, see the range of 45 for FG 1 in 2006 vs. 25 for FG 2). The input probabilities indicate that the frequency of variant X has increased overall in this register, (.53 in 1956 to .77 in 2006), even though the proportions for variant X in the two times are very similar. Assuming FG1 and FG2 are internal linguistic factors, then this is an indication of grammatical change.

Alternatively, looking only at Table 2-3, “writing”, the constraint hierarchies, magnitude of effects and significance of the internal linguistic factors show no difference between the two time periods, but the input probabilities show that the frequency of variant X has increased overall in this register, and the proportions show how variant X has increased uniformly across all contexts.
This would be an indication of register change — no grammatical change is happening, so we must turn to an external factor.

For instance, if the variable in question were the use of that as a relative clause pronoun and FG1 were the internal linguistic factor group animacy of the (nominal) antecedent, we could infer that in speech (Table 2-2), although the variant has increased overall (see the higher input), and although it is being used at the same proportions in each grammatical context, the way it is being conditioned by those grammatical contexts has shifted due to a grammatical change. In Table 2-3, the variant is increasing in proportion, becoming more like the “speech” in Table 2-2. However, the conditioning effect of the grammatical contexts has not shifted, since the constraint hierarchies are consistent for the conditioning factor groups. It is only by comparing the results for the two registers side by side that the evidence for both register change (in Table 2-3), and grammatical change (in Table 2-2) can be seen. By looking at only one register or the other, we come to different interpretations.

2.4.2 Part 2: register-internal constraints

A second piece of evidence for register change comes from the domain of corpus linguistics, and involves testing factor groups that tap specifically into the information structure of a register. Such factors allow us to test for whether there is a shift in how the register makes use of the variable (e.g. Table 2-3, above), while grammatical constraints on it remain constant.

One such factor is lexical density, sometimes referred to as type-token ratio (TTR). A high TTR means that there are a high number of different lexical categories present in a text, something that is indicative of the informational end of the involved/informational register dimension (Biber, 1995:117). If the lexical density of texts in a register are increasing or decreasing over time, this is evidence of register change. Hinrichs & Szmrecsanyi (2007:457–8) and Szmrecsanyi & Hinrichs (2008:300–1) operationalize such a factor group within a study of the s-/of-genitive by calculating “the number of different word types” that are present within 50 words on either

---

25 Whether or not this is a likely scenario is a valid question. Realistically, we expect an increasing variant to appear first in different registers in the same linguistic contexts before any grammatical change takes place.
side of an instance of an interchangeable genitive token. First, each token of a variable is annotated in the data file, then the data is tagged for part of speech using the “Tree-tagger” part-of-speech tagging software (Schmid, 1994a, 1994b, 1995). Since the multivariate analysis is to be run in GoldVarb, which uses discrete, rather than continuous variables, the initial Tree-Tagger output of 36 different parts of speech was further reduced into coarser distinctions of a maximum of 16 unique lexical categories. This meant, for instance, collapsing number distinctions for nouns, tense distinctions for the verbs, etc. (see Appendix B for details). Some of the pros and cons of this method are presented in Appendix B, along with a sample output from the software.

Perl scripts are then employed to count the number of unique lexical categories within 50 words on either side of each token of a variable, and each token is assigned an integer value between 1–16. The factor group is then run in GoldVarb in the same way as internal linguistic factors described previously. If lexical density is found to exert a statistically significant effect on the variable, and that effect is found to be changing over time using the same “three lines of evidence” of the comparative method described above (Poplack & Tagliamonte, 2001:92–5; Tagliamonte, 2002a:750–53), this is evidence of register change.

---

26 These studies investigate the $s$-/of–genitive variation. While this factor will certainly be relevant to the empirical study of genitives presented in Chapter 4, it is also relevant to the relative clause study, and so the variable was included for this study as well.

27 My method differs from that of Hinrichs & Szmrecsanyi (2007) and Szmrecsanyi & Hinrichs (2008). I took their description of counting “the more different word types we find in a given passage” (Hinrichs & Szmrecsanyi 2007: 457, bolding mine) to refer to the number of unique lexical categories in a passage, when in fact they counted unique lexical items. Although the two counting methods differ, a high number of unique lexical categories is also indicative of high lexical density. The effect of counting lexical categories, as was the case in the present study, is that the available range measured was narrower than the possible range for number of lexical items. This would underestimate any effect of the factor group, making my method more conservative in its assessment of lexical density. Testing the impact of these two different methods of calculating lexical density is for future research.

28 Hinrichs & Szmrecsanyi (2007) and Szmrecsanyi & Hinrichs (2008) use continuous variables in their model of genitive variation. However, since the possible range here is whole integers from 8–15, this factor can just as easily be treated as discrete. While transforming continuous variables into discrete is not a favoured statistical practice (Johnson, 2009:362), the main complaint seems to be loss of predictive power, specifically that “...it results in underestimating effect sizes and reducing the power of statistical hypothesis tests” (Cohen, 1983:249). If the best model fit still returns significant and interpretable results in spite of that (e.g. the different frequency values are hierarchically ranked in order by their factor weight), then the analysis should be considered sound.
In addition to being a rubric of possible changes to the information structure of a register, lexical density (TTR) can also be used to tap into the notion of “economy” described earlier in this chapter. This is the idea that it is preferable in some registers to encode as much information as possible in less space, and is thought to be particularly active in journalistic prose in the later 20th century (Biber, 2003). A high lexical density means that there are a high number of different lexical items in a text, something that is indicative of the informational end of the involved/informational register dimension (Biber, 1995:117). In variables where one variant can be said to be more “economical” than another, as in the case of the s-genitive (over the of-genitive) or the Ø-relative pronoun (over any other variant where the pronoun is realized), the need for economy dictates that the more compact variant be used. A change over time for such a factor group indicates that the pressures of economy are changing, which would then force the register to change.

Hinrichs & Szmrecsanyi (2007:457–8) and Szmrecsanyi & Hinrichs (2008:300–1) operationalize such a factor group within a study of the s-/of-genitive by calculating how many different parts of speech are present within 50 words on either side of an instance of an interchangeable genitive token. In general, the higher the type-token ratio, the more likely the need should be to use the more “economical” s-genitive.29

A second factor that can tap into changes to the informational end of Biber’s Dimension 1 (1995:117) is information status. This refers to the givenness or newness of an item in the preceding discourse. This is of particular relevance for variables involving nominal constructions (such as the variable genitives and variable relative clause pronouns), since these often could, in later discourse, be replaced by referential pronouns. Hinrichs & Szmrecsanyi (2007:451) call this “a higher-level pragmatic factor.”30 More nouns are indicative of informational production, while more pronouns indicate a less-informational production (see Ch. 1, Fig. 2). Further, new

29 This is also testable with the variable relative pronouns: the higher the number of different parts of speech within 100 words (50 to either side) of a variable relative clause, the more likely the need should be to use the Ø-variant to economize on space rather than an overt forms such as that or a wh-relativizer. However, the relevance of lexical density is greater than simply as a measure of the need for economy, as will be demonstrated in Ch. 3 for the variation between that and which.

30 Again, their study investigates the genitive s-/of– variable. This factor, coded the same as for the relatives study here, will also be relevant to the empirical study of genitives presented in Chapter 4.
material introduced into the discourse is more likely to appear as a noun than a pronoun (Fox & Thompson, 1990:304).

Therefore, if a register comes over time to use more nominal constructions that are New information as opposed to Given (or vice-versa), this is evidence of register change. The factor group can be coded by implementing Perl scripts to search for any occurrence of a head noun in a given nominal construction within the 100 words prior in the discourse (or text excerpt). If the effect of such a constraint is found to be changing over time — again, visible by a change in the significance, constraint ranking or constraint hierarchy of the factor groups in a multivariate analysis — then this would in and of itself be strong evidence for register change, since it speaks directly to the co-occurrence of linguistic features that make up the register. Further, register changes like these — changes in lexical density, the propensity for new vs. given information status in a text, or changing pressures of economy — could in fact account for the actuation of some types of change; as the linguistic makeup of a register changes, so then certain variants become more pre-disposed.

As discussed above in section 1.4, Szmrecsanyi, coming from the perspective of a historical corpus linguist, makes the strong argument that “subtle shifts in the stochastic [random, statistically analyzable] effects of conditioning factors” — i.e. internal linguistic constraints such as those described in section 2.4.1 — should be used to test for grammatical change “instead of focusing solely on text frequencies”, since “fluctuating frequencies of grammatical variants are a function not only of changing grammars, but are also conditioned by environmental changes in the textual habitat” (2013:2). This is, in fact, exactly what variationist methodology is, with many decades of research behind it (see, e.g. Cedergren & Sankoff, 1974; Labov, 1982; Poplack & Tagliamonte, 2001; D. Sankoff, 1988a).

I would further argue, then, that what I have been calling “register-internal constraints” here, which tap into text frequencies, should be incorporated into variationist sociolinguistic methodology where appropriate, and not only left to the purview of corpus linguists, as they tend to be. This tendency again comes back to the idea of vernacular speech data as the ideal data for sociolinguistic study. However, once we acknowledge that vernacular speech is not the only register in which language change can originate (see Chapter 1 pp. 1–6), and that there are times when an edited written record is all that is available, then we are left to conclude that we must
also attempt, where possible, to take into account the frequency effects in texts as well as speech.  

2.5 Summary

In this chapter, the use of three language registers (two text and one spoken) was motivated for investigating the relationship between speech and writing in the course of language change. Diagnostic linguistic variables that will be used to further investigate this relationship were introduced: variable $s$-/of-genitive, and variable relative pronouns. The comparative, quantitative variationist methodology that will be used in the empirical studies was described and demonstrated in abstract, in particular how it will be used to identify and separate potential grammatical change from register change.

All of this paves the way for the first of the empirical studies in Chapter 3: variable relative clause pronouns. While relative pronoun variation has long been a widely-studied linguistic phenomenon, the perspective presented here will add not only to our understanding of the linguistic variable itself, but to the larger picture of variation and change across speech and writing. Using this methodology, the hope is that if the starting point of a change can be captured, we will be able to determine whether or not that change is originating in the spoken language, or in some other register. Further, this methodology can provide evidence whether or not we have “change from above” or “change from below” (Labov, 1994:78). Bell’s Style Axiom (1984:151) states that “Variation on the style dimension within the speech of a single speaker derives from and echoes the variation which exists between speakers on the ‘social’ dimension.” This suggests that a variable that operates across the style dimension — such as one

---

31 A question that arises is whether register change involves a change of a register, or a change within a register. Some properties of register change appear to be inherent, or internal to the register (i.e. changes to the linguistic makeup of the co-occurring groups of features). Factors such as type-token ratio tap into this. When such changes take place they are brought about by external, social forces, including (as will be shown in Chapter 3), changing editorial policies, stylistic change, colloquialization, etc. These influences are also referred to as register change, but are more along the lines of what Görlach calls “changes in communicational intentions, the media, educational facilities and other social factors” (2002:18). As described in 1.2, it is difficult to quantify notions such as “changes in communicational intentions” and “changing editorial policies.” Thus, this study uses register-internal factors as a means to test register change generally and appeals to external social forces to motivate such internal register changes.
that has correlations with the register-internal constraints — will be one that operates above the level of conscious (social) awareness due to increased self-monitoring of language in certain registers. The register-internal constraints test whether there is change within the style dimension. If correlations with the register-internal constraints are found, implicating them in the ongoing change, this is evidence that the change is taking place above the level of conscious awareness, or a change from above.
Chapter 3
Relative pronoun variation

The first of the two empirical studies, which will serve to test the claims and demonstrate the methodology described abstractly in Chapter 2, is the relative pronouns. This variable will be used to demonstrate a case of register change without accompanying grammatical change, before addressing a case that shows both grammatical and register change in Chapter 4.

3 Introduction

Relativization in English has been described as “one of the most researched topics in the history of the English language” (Nevalainen & Raumolin-Brunberg, 2002:109). As motivated above in section 2.1 (see example (3)), the variation in English relative pronoun choice has been shown to be tied closely to register and formality (see also Ball, 1996; McDavid, 1977; Montgomery, 1989; Romaine, 1982; Tottie, 1997a, 1997b) since their introduction into English. Further, recall that the introduction of the WH-forms of the relative pronouns is one of the examples of a change that is believed to have entered the language first in formal, written registers from contact with French and Latin grammar to compete with that and Ø before eventually spreading to vernacular speech in Middle English (Romaine, 1982; Tottie 1997a:84). These forms never replaced the native English that-relative, which continues to compete with who and which to the present day in vernacular speech (D’Arcy & Tagliamonte 2010, Tagliamonte et al., 2005). The inflected forms whom and whose are restricted in use to the object and genitive case, respectively, and whom in particular is considered to be restricted to the most formal of written registers, and to have fallen out of regular use in vernacular speech (Quirk, Greenbaum, Leech & Svartvik, 1985:367–8).

Previous studies (see D’Arcy & Tagliamonte, 2010:383–4 for discussion) have revealed long-standing grammatical conditioning on the choice of variants; grammatical constraints that are still in effect in modern English. The conclusion that “the infiltration of WH into the relative system can be seen as completed in the modern written language […] but it has not really affected the spoken language” (Romaine, 1982:212) might indeed hold for vernaculars, but those studies that look at either more educated or more formal speech have shown that the WH-forms are certainly present outside the written language (D’Arcy & Tagliamonte, 2010:383–4). In other
words, this variable is known to be strongly conditioned not only by internal grammatical factors, but by stylistic and register-based factors. This makes it an ideal testing ground for possible co-occurring processes of grammatical change and register change using the methodology described in Chapter 2. D’Arcy & Tagliamonte (2010:386, bolding mine) note:

> Grammatical descriptions of the relative system neatly circumscribe the use of particular forms to specific linguistic contexts [...] who(m) for persons, which for things (see e.g. Curme 1947:166, Denison 1998:278, Jespersen 1961:80, Swan 1995:473), and that (which carries no case or person inflection) for either persons or things (Swan 1995:473). According to Montgomery (1989:136), these prescriptions were in place by the eighteenth century. We take the sum of these facts to imply stable variation among the relative pronouns in present-day English.

Given the historical context and such predictions of stability with this variable, we can hypothesize that we will find a highly constrained grammatical system for the relative pronouns in *Maclean’s* and *Hansard*, but no evidence of grammatical *change* over time. Rather, any changes in distribution over time are expected to be due to external factors, such as a change in the makeup of a register (i.e. its co-occurring features, as per Biber, 1988, 1995).

In addition, this is the first study on relative pronoun variation — to my knowledge — that focuses on the *which* variant in formal spoken and written Canadian English (for discussion of *which*\(^1\) in other varieties of English including British, American and New Zealand, see, e.g., Ball, 1996; Hundt & Leech, 2012; Hundt, Denison & Schneider, 2012; McDavid, 1977), making it an excellent beginning point for cross-variety comparison for major varieties of English across speech and writing.

### 3.1 The variable context

In circumscribing the variable context with the express purpose of this study being as comparable as possible to a study of the same variable in vernacular speech, this study follows the methodology of D’Arcy & Tagliamonte (2010:387–8) (modeled in turn after Tagliamonte, 2002b:151–2; Tagliamonte et al., 2005:84–6; and Tottie & Harvie, 2000). The study looks only at restrictive relative clauses, as in (1), and not non-restrictives, as in (2).

\(^1\) The decision to focus on this variant will be motivated below.
(1) Restrictive use: “it’s the man who prints a newspaper” [Maclean’s 1906/a]
(2) Non-restrictive use: “The market for his slippers was found by his good wife who sold them on the streets of the town” [Maclean’s 1906/d]

Non-restrictive relatives are known to pattern differently from restrictives in relative pronoun use, in that they are almost categorically found with a wh-form, and unlikely to be found with that or Ø. Non-restrictive relatives also provide more supplementary information than restrictive relatives (e.g. Huddleston & Pullum, 2002; Quirk et al., 1985).

Adverbial relatives such as (3) serve a different grammatical function, allowing where or when to be substituted. These were also not extracted, nor were cases of WHIZ-deletion, where the relative pronoun and the verb to be are deleted.²

(3) a. adverbial use (i.e. ‘when’): “By the time Ø we got to the 400-foot-long breach in the 17th Street Canal…” [Maclean’s 2006/z]
   b. adverbial use (substitute where): “There are dramatic groups in Canada without halls in which to perform…” [Maclean’s 1956/s]

Table 3-1: Overall distribution of relative clause pronouns, *Maclean’s* and *Hansard*

<table>
<thead>
<tr>
<th>Relative marker</th>
<th>Maclean’s</th>
<th>Hansard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>that</td>
<td>35</td>
<td>435</td>
</tr>
<tr>
<td>who</td>
<td>27.5</td>
<td>345</td>
</tr>
<tr>
<td>Ø</td>
<td>21</td>
<td>265</td>
</tr>
<tr>
<td>which</td>
<td>13.5</td>
<td>172</td>
</tr>
<tr>
<td>whose</td>
<td>2</td>
<td>24</td>
</tr>
<tr>
<td>whom</td>
<td>1</td>
<td>13</td>
</tr>
</tbody>
</table>

| Total N         | 40   | 1254 | 60   | 1894 |

This resulted in a total of 3148 tokens of restrictive relative clauses, 1254 from *Maclean’s* and 1894 from *Hansard*. The overall distribution of the forms in the two corpora is given in Table

² WHIZ deletion involves both deletion of the relative pronoun and the auxiliary to be, as in: They walked to the store [that is] up the road or There’s a man [who is] planning to visit tomorrow. Tottie & Harvie (2000:206) exclude these constructions on the basis that they are often not classified as relative clauses because they are non-finite, and that the construction also involves auxiliary deletion. Tagliamonte et al. (2005:86) also adopt this methodology, as do D’Arcy & Tagliamonte (2010). For the sake of comparison, these constructions were not extracted for the present study. At the very least, their status as non-finite constructions indicates that they would need to be considered separately from finite relative clauses.
3-1. The older *that* and Ø forms appear stable in this initial cross-register comparison, while the WH-forms already exhibit cross-register variation — a higher proportion of *who* in *Maclean’s*, but a higher proportion of *which* in *Hansard*.

Numerous studies (see those previously cited) have shown that the grammatical function of the antecedent distinguishes the use of certain forms. The Ø variant is rarely found with subject relatives (4) and is prescribed against in that position (Quirk et al., 1985:420), while *who* is almost never found with direct object function (5). The two *wh*-forms that still carry overt grammatical inflection, *whose* (6) and *whom* (7), are reserved primarily for genitive and indirect object function, respectively. *Who* is never found as the object of a preposition (8a). ³ Rather, *which* is often used in this way to indicate an adverbial sense of ‘where’ (8b) or ‘when’ (8c), as is *whom* (8d), with its holdover of overt English dative case marking.

(4) Subject: “it’s *the man* **who** prints a newspaper” [Maclean’s 1906/a]
(5) Object: “I should like to add something to the remarks Ø I have already made” [Hansard 1956/N]
(6) Genitive: “he is defaming a man *whose* shoe laces he is not worthy to unloosen” [Hansard 1956/K]
(7) Indirect Object: “it may go back to *the people to whom* he says it belongs.” [Hansard 1906/F]
(8) a. Object of Preposition (stranded): “I venture to submit that *this particular class of legislation that* we are now dealing with will not be conducive” [Hansard 1906/b]
b. Object of Preposition, adverbial where use: “Canada is the only country in **which** the railways are allowed to go practically untaxed.” [Maclean’s 1906/j]
c. Object of Preposition, adverbial **when** use: “returns for 1907, the latest year for **which** detailed returns are at present available.” [Maclean’s 1906/j]
d. Object of Preposition, whom: “or did the agent of the broker through whom they were bought examine them? [Hansard 1956/N]

Table 3-2 summarizes these patterns. Given these predictions, it becomes clear that the data must be separated by grammatical function of the relative pronoun. Although the lexemes overlap, they group together in unique sets according to the different antecedent functions of the relative clause domain.

³ There is also a distinction in permissible forms between when the preposition is stranded or fronted. Compare the use of *that* in (8a), which has a stranded proposition, to the same sentence with the preposition fronted, where *that* is not permitted: “*this particular class of legislation with which* we are now dealing...”
Table 3-2: Possible relative pronouns by grammatical function
(adapted from Sigley, 1997:209, in turn adapted from Quirk et al., 1985:366f)

The distribution of all restrictive relative clause tokens by grammatical function in the current data is shown in Table 3-3 (Maclean’s) and Table 3-4 (Hansard). From these tables, we see of the grammatically uninflected forms (that, who, Ø, which), no more than three are in active use for any one single grammatical function at a time. No single analysis could therefore address all variants across all grammatical functions. In subject function, only that, who and which are found. For direct object and prepositional object function, the variants are that, Ø and which, but at very different rates — Ø is the preferred variant for direct object function, and which for prepositional object function. Both indirect object and genitive function are sparsely represented in the data. As subject and (direct) object function account for approximately 89 percent (Maclean’s) and 90 percent (Hansard) of the data, analysis from this point will focus on these two functions, starting with subject function.

Table 3-3: Relative clause pronouns in Maclean’s by grammatical function

<table>
<thead>
<tr>
<th>Subject</th>
<th>Direct Object</th>
<th>Prep. Obj.</th>
<th>Ind. Obj.</th>
<th>Genitive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>that</td>
<td>48</td>
<td>379</td>
<td>14</td>
<td>43</td>
</tr>
<tr>
<td>who</td>
<td>43</td>
<td>345</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Ø</td>
<td>--</td>
<td>--</td>
<td>76</td>
<td>235</td>
</tr>
<tr>
<td>which</td>
<td>9</td>
<td>73</td>
<td>10</td>
<td>31</td>
</tr>
<tr>
<td>whose</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>whom</td>
<td>--</td>
<td>--</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total N</td>
<td>797</td>
<td>310</td>
<td>111</td>
<td></td>
</tr>
</tbody>
</table>
Table 3-4: Relative clause pronouns in *Hansard* by grammatical function

<table>
<thead>
<tr>
<th></th>
<th>Subject %</th>
<th>Direct Object %</th>
<th>Prep. Obj. %</th>
<th>Ind. Obj. %</th>
<th>Genitive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>that</td>
<td>49</td>
<td>554</td>
<td>27</td>
<td>153</td>
<td>14</td>
</tr>
<tr>
<td>who</td>
<td>29</td>
<td>331</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Ø</td>
<td>--</td>
<td>--</td>
<td>55</td>
<td>312</td>
<td>21</td>
</tr>
<tr>
<td>which</td>
<td>21</td>
<td>240</td>
<td>18</td>
<td>103</td>
<td>61</td>
</tr>
<tr>
<td>whose</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>whom</td>
<td>--</td>
<td>--</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Total N</td>
<td>1125</td>
<td>570</td>
<td>183</td>
<td>5</td>
<td>11</td>
</tr>
</tbody>
</table>

Figure 3-1 and Figure 3-2 give the overall distribution of *that*, *who* and *which* in subject function relative clauses across the three time periods and two corpora. In the press register of *Maclean’s*, *that* and *who* parallel one another in proportion of use over time until 2006, where they begin to diverge.\(^4\) *Which*, although in competition with *that* and *who* in 1906, had already all but disappeared by the mid-20\(^{th}\) century. *Hansard* holds on to restrictive *which* for longer, lagging

---

\(^4\) The difference between the proportions for *that* and *who* in Figure 3–1 is significantly different. The lower confidence interval limit for *that* is 52 percent, and the upper confidence interval limit for *who* is 48 percent. Since the confidence intervals do not overlap, the proportions (58 percent *that* and 41.5 percent *who*) can be interpreted as being significantly different at the p < 0.05 level.
behind Maclean’s in the loss of this form until the late century, when which drops off abruptly. Given claims of the press register being more “innovative” and “agile” (Hundt & Mair, 1999:235), this is the predicted result.

**Figure 3-2:**

![Relative pronouns in Hansard over time, subject function](image)

What is more interesting for the present study, however, is that there is a visible cross-register difference in the use of that and who by 2006. Even with the loss of which in both registers by 2006, there are still register differences — the gap between the rates of use of these two forms is considerably more pronounced in Hansard, with a noticeably lower proportion of use of who (41.5 percent in 2006 Maclean’s vs. 30% in 2006 Hansard). Again, this is an expected result when we consider the oft-cited finding that, despite centuries of co-existence and competition with the older non-wh forms, the wh-forms of the relative pronouns, which and who, are generally not found in speech (D’Arcy & Tagliamonte, 2010:383–4). Hansard is the more “oral” of the two registers (i.e. ‘written-to-be-spoken’, or edited transcripts of speech), while Maclean’s, being a purely written register, would be expected to more closely follow the prestige standard of the written language.

Figure 3-3 and Figure 3-4 plot the overall distribution of the relative pronouns found in object function relative clauses across the three time periods and the two corpora. Here, the variants in play are that, Ø and which:
Again, *which* declines first in *Maclean’s* and is completely lost by 1956, while *Hansard* retains the form until the latest time period. The upward trajectory for Ø in *Maclean’s* is a mirror image of *which*, indicating that *which* is simply replaced here by Ø. The trajectory of *that* in *Maclean’s* is stable throughout. Overall, there is little evidence for ongoing change as there is with *that* for subject relatives in *Maclean’s*.

In *Hansard*, it is the Ø variant that remains fairly stable across the century, with a brief increase in 1956, in parallel to the beginning of the drop in *which*, much like in *Maclean’s*. *That* does
appear to be in the rise in 2006. However, compare the trajectories of *that* and *Ø* in Figure 3-3 and Figure 3-4 to Figure 3-5, the non-subject function relatives in Toronto (D’Arcy & Tagliamonte, 2010:393).

**Figure 3-5: Distribution of relative pronouns by Non-Subject Function over apparent time in Toronto (from D’Arcy & Tagliamonte, 2010:393)**

The Toronto data from D’Arcy & Tagliamonte (2010) is apparent time, collected from 2002–2004. This means that the 30–59 year olds in the Toronto data are approximately the same age range as the 2006 *Hansard* and *Maclean’s* informants, who were 30–65 at the time of their data sample in those corpora. The 1956 *Hansard* and *Maclean’s* data is then roughly comparable to the 60 and over speakers in Toronto. The oldest speaker in the Toronto Corpus was age 92 in 2002. The Toronto Corpus is the larger of the Ontario Dialect Corpora described above in section 2.2, containing data from 214 speakers. Data from the smaller city of Belleville and the towns of Lakefield and Burnt River (a combined 68 speakers), were not used in D’Arcy & Tagliamonte’s (2010) analysis, and only a subsample of the 214 speakers from the Toronto Corpus. In addition, data from female speakers is used in the vernacular speech study, unlike for the analyses of *Maclean’s* and *Hansard* (see section 2.2.2 for discussion). The reader is referred to D’Arcy & Tagliamonte (2010) for specifics on the sample size used in that study. The inclusion of data from women might be expected to increase the amount of standard variants in the vernacular speech data, given Labov’s widely cited Principle I (2001: 266) that “in stable sociolinguistic stratification, men use a higher frequency of nonstandard forms than women. However, the overall low rates of *wh*-forms in vernacular speech reported by D’Arcy & Tagliamonte (2010) indicate that differences across register for this variable are more salient than those across the gender dimension.
The trajectory for *that* and Ø in non-subject relatives for speakers under 60 in Toronto is completely stable, with no ongoing change indicated. In *Maclean’s*, the pattern of stability started earlier, in 1956, the time period equivalent to the over 60 speakers in apparent time. This provides evidence to project that *Hansard* will likely remain stable into the future as well. It is worthy of note, though, that it is *Maclean’s*, the press register, that stabilized earliest. For the remainder of the chapter, then, we will turn to further analysis of the subject relatives (see Figure 3-1 and Figure 3-2), where there is evidence of ongoing change and increasing polarization between the forms.

### 3.2 Internal factors: Animacy and grammatical number of antecedent

One of the most important factors determining the use of *that*, *who* and *which* in subject relative position is animacy. Ball documents the rise of *who* at the expense of *which* for human reference beginning in the 16th century, as well the rise of *that* to compete with *which* for non-human reference beginning in the 17th century (1996:246–51), citing the observation that “There was formerly a tendency to use ‘that’ for everything: the tendency now is to use ‘who’ and ‘which’ for everything.” (Fowler & Fowler 1973:88) as an indication of the state of this variation in the early 20-century (Ball, 1996:248).

However, by the 1970s to 1980s, a strong prescription had arisen in North America against the use of restrictive *which* in writing. Tottie (1997a:84–7) and McDavid (1977) attempt to date this trend. Both authors point to the suggestion of H. W. Fowler’s *Modern English Usage* (first published in 1926, re-issued in 1965) that:

…perhaps the line of improvement lies in clearer differentiation between *that* and *which*, and restoration of *that* to the place from which, in print, it tends to be ousted. […] [I]f writers would agree to regard *that* as the defining [i.e. restrictive] relative pronoun, and *which* as the non-defining, there would be much to gain both in lucidity and ease. (Fowler, 1965:625f)

Tottie concludes that the North American usage guide writers took this suggestion to heart, and began to include this as a prescriptive rule in their publications, where it began to appear in the style guides of major publishers and newspapers by the mid 1970s and continued through the
Tottie’s conclusion is supported by McDavid (1977), who reports the proportion of restrictive which at 72.7 percent of all which-relatives in a 1955–1965 corpus of American prose, and nearly 74.75 percent in the original BROWN corpus (described in sections 1.2 and 2.2.1), which represents a balanced sample of 15 standard American text registers from 1961. McDavid concludes that at that time, “The advice that which be used only nonrestrictively is thus so far from reflecting actual practice that it should be abandoned” (1977:81).

Each token in the data was coded for animacy of the antecedent based on an expanded version of the coding scheme used by Tagliamonte et al. (2005:90), incorporating the animacy hierarchy of Zaenen et al. (2004:11). Animate categories included humans (personal names and otherwise, singular and plural), non-human animates and collective humans. Inanimate categories included organizations, “municipal” places (such as specific towns, countries, provinces, streets and political entities), non-municipal places (such as generic landscapes and planetary bodies), inanimate objects, activities, states, body parts and units of time. The full coding scheme for inanimates with examples is provided in Appendix D.

Organizations, as defined here, are technically inanimate entities, but are arguably similar to collective humans and plural humans. Zaenen et al. (2004:10–11) discuss the inherent difficulties in coding for this category:

Especially prevalent cases of metonymy are the use of names to refer both to organizations (e.g. IBM) and to characteristic members of them, and the use of place names (e.g. Russia) to refer both to organizational entities and geographical places or inhabitants of them. Terms belonging to these semantic classes are systematically ambiguous. […] organizations are often presented as groups of humans engaging in actions that are typically associated with humans. […] The cut-off point between HUMAN and ORG[anization] was put at ‘having a collective voice/purpose’: so a group with collective voice and purpose is deemed to be an ORG[anization], a group with collective action, such as a mob, is not.

Recall also from Chapter 2 that the Canadian Hansard style guide has such a prescription, stating that which “introduces a nonrestrictive clause” and “is always set off by commas,” while that “is the relative pronoun used to introduce a clause that restricts or defines its antecedent” and “should never be set off by commas.” (Hansard Association of Canada Style Guide:32).

Human >> other animates (includes organizations, animals, intelligent machines and vehicles) >> inanimates (includes concrete inanimate, non-concrete inanimate, places, time.) (Zaenen et al. 2004:11)

Cross-tabulation of abstract inanimate antecedents by grammatical number reveals little difference in the proportion of that vs. which between singular and plural abstract inanimate antecedents in either Maclean’s or Hansard.
Based on this, a “group [of humans] with a collective action” would be coded as plural human. I have made the further distinction between these (coded as collectives) and simply plural humans (i.e. men/adults/girls). As well, the category of places is not always straightforward to code:

There are two different problems with the delimitation of place. On the one hand, any location can be a place, e.g. a table, a drawer, a pinhead, … The coding scheme takes the view that only potential locations for humans are thought of as ‘places’. On the other hand some places can be thought of as ORG[anization]s. (Zaenen et al., 2004:11)

Table 3-5 (Maclean’s) and Table 3-6 (Hansard) show the subject relatives in the data by animacy and number of the antecedent. It is immediately apparent that human antecedents are overwhelmingly realized by who in the data. As Ball (1996:252) observed, “The work that who performs seems to be fairly clear: in standard English it marks personal subjects, at such a rate as to be almost categorical”, and this observation is again borne out in this data. Non-human antecedents demonstrate a preference for that, in competition with which. Despite differences in

### Table 3-5: Animacy of antecedent in Maclean’s, Subject function

<table>
<thead>
<tr>
<th></th>
<th>Human, pl. %</th>
<th>Human, sg. %</th>
<th>Collective %</th>
<th>Organization %</th>
<th>Places %</th>
<th>Concrete Inanimate %</th>
<th>Abstract Inanimate %</th>
</tr>
</thead>
<tbody>
<tr>
<td>that</td>
<td>1</td>
<td>5</td>
<td>53</td>
<td>74</td>
<td>78</td>
<td>78</td>
<td>87</td>
</tr>
<tr>
<td>who</td>
<td>99</td>
<td>95</td>
<td>25</td>
<td>11</td>
<td>6</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>which</td>
<td>--</td>
<td>--</td>
<td>22</td>
<td>15</td>
<td>16</td>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td>Total N</td>
<td>195</td>
<td>144</td>
<td>32</td>
<td>27</td>
<td>32</td>
<td>101</td>
<td>250</td>
</tr>
<tr>
<td>Total %</td>
<td>24.5%</td>
<td>18.5%</td>
<td>4.5%</td>
<td>3.5%</td>
<td>4%</td>
<td>13%</td>
<td>32%</td>
</tr>
</tbody>
</table>

### Table 3-6: Animacy of antecedent in Hansard, Subject function

<table>
<thead>
<tr>
<th></th>
<th>Human, pl. %</th>
<th>Human, sg. %</th>
<th>Collective %</th>
<th>Organization %</th>
<th>Places %</th>
<th>Concrete Inanimate %</th>
<th>Abstract Inanimate %</th>
</tr>
</thead>
<tbody>
<tr>
<td>that</td>
<td>2</td>
<td>--</td>
<td>60</td>
<td>57</td>
<td>62</td>
<td>64</td>
<td>71</td>
</tr>
<tr>
<td>who</td>
<td>98</td>
<td>100</td>
<td>30</td>
<td>5</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>which</td>
<td>--</td>
<td>--</td>
<td>10</td>
<td>38</td>
<td>38</td>
<td>35</td>
<td>29</td>
</tr>
<tr>
<td>Total N</td>
<td>250</td>
<td>72</td>
<td>40</td>
<td>42</td>
<td>34</td>
<td>176</td>
<td>497</td>
</tr>
<tr>
<td>Total %</td>
<td>22.5%</td>
<td>6.5%</td>
<td>3.5%</td>
<td>4%</td>
<td>3%</td>
<td>16%</td>
<td>44.5%</td>
</tr>
</tbody>
</table>
proportions between the registers, such as the higher proportions of which for collectives in Maclean’s (22 percent vs. 10 percent in Hansard) and the higher proportion of which for places and organizations in Hansard (both 38 percent vs. 15-16 percent in Maclean’s).

A small number of tokens were found of relative pronoun that with human antecedents (Maclean’s N = 8; Hansard N = 5). One of the factors that predisposes a relative pronoun to be that is if the antecedent NP is either a demonstrative, as in (9a), or contains quantification, as in (9b). This will be discussed further in section 3.3. below, but a check of each of the 13 tokens of that with humans found that all the antecedent NPs were either demonstrative or quantified.

(9) a. Demonstrative: “On his feet were moccasins, the only feature of his dress that marked the Indian that was in him. [Maclean’s 1906/h]
   b. Quantifier: “every witness that had been asked for by the committee had been subpoenaed…” [Hansard 1906/F]

It is also interesting to see that organizations and places pattern together in the data, with near-identical proportions of which and that in both corpora. In the end, the determination of how best to group categories for multivariate analysis comes down to how those categories pattern with respect to one another in the data. Here, the question was whether the distribution of organizations would be more similar to plural humans or to places. Given these results, the categories of organizations and places will be combined in the multivariate analysis.

In comparison to Table 3-5 and Table 3-6, Table 3-7 shows the distribution of relative pronouns in the vernacular speech data of the Toronto corpus, from D’Arcy & Tagliamonte 2010:

**Table 3-7: Animacy of antecedent in Toronto, Subject function (adapted from D’Arcy & Tagliamonte, 2010:392)**

<table>
<thead>
<tr>
<th></th>
<th>Humans %</th>
<th>people %</th>
<th>Collectives %</th>
<th>Animals %</th>
<th>Things %</th>
</tr>
</thead>
<tbody>
<tr>
<td>that</td>
<td>45.2</td>
<td>41.9</td>
<td>71.4</td>
<td>87.1</td>
<td>96.2</td>
</tr>
<tr>
<td>who</td>
<td>50.8</td>
<td>54.6</td>
<td>24.3</td>
<td>6.5</td>
<td>--</td>
</tr>
<tr>
<td>Ø</td>
<td>3.7</td>
<td>3.1</td>
<td>5.7</td>
<td>6.5</td>
<td>3.5</td>
</tr>
<tr>
<td><strong>Total N</strong></td>
<td><strong>677</strong></td>
<td><strong>291</strong></td>
<td><strong>70</strong></td>
<td><strong>31</strong></td>
<td><strong>606</strong></td>
</tr>
<tr>
<td><strong>Total %</strong></td>
<td><strong>40.4%</strong></td>
<td><strong>17.4%</strong></td>
<td><strong>4.2%</strong></td>
<td><strong>1.9%</strong></td>
<td><strong>36.2%</strong></td>
</tr>
</tbody>
</table>

Not shown: whose and which (N = 5).
The first major difference is that which is nearly non-existent in their data, with inanimate “things” realized near-categorically with that. The second is that there is variation between that and who with human antecedents. When compared to Maclean’s and Hansard, these results once again highlight the centuries-long cross-register differences in the use of the relative pronouns that has been reported so often (Ball, 1996; Biber et al., 1999; Montgomery, 1989; Nevalainen & Raumolin-Brunberg, 2002; Romaine, 1982; Sigley, 1997; Tottie, 1997a; Tottie, 1997b). While the locus of variability in the vernacular system of subject relatives lies with the competing use of who and that with human antecedents, in the written registers, the variability lies between that and which.

It is of interest that the use of who and that for collective antecedents is similar in all registers. In Maclean’s and Hansard, the higher proportion of who with collectives (10a) sets them apart from organizations (10b) and places (10c). Use of who with such antecedents demonstrates the fuzzy nature of animacy distinctions.9

(10)  a. Collective: “Thousands of families who escaped the city in the late Forties have since been engulfed by the city’s expansion.” [Maclean’s 1956/u]
b. Organization: “…even having escaped the attentions of the universities who would have been only too happy to add him to their LL. D. lists” [Maclean’s 1906/b]
c. Place: “It was the United States who built bases in Saudi Arabia, who invaded Afghanistan and Iraq.” [Maclean’s 2006/w]

From this point, I will set aside the handful of tokens realized as who for collectives (Maclean’s N = 8, Hansard N = 12), organizations (Maclean’s N = 3, Hansard N = 2) and places (Maclean’s N = 2, Hansard N = 0), and address the more robust variation between that and which. As this variation appears only with inanimate antecedents, this subset of the data (Maclean’s N = 442, Hansard N = 789) will be the focus of the analysis in sections 3.3 – 3.6.

9 “Whereas it is true that animacy can be determined by looking at the entity an expression refers to, in practice it is not always clear what the referent of an expression is.” (Zaenen et al. 2004:10)
3.3 Internal factors: Antecedent definiteness, NP vs. Pronoun, NP modification

A number of other factors have been implicated in relative pronoun choice, and particularly in the choice between *that* and *which*. Tagliamonte et al. (2005:98) test for the effect of definiteness and whether the antecedent NP is a noun or pronoun (after previous mixed results were reported in studies by Guy & Bayley, 1995:151; Huddleston, 1971; Kautzsch, 2002:178, and Tottie & Harvie, 2000:214). They found that indefinite antecedents and pronouns favoured the $\emptyset$ form over overt *wh*-forms or *that*. The factor group of definiteness is therefore not expected to be relevant to the use of subject relatives in Maclean’s or Hansard, where the $\emptyset$ variant is never found, nor with inanimate antecedents, where *that* and *which* are the primary choices.\(^{10}\)

According to Ball (1996:234–5, with data in part from Olofsson, 1981) certain types of modification with inanimate antecedents can affect whether the relative pronoun is realized as *that* or *which*. In particular, NP antecedents containing some kind of quantification (11a) are more likely to favour *that* as the relative pronoun, while NP antecedents containing a demonstrative (11b) are more likely to favour *which*. These will be tested against NPs with neither demonstrative nor quantifier modifiers.

(11) a. NP antecedent with quantification: “it cannot be expected that any member of this House can master the details of every subject *that* comes before the House,” [Maclean’s 1906/E]

b. NP antecedent with demonstrative: “what the country needs is an expansion of funds for those institutions *which* require help” [Maclean’s 1956/J]

While there are some pronominal antecedents (Maclean’s N = 19, Hansard N = 54), they will be excluded for the purposes of this factor group, as NP antecedents make up the bulk of the data at 96 percent in Maclean’s (Total N = 452) and 93 percent in Hansard (Total N = 791).

\(^{10}\) Definiteness would be expected to interact heavily with the factor group of Antecedent Modification, a factor group that also incorporates an NP vs. Pronoun distinction. Some antecedent modifiers, such as comparatives (*the better $\emptyset$ he did*) and superlatives (*the best that there was*) are always definite.
3.4 Internal factors: Length and Complexity of Relative Clause, Adjacency

Two other internal linguistic factors have been tested. The first of these is adjacency of the relative pronoun to its antecedent. Again according to Ball (1996:236), Olofsson (1981) reported that it is possible which is slightly more preferred over that when there is intervening material between the relative clause and its antecedent (12).11

(12) a. Non-adjacent: “We are dealing with a section of the Bill which has no basis from the point of view of crime.” [Hansard 1906/B]
   b. Adjacent: “There are two religious sections that have these conscientious feelings …” [Hansard 1906/B]

Length of the relative clause has also been reported to be a factor for that vs. which, with which reportedly favoured for longer clauses (Ball, 1996:236–7). Further, Tagliamonte et al. (2005:97) suggest that this factor is also intertwined with whether the relative clause has a post-modifying clausal argument or not. The data was therefore coded for the two-way distinction of length and complexity (13).

(13) a. Short (1-3 words) and Simple (no clausal arguments): “…telling someone all about the things that are happening to us” [Maclean’s 1956/q]
   b. Long (4+ words) and Simple (no clausal arguments): “…and negotiated with Mr. Blaine the treaty which established the Paris tribunal” [Maclean’s 1906/m]
   c. Long (4+ words) and Complex (contains clausal arguments): “their apathy was replaced by an interest which soon removed many of the abuses of which Dickens wrote” [Maclean’s 1906/n]

3.5 Register factors: Lexical density (type-token ratio) and Information Status (New vs. Given)

Despite testing numerous internal linguistic factors for relevance in explaining the variation between which and that, Ball (1996:250–1, emphasis added) concludes:

…which has found little or no place in the restrictive relative paradigms of nonstandard varieties of English. It must, however, serve some functions in standard English. Olofsson (1981),

11 Tagliamonte et al. (2005:98) also consider this factor to be relevant in the case of the Ø variant: namely that an overt relative pronoun helps to delineate the boundary of the clause, and so would be preferred when the antecedent is not adjacent in order to facilitate processing. This hypothesis is not relevant for the subject relatives in the current study, since the Ø variant is never found in those contexts as it is in the British English vernacular speech data of Tagliamonte et al. (2005).
Romaine (1982), and others suggested that the relevant factor is level of formality: if this is correct, the story of *which* in the written data may simply involve stylistic change. The notion of formality, however, deserves closer scrutiny than it has received, and the stylistic value of linguistic features cannot be assumed to remain constant over time. In any case, it is evident that *which* offers many areas for further research.

Ball suggests that factors such as “informativeness” and “information status” could be part of the explanation for the variation between *that* and *which* in restrictive relative clauses with inanimate antecedents (1996:252–3, emphasis added).

…the functions of *which* in 20cE RRs remain obscure, and after factoring out several obvious sources of variation between *which* and *that*, there is considerable variation left to be explained. The time has come to move the focus of research beyond obvious, but superficial factors (e.g., clause length, distance between head and antecedent, and even genre) and to consider the effect of pragmatic properties, such as informativeness, information status, and focus.

To this end, the data has been coded for two factors that are designed to tap specifically into the information structure of the register. Recall from the discussion in section 2.4.2, testing such factors in combination with testing for the effects of internal linguistic factors allows us to determine whether the underlying grammatical system of a variable is changing (i.e. grammatical change), or whether there is a shift in how a register makes use of the variable, while the grammatical system remains constant. It is also possible for both types of change to happen simultaneously.

The first factor is lexical density, also referred to as type-token ratio (TTR). A high TTR means that there are a high number of different lexical categories present in a text, something that is indicative of the informational end of the involved/informational register dimension (Biber, 1995:117). If the lexical density of texts in a register is decreasing (or increasing) over time, this would be evidence of register change. As discussed in Chapter 2, lexical density, or type-token ratio, is also relevant in terms of measuring the likelihood of economy being a factor in variant choice (Biber, 2003; Hinrichs & Szmrecsanyi, 2007; Szmrecsanyi & Hinrichs, 2008). This need to pack as much information as possible into a finite text space is expected to lead to increased use of a more compact variant.12

While economy could be a factor for certain functions of relative pronouns (i.e. object function relatives where there is variation between overt forms *that* and *wh*-relativizers and the Ø variant),

12 This will be discussed further as it applies to s-/of-genitive variation in Ch. 4.
economy is not expected to be a factor in the variation between that and which. However, the relevance of lexical density is more than simply as a measure of the need for economy. It serves as a measure of the informational density of a text, which is indicative of the informational end of the involved/informational register dimension (Biber, 1995:117). If the lexical density of texts in a register are increasing or decreasing over time, this is evidence of register change.

This factor was coded, as described in 2.4.2 (adapted from Hinrichs & Szmrecsanyi (2007:457–8) and Szmrecsanyi & Hinrichs (2008:300–1, but also see fn 27, pg. 39), by tagging the data for part of speech and then reducing the possibilities to a coarse list of 16 lexical categories. For each token, Perl scripts were then used to count the number of unique lexical categories within 50 words on either side of a relative clause. For this data and variable, a range of between 7–15 unique lexical categories was found in the 100-word window around each relative clause token. Figure 3-6 and Figure 3-7 show the distribution of lexical density with respect to relative which in Maclean's and Hansard, respectively.

Relative which has declined 2006 in both registers, but the relationship between the use of which (versus that) and lexical density appears to be steady across time. As expected, all lines remain at roughly the same height in each time period, although in 1906 Maclean’s there is more of a gap between the highest lexical densities (12-15 unique parts of speech) and the others, while in 1956 Hansard, the highest lexical densities appear to have a slightly steeper upward slope than the others. Lexical density is not expected to be a highly relevant factor influencing variant choice. Whether this is statistically significant or not will be tested in the multivariate analysis.

---

13 An example of how the data was tagged and counted is shown in detail in Appendix B.
14 The final grouping of this factor group into the three categories of 7–10, 11, and 12–15 was somewhat arbitrary, and an artifact of how the factor group was coded. A good amount of the data clustered at the mid-point value of 11, and the categories above and below that had considerably fewer data points each than the data with the value of 11. This was also the case for this factor group in Ch. 4, and so the same grouping was used for the data there for this factor group.
The second factor that taps into changes to the informational end of the involved/informational register dimension (Biber, 1995:117) is the factor of information status, in this case, the givenness, or not, of the antecedent noun of the relative pronoun. As described in 2.4.2, more nouns (as opposed to pronouns) are indicative of “informational production” (see Ch. 1, Fig. 2), and material new to the discourse is more likely to appear as a noun than a pronoun (Fox & Thompson, 1990:304). Therefore, if there is a change over time in which a register comes to use more or fewer relative clause antecedents that are New information as opposed to Given (or vice-versa), this can be considered evidence for register change. In addition, this factor group has been hypothesized to be directly relevant to the variation between *that* and *which*. Geoffrey Pullum, in a post on Language Log says (somewhat indecisively):

> I think good writers may be choosing how to begin their integrated (= restrictive) relative clauses on the basis of a fairly subtle semantic distinction (*which* is preferred when introducing something new, and dispreferred when the material is presupposed), but I'm not sure about that. (Pullum, 2009)

The factor group was coded by use of implementing Perl scripts to search for any occurrence of the antecedent head noun of each relative clause token within the 100 words before the relative clause (see 2.4.2 and Appendix A for further details).

Figure 3-8 and Figure 3-9 show the distribution of information status with respect to relative *which* in *Maclean’s* and *Hansard*, respectively. Recall from Figure 3-2 that there are no
occurrences of relative *which* in *Maclean’s* 2006, and very few in *Maclean’s* 1956. Given the sparsity of relative *which* in 2006, we would not expect this factor group to be significant there. In *Hansard*, the proportions across time of new vs. given antecedents are near-identical in general. There is a potential for significance in 1906 *Maclean’s*, though, as the New and Given points are separated by a relatively wide margin in 1906. However, there is a crossover in 1956. If this difference over time is significant, this would be evidence of register change.

**Figure 3-8: Information Status in *Maclean’s* within previous 100-words of relative *which***

**Figure 3-9: Information Status in *Hansard* within previous 100-words of relative *which***

---

3.6 Multivariate analysis of grammatical factors: Subject relatives in *Maclean’s* and *Hansard*

Table 3-8 presents the results of the multivariate analysis of internal factors for *Maclean’s* in 1906.\(^{15}\) The analysis is presented with *which* as the application value — recall that at this point in the analysis, we are looking only at the variation between *which* and *that*. It was determined in the distributional analyses above that there was more evidence for ongoing change in the subject relatives (section 3.1) than for other functions, that within the subject function domain, it was the inanimate antecedents that exhibited the most variability (animates were overwhelmingly found

---

\(^{15}\) The external register factors were run in the same analysis, but will be presented in Table 3-10 and 3-11 below. Table 3-8 and 3-9 show only the internal grammatical factor portion of the analysis. Significant effects of the register factors will also come into play and help to further hone the model of this variation.
with *who*, and that this variation was primarily between *that* and *which* (section 3.2). Running *which* as the application value allows us to examine the patterning of the internal and external constraints on a form in the later stages of obsolescence. Proportions are given in Table 3-8 for 1956 (*Maclean’s*), where the data is too sparse to run multivariate analysis. For 2006, Ns are provided for completeness, even though only *that* is found with non-human antecedents at that time. Table 3-9 presents the multivariate analysis of *which* in the three time periods in *Hansard*.

**Table 3-8:**

<table>
<thead>
<tr>
<th>Contribution of grammatical factors to the probability of <em>which</em> as relative pronoun with NON-HUMAN antecedents: <em>Maclean’s</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1906</strong></td>
</tr>
<tr>
<td><strong>Input:</strong> 0.41</td>
</tr>
<tr>
<td><strong>ANIMACY</strong></td>
</tr>
<tr>
<td>Concrete Inanimates</td>
</tr>
<tr>
<td>Organizations &amp; Places</td>
</tr>
<tr>
<td>Abstract Inanimates</td>
</tr>
<tr>
<td>Collectives</td>
</tr>
<tr>
<td><strong>range</strong></td>
</tr>
<tr>
<td><strong>ANTECEDENT MODIFICATION</strong></td>
</tr>
<tr>
<td>Demonstrative</td>
</tr>
<tr>
<td>NP without modification</td>
</tr>
<tr>
<td>Pronoun without modification</td>
</tr>
<tr>
<td>Quantifier</td>
</tr>
<tr>
<td><strong>range</strong></td>
</tr>
<tr>
<td><strong>ADJACENCY</strong></td>
</tr>
<tr>
<td>Adjacent</td>
</tr>
<tr>
<td>Non-adjacent</td>
</tr>
<tr>
<td><strong>range</strong></td>
</tr>
<tr>
<td><strong>LENGTH x COMPLEXITY</strong></td>
</tr>
<tr>
<td>4+ words, post-modification</td>
</tr>
<tr>
<td>4+ words, no post-modification</td>
</tr>
<tr>
<td>1-3 words, no post-modification</td>
</tr>
<tr>
<td><strong>range</strong></td>
</tr>
</tbody>
</table>

Excluded N = 8 tokens of *who* with Collectives, N = 3 with Organizations and N = 2 with Places

* = these tokens were excluded for the multivariate analysis of this factor group only due to low Ns

---

16 Tokens that have removed from the multivariate analysis from within a single factor group (indicated by * in Table 3-8, as well as in Table 3-9 and Table 4-2, below) were removed using the ‘slash’ function of GoldVarb. The marginal numbers have been left in the tables for illustrative purposes.
Comparing the two analyses, we see that length-by-complexity is the only significant factor group, and it is only significant in *Hansard*. This is likely due to the fact that this register, being edited transcripts of speech, does involve an element of on-line processing. This factor group remains significant and with its constraint hierarchy intact in 1956. *Which* continues to be favoured by longer and more complex clauses. By 2006, where the input value of *which* has decreased dramatically, no grammatical factors are significant. In *Maclean’s*, where *which* is already almost gone by 1956, none of the internal grammatical factors are significant in 1906, despite the respectably high input probability of 0.41, indicating that the loss of grammatical
conditioning for *which* was more advanced in that register — this abrupt decline of *which* will be discussed further below. Overall, the multivariate analysis for *which* confirms and shows retreat of a variant that is on its way out and being replaced by *that*. We are seeing the end of this process, with *Maclean’s* in the lead.

**Table 3-10:**

| Contribution of register factors to the probability of *which* as relative pronoun with NON-HUMAN antecedents: *Maclean’s* |
|---|---|---|
|  | 1906 | 1956 | 2006 |
|  | Input: 0.41 | Input: 0.07 | N: 156 |
|  | N: 139 | N: 141 |
| LEXICAL DENSITY (TTR) |  |  |  |
| 7–10 | [.55] | 45% | 51 |
|  | [.42] | 34% | 65 |
| 11 | [.57] | 48% | 40 |
| 12–15 | range |  |  |
| INFORMATION STATUS |  |  |  |
| Given | .68 | 60% | 25 |
| New | .46 | 37% | 131 |
| range | 22 |

Moving on to the analysis of the register-internal factors, Table 3-10 presents the multivariate analysis of register-internal factors for *which* in *Maclean’s*. The most interesting result here is the significance of information status in 1906 *Maclean’s* in Table 3-10, with *which* being preferred in contexts where the antecedent head noun was previously given in the discourse (here, within the previous 100 words), as in (14):

(14) A bigger field lay before him; a field *which* seemed to possess boundless possibilities and called for an entire concentration of his energies on practical development.

[*Maclean’s 1906/a*]

By 1956, a notable change has occurred. The form *which* has dramatically declined in frequency (see the input value of 0.07, vs. 0.41 in 1906) to the point where none of the register factors tested are significant (and the internal grammatical factors, shown in Table 3-8, could not be run at all due to the sparsity of the data). Some insight into the nature of this abrupt decline is visible in the hierarchy of constraints in the proportional marginal data for the register constraints in Table 3-10. The information status constraint is reversed: the few tokens of *which* that are present (N = 9/139) are only ever found with information that is new to the discourse (no factor...
weights are provided because of the categoricity). Lexical density is never significant in the analysis. Taken with the steep obsolescence of *which* over the first half of the 20th century, this is evidence that the loss of *which* is a register change in *Maclean’s*.

One possible explanation is that this is due to editorial conventions that came into play in the mid-20th century, when restrictive *which* is known to have begun being prescriptively edited out of the press writing (Fowler, 1926:635; Sigley, 1997:208; Tottie, 1997a:85–7; see also the discussion above in section 3.2). The results of this are visible by 2006 *Maclean’s*, where restrictive *which* has been completely eliminated. Prior to that, when it was still used, it was not grammatically conditioned as it was in 1906 and 1956 *Hansard* (see Table 3-9). Rather, it was conditioned in *Maclean’s* by a higher-pragmatic factor more closely tied to the information structure of the discourse and register, namely whether the antecedent head noun was Given (shown by the favouring factor weight of .68) or New (shown by the disfavouring factor weight of .46).

It is interesting to note that the effect of information status in 1906 *Maclean’s* is opposite what was predicted by Pullum (2009); *which* is preferred there for information that has been previously given. It then changes over in 1956 to be consistent with Pullum’s hypothesis. The

**Table 3-11:**

<table>
<thead>
<tr>
<th>INFORMATION STATUS</th>
<th>1906</th>
<th>1956</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GLEXICAL DENSITY (TTR)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7–10</td>
<td>[.44]</td>
<td>33%</td>
<td>63</td>
</tr>
<tr>
<td>11</td>
<td>[.56]</td>
<td>45%</td>
<td>96</td>
</tr>
<tr>
<td>12–15</td>
<td>[.48]</td>
<td>37%</td>
<td>90</td>
</tr>
<tr>
<td><strong>INFORMATION STATUS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Given</td>
<td>[.46]</td>
<td>36%</td>
<td>76</td>
</tr>
<tr>
<td>New</td>
<td>[.52]</td>
<td>41%</td>
<td>173</td>
</tr>
</tbody>
</table>

Excluded N = 12 tokens of *who* with Collectives and N = 2 with Organizations
picture is somewhat different for *Hansard*, in Table 3-11. There we find no significant effects of the register factors.

The question is, what does this tell us about the origin of the register change leading to the loss of *which*? Is it possible that *Hansard*, with its stable constraint hierarchy for information status, is ahead of *Maclean’s* in the progress of the change? The important piece of the puzzle here is that it is quite possible that the effect of information structure was always different in the two registers. *Hansard*, according Biber’s register dimension described in chapter 1 (see Table 1-1) is not as “informational” as *Maclean’s* would be expected to be, meaning that a factor group that taps into nominal features would not be expected to be as relevant in *Hansard* as in *Maclean’s*. The multivariate analysis of register-internal factors in *Hansard* in Table 3-11 confirms this. Despite consistency in the constraint hierarchies for the register-internal factors, none of those factors is significant in any time period. Taken together, then, the results from Table 3-10 and Table 3-11 are evidence that the register change leading to the loss of *which* is led by the press register of *Maclean’s*.

Further evidence of *Maclean’s* leading the loss of *which* comes from the grammatical conditioning in *Hansard*. The use of *which* was grammatically conditioned only by the processing factor constraint — length and complexity of the relative clause. This formal spoken register seems to represent an earlier state of the conditioning of the variable, lagging behind and perhaps more formal than *Maclean’s*. Restrictive *which* is still found in *Hansard* in 2006, but only with abstract and concrete inanimates. It is no longer found with places, organizations, or collectives. The grammatical conditioning there has weakened, likely due to the extremely low frequency (Cukor-Avila, 1997), to the point that no factors are significant in multivariate analysis. Register-internal factors are never significant.

More evidence of the weakening of the grammatical conditioning of *which* comes from Tagliamonte & D’Arcy (2010:389–91), who find only four instances of restrictive *which* out of a total of 3220 restrictive relative clauses (1675 subject function). All four tokens of *which* are
subject function. Interestingly, these four tokens represent a 4-person spread by age (young vs. old) and gender (male vs. female):  

(15) a. And I would try to ignore it, but some- some of my malicious friends which say “(singsong) W----, your father’s calling you!” (N/ø, female age 78)  
b. People corrupt their damn selves or they hang out with the wrong people which do not exclusively live in the city… (N/r, female age 22)  
c. Yeah, well fortunately uh, I had these little jobs which gave my uh, maybe a dollar a week or something like that… (N/d, male age 83)  
d. Which is another thing which makes me upset. Like, when people go to like, American franchises, like Nike and Adidas and stuff. (2/c, male age 16)  

This indicates that although restrictive which is edited out of writing and is no longer a productive variant in either formal speech or the vernacular, it still persists under certain circumstances. It is not, however, restricted to a particular context by animacy, as we might expect: in (14a), where the antecedent is plural human, and in (14b), the antecedent is the collective plural ‘people’. In (14c) and (14d), the antecedents are inanimate — an activity (job) and the abstract inanimate thing (also referring to an activity).

3.7 Recap and conclusions

What is the story of which? We have seen that the main grammatical constraints of antecedent function and animacy are so well-entrenched that they are near-categorical for which — and indeed for the relative pronouns in general — such that these contexts constitute separate variable contexts (i.e. who for animates, that and which for inanimates). Further, there is no evidence of grammatical change due to the influence of other grammatical factors tested (see Table 3-8 and Table 3-9). Rather there appears to be a general weakening of the grammatical constraints on which — a pattern that is expected as one form replaces another (Kroch, 1989) — as it disappears in all grammatical contexts and undergoes lexical replacement by that. The effect of information status has changed over time in Maclean’s, which is evidence of register change, and the cross-register comparison shows that it is Maclean’s leading this change.

17 These tokens of which are not discussed in D’Arcy & Tagliamonte (2010) beyond the N being included in their overall distribution, as the focus of that study was restrictive who. These examples have been extracted here from their original data file.
Is *which* completely disappearing? It would seem not — *which* is currently only completely eradicated in the more heavily prescriptively edited *Maclean’s*. In the spoken registers, both in *Hansard* and Toronto vernacular speech, *which* persists, but without clear grammatical conditioning, as might be expected given that “variation among relative pronouns in contemporary spoken English is a highly marked sociolinguistic feature” (D’Arcy & Tagliamonte, 2010:396, emphasis added). Moreover, once an obsolescing form drops below a certain level of frequency, the grammatical constraints on its use may cease to operate in a significant way.\(^{18}\) Indeed, we are once again brought back to Ball’s (1996:250) point that *which*: 

... must, however, serve some functions in standard English. Olofsson (1981), Romaine (1982), and others suggested that the relevant factor is level of formality: if this is correct, the story of *which* in the written data may simply involve stylistic change.

The analysis presented here shows some quantitative evidence of stylistic change. First, lexical replacement of *which* by *that*, possibly due to editorial standardization and ‘colloquialization’, or the shift towards a register that more closely approximates speech, within press reporting (cf. Hundt & Mair, 1999; Leech, Hundt, Mair & Smith, 2009: 239). The end result is a(n internal) register change. This is visible at least in part though two pieces of evidence: 1) the complete loss of *which* overall in *Maclean’s* as opposed to its continued presence in *Hansard* (without grammatical conditioning) and in vernacular speech, and 2) the shift of the once-conditioning factor of Information Status in *Maclean’s* (Table 3-10) to briefly favour new rather than given information in 1956, before the more stylistically formal *which* disappears completely by 2006.

The grammatical system of the relative pronouns was stable before the loss of *which*, and there is no evidence here to support an analysis of current grammatical change, only the weakening of grammatical conditioning (Tables 3-8 and 3-9) as *which* disappears. This variable is therefore an example of register change without a co-occurring process of grammatical change, a distinction

\(^{18}\) Note, though, that the patterning of those constraints can sometimes still be visible in the constraint hierarchy of the non-significant factor weights (Tagliamonte & Jones, 2004:119; Cukor-Avila 1997). See, for example, the consistency of the constraint hierarchy of the non-significant factor weights for Length x Complexity in 2006 *Hansard* compared to the previous time periods (Table 3-9), and the same hierarchy for this factor in *Maclean’s* 1906 (Table 3-8). Also, the non-significant trend for new information to favour *which* in *Hansard* (Table 3-11), in the direction of the effect predicted by Pullum. The consistency of this trend is all the more surprising because of the low amount of *which* in *Hansard* at this time.
that could not have been made if register-internal factors were not taken into account, and a cross-register perspective presented.

The apparent trajectory of the decline of restrictive *which*, though, is interesting. As one of the WH-relatives, *which* was always a higher-prestige form in English. It originated as a change from above that appeared first in written registers, yet never completely took hold in the vernacular (Romaine, 1982; Tottie, 1997a). We now have further quantitative evidence that the decline of *which* is also a change from above, enforced by socio-stylistic, register-based constraints and not internal grammatical ones. In chapter 4, we look at a different type of situation in which grammatical and register change can be entangled, with the analysis of *s-* vs. *of-*genitive.
Chapter 4
Genitive variation

4 Introduction

The empirical study presented in this chapter will provide a second example of the methodology and hypotheses laid out in chapter 2 to test for both grammatical and register change. While the s-/of-genitive variable has been widely studied in American and British English varieties (Hinrichs & Szmrecsanyi, 2007; Ljung, 1997; Rosenbach, 2002, 2003, 2005; Szmrecsanyi & Hinrichs, 2008), it is only now beginning to be studied in Canadian English (Tagliamonte & Jankowski, forthcoming, for vernacular speech).

4.1 Variable s– vs. of-genitive

The s-genitive has been increasing in the 20th century (see Hinrichs & Szmrecsanyi, 2007; Ljung, 1997; Rosenbach, 2002, 2003, 2005; Szmrecsanyi & Hinrichs, 2008; Tagliamonte & Jankowski, forthcoming), and importantly, it varies by register (see section 2.1 and examples therein).

All tokens of variable s- and of-genitive were extracted from the Maclean’s (N=1318) and Hansard corpora (N=1308). Further examples of the variable are given in (1)–(4):

(1) … all Canadians should stand equal before the trials of life and that all Canadians should benefit equally from life’s opportunities. [Maclean’s 2006/0]
(2) I desire to call the attention of the minister to the subject of dredging in Toronto harbour. I called the minister’s attention to this some time ago…[Hansard 1906/C]
(3) When he joined the board of GM, Sullivan launched a crusade […] Sullivan used his position on GM’s board to tear away at apartheid’s barriers […] In 1990, when the legal structure of apartheid was officially dismantled… [Maclean’s 2006/Z]
(4) a. I understood the policy of the government to be… [Hansard 1906/C]
b. …opposing the government’s present policy on the marketing of butter… [Hansard 1956/O]

Only fully interchangeable genitive constructions were used in the analysis. Yet there are many possessive constructions in English grammar that cannot be used with either form. Any instance of genitive s- or of- that could not be substituted for the other variant with no more than the removal of the determiner on the possessum for the s- variant (or the addition of one for the of-variant), was excluded, as were set phrases (i.e. University of Toronto, Peggy’s Cove, Murphy’s
A further criterion was that there could be no change in meaning; for instance, *hothouse growths of the Old World* does not mean the same thing as *the Old World’s hothouse growths* (similarly, *other facets of the debate ≠ the debate’s other facets*), since the possessum is always definite in the s-genitive.²

Figure 4-1:

Distribution of the s-genitive across time and register is shown in Figure 4-1. Overall, the s-genitive has increased in proportion over the 20th century in both registers, with *Maclean’s* using

1 The possessive pronoun alternation (i.e. *my book vs. the book of mine, its problems vs. the problems of it*, etc.) is not part of the variable context. It is worth noting that that these constructions can serve as the nominal head of an interchangeable genitive construction themselves (i.e *my book’s pages/the pages of my book* (which are considered part of the variable context), vs. *the book of mine’s pages/the pages of the book of mine* (which would be considered a phrasal genitive with post-modification according to the exclusion criteria discussed above).

2 Exclusions follow the methodology of Hinrichs & Szmrécsey 2007 and Ljung 1997. They also include units of measure (“the size of a pea”), partitives (“the middle of the road”), so-called “descriptive” genitives (“parents’ night at the school”), and (forests of black spruce and maples). Set phrases included proper municipal and corporate names, well-known idioms (i.e. ‘Murphy’s Law’), and a few high-frequency items that were found to never vary within a corpus — other notable set phrases that were excluded were “people/citizens of Ontario/Toronto” and “the (hon.) Member of this House”, which never vary in *Hansard*. “Phrasal genitives,” or of-genitives with post-modification (i.e. “humble in the knowledge of the responsibilities which accompany that membership”, “increases to the range of community resources that are available”, “to determine the extent of the meth problem in Ontario”), were also excluded. Hinrichs & Szmrécsey (2007:447) and Ljung (1997:30) point out that these are proscribed against in standard English and so would be edited out of press data. Using native speaker judgments, I often found these acceptable for interchangeability, even in written English. However, they were excluded so as to maintain a methodological consistency.
it more. This is consistent with previous studies that show the s-genitive increasing in real time over the 20\textsuperscript{th} century in press and spoken data (Hinrichs & Szmrecsanyi, 2007; Szmrecsanyi & Hinrichs, 2008; Tagliamonte & Jankowski, forthcoming), and further supports the hypothesis that the press will lead the expansion of the change into written registers. Hansard, though a partially spoken register, appears to be the more conservative with regard to the increase of the incoming s-genitive.

4.2 Internal factors

Tokens were coded for four internal linguistic factors known to influence variant choice: animacy of the possessor NP, end weight of the possessor and possessum phrases, and the phonological factor of final sibilant in the possessor. The first of these, animacy of the possessor NP, is perhaps the best known constraint on variant choice. The s-genitive is expected to be favoured by possessors that are human, or following that animate in some way (Ljung, 1997:25), though Hinrichs & Szmrecsanyi (2007:440) point out that it has also been suggested that this constraint is shifting, and that s-genitives are coming to be used with inanimates and collective nouns:

There is no consensus whether the shift from of to 's is due to changes in the animacy constraint: some authors attribute it to a spread of the form to inanimate possessor noun phrases (tentatively also Denison, 1998, e.g. Jespersen, 1909-49:VII, 327–328), while Mair (2006a:147, 2006b) claims that the animacy constraint is currently being loosened for collective nouns, not inanimates, and that furthermore, the more significant causes of the spread of the s-form lie in the area of discourse practices, \textsuperscript{[3]} not the underlying constraint grammar.

The animacy of each possessor NP was coded according to categories discussed by Ljung (1997:24–25) and Hinrichs & Szmrecsanyi (2007:449–50), and based on the same animacy hierarchy of Zaenen, et al. 2004 as the relative clause study of the previous chapter (see section 3.2, fn 4 and Appendix D).

While previous studies have, for the most part, grouped together all inanimate categories (Hinrichs & Szmrecsanyi, 2007:449; Ljung, 1997:30; Szmrecsanyi & Hinrichs, 2008:298, but see

\textsuperscript{3} By “discourse practices”, Mair (2006a:147) is referring to a need “for compressing information” in the press reportage genre, described previously as economy. This factor will be tested in the analysis.
Tagliamonte & Jankowski, forthcoming, discussed below in section 4.6), there are claims (Rosenbach, 2005:615, 2003: 386), that the $s$-genitive is common with certain subsets of inanimate nouns, such as places (in Rosenbach: “geographic locations”, i.e. Toronto’s) and units of time (i.e. tomorrow’s). In the Maclean’s and Hansard corpora, only the categories of humans (singular & plural), organizations and places\(^4\) show variation for all time periods. Both Maclean’s and Hansard are fortuitously rich in the inanimate category of places — this category comprises a full 30 percent of the $s$-genitive tokens in both corpora — so Rosenbach’s claim can be tested in a way not previously done, namely over a 100-year time span and across multiple registers. All other inanimate categories, as well as non-human animates, are overwhelmingly realized with the of-genitive.\(^5\) Only the categories of singular humans, plural humans, collective humans, organizations and places are therefore considered for analysis.\(^6\)

Examples of how possessor and possessum weight were counted are shown in (5), for possessors that are shorter than the possessums and (6), for possessors that are longer than possessums. Possessors are in bold and possessums underlined. According to \textbf{end weight}, longer and more complicated constituents will come after shorter ones. A shorter possessor is more likely to take an $s$-genitive and a longer possessor to take an of-genitive (5a–c), while a shorter possessum will be more likely to take an of-genitive, and a longer possessum to take the $s$-genitive (6a–c).

\begin{itemize}
  \item \textbf{In particular, municipal places} such as Toronto’s or Canada’s (N = 220 $s$-genitive/705) currently make up most of the place possessors, and are generally shorter in length than non-municipal places (N = 2 $s$-genitive/104), which show very little variation. Non-municipal places have therefore been excluded from the analysis. In addition, in many cases, non-municipal places could arguably have been coded as geographic features (i.e. riverbank) or inanimate objects (i.e. low-rise apartment complex).
  \item Total Ns for the $s$-genitives in these categories, for both corpora and all time periods combined, are also quite low: temporal (N = 2/165), inanimate objects (N = 27/712), activities (N = 16/576), states (N = 4/233), body parts (N = 0/7), and animate non-humans (N = 7/21). These categories were therefore excluded from the analysis.
  \item Total Ns for the $s$-genitives in these categories are: singular humans (N = 269/358 Maclean’s; 103/273 Hansard), plural humans (N = 20/97 Maclean’s; 10/116 Hansard), collectives (N = 31/91 Maclean’s; 11/142 Hansard), organizations (N = 94/290 Maclean’s; 48/346 Hansard) and places (N = 20/97 Maclean’s; 65/285 Hansard). These include proper nouns in both animate (proper names of humans (N = 247/391) and inanimate (proper/municipal places names (N = 220/633) contexts. When compared, proper nouns and non-proper nouns patterned together in both Maclean’s and Hansard for animates and so were not separated for analysis. This differs from patterns found in vernacular speech (Tagliamonte and Jankowski, forthcoming), where proper names for humans are near-categorically found with the $s$-genitive, e.g. Nancy’s friend, my sister’s friend, etc.
\end{itemize}
However, this does not always happen — see (7a), and the more extreme case in (7b) (Hinrichs & Szmrecsanyi, 2007:453–5; Szmrecsanyi & Hinrichs, 2008:292–3, 299–300):

(5) a. life's opportunities (1-word possessor, 1-word possessum)  
   b. Canada's first responsible government (1-word possessor, 3-word possessum)  
   c. the Ontario Provincial Police's crime prevention and electronic crime section (4-word possessor, 6-word possessum)

(6) a. the son of Hiram and Martha Fulford (4-word possessor, 1-word possessum)  
   b. the gravest evils of the situation (2-word possessor, 2-word possessum)  
   c. the backs of the Canadian farmer and artisan (5-word possessor, 1-word possessum)

(7) a. a big man's share (3-word possessor, 1-word possessum)  
   b. the 21-year-old Thunder Bay, Ontario reservist's grieving father (8-word possessor, 2-word possessum)

Figure 4-2:

---

7 In 1906 Hansard, the only 1-word possessors found were in fixed contexts such as citizens of Canada, which never appear with an s-genitive.
4.2.1 Interaction of internal factors

It is not possible to separate the factor of possessor length from animacy. Figure 4-2 (*Maclean’s*) and Figure 4-3 (*Hansard*) show the remaining relevant animacy categories used in the analysis by possessor length. Several observations can be made from these figures. First, plural humans and collectives pattern together, no matter the length of the possessor. Places and organizations also pattern together, with the exception that 1-word organization possessors appear more with s-genitive in *Maclean’s*. Keeping this exception in mind, plural humans and collectives will be grouped together for the remainder of the analysis, as will organizations and places.  

Figure 4-4 and Figure 4-5 demonstrate that this increase in the proportion of s-genitive is not due to an overall increase in proportion of place- (Figure 4-4) or organization-possessors (Figure 4-5) over time.

---

8 The relevance of grouping together places and organizations was further motivated in the relative clause study in Chapter 2 (see section 3.2), based on the potential of places to sometimes be thought of as organizations. It is important to note, though, that decisions about collapsing factor groups are based on the distribution of data for each variable independent of the other. For instance, in the relative clause study, collective and plural human antecedents were not grouped together, because their distributions of use for the relative pronoun variants differed considerably.
However, there is an overall tendency in both corpora for one-word possessors to increase over time, as shown in Figure 4-6.
Figure 4-6:

Investigating this further shows that, in fact, it is the one-word organization and place-possessors that are responsible for this increase. Figure 4-7 plots the proportion of one-word organizations and places realized with s-genitive over time in both corpora. The sharp upward slope over time for all categories shows that for these animacy categories, one-word possessors are on the rise over time. The increase happens first in *Maclean’s*. As per the discussion in section 2.4.1, this change in proportion of one-word possessors within a particular sub-context of the overall genitive variable context is an indication of register change. This observation will be tested in multivariate analysis.
Figure 4-7:

Figures 4-4–4-7 demonstrate the interaction between animacy and possessor length — specifically, the combined tendency of place-and organization-possessors to be one word in length, and for one-word possessors to be more frequent in recent years. In order to take this interaction into account, animacy and possessor length were combined and, after further comparison collapsed down to five categories, with the major distinction falling between 1-word and 2-word place possessors.

Figure 4-8 and Figure 4-9 show the animate possessor categories (humans and collectives) by possessor length in both corpora over time. The expected animacy constraint on genitive choice holds: singular animates in Maclean’s (Figure 4-8) are most often used with the s-genitive in all time periods. This constraint is also operational in the Hansard data (Figure 4-9), but not to the same degree. Figures 4-8 and 4-9 show that the proportion of s-genitive increases steadily over time in all singular animate categories, but not in plural animate categories. These categories will be kept separate in the multivariate analysis over time. There is also a strong tendency for singular animate possessor NPs to be one-word in length, while plural animate possessors show no such tendency.
Figure 4-8:

Proportion of s-genitive for animate categories by year in Maclean's

Figure 4-9:

Proportion of s-genitive for animate categories by year in Hansard

Figure 4-10 and Figure 4-11 are the inanimate possessor categories by possessor length in both corpora across time. As expected (given Figure 4-7, above), the sharpest increases are in the proportion of s-genitive used for 1-word inanimate organization- and place-possessor NPs, and
Figure 4-10:

Proportion of s-genitive for inanimate categories by year in Maclean's

Figure 4-11:

Proportion of s-genitive for inanimate categories by year in Hansard

1-word inanimate organization- and place-possessor NPs. These results, along with those of Figure 4-8 and Figure 4-9, are also in line with results from spoken vernacular Ontario English in apparent time, where the s-genitive is showing stability over time with human possessors, and increase over time with inanimate possessors (Tagliamonte & Jankowski, forthcoming):

---

9 Organizations and places were not broken down further into singular and plural, as the two are, for the most part, semantically singular.

10 This comparison will be discussed specifically in section 4.6, below.
Finally, Figures 4-8–4-11 also show that plural animates (see Figure 4-8 and 4-9) and organizations (see Figure 4-10 and 4-11) most often use the *of*-genitive in all time periods. Overall, these distributional differences by animacy make it apparent that animate and inanimate contexts are separate variable contexts for genitive variation. Animate and inanimate possessors will therefore be treated separately for the remainder of the analysis.

### 4.2.2 Prototypical vs. non-prototypical possession and presence of final sibilant in the possessor

Another distinction to consider in genitive variation involves the semantics of the possessor relationship. Wolk, Bresnan, Rosenbach and Szmrecsanyi (to appear:14) make a distinction between *prototypical* vs. *non-prototypical* possessive relation, based on Rosenbach 2002. Examples, based on their classifications, are given in (8). All other possessor types are considered “non-prototypical.”

(8)  
<table>
<thead>
<tr>
<th>Possessor Type</th>
<th>Example</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. legal ownership</td>
<td>“…University of Saskatchewan history professor and <em>author of ’So Little For The Mind’</em>”</td>
<td>Maclean’s 1956/o</td>
</tr>
<tr>
<td>b. body parts:</td>
<td>“…resulting in <em>partial paralysis of the lower limbs</em>.”</td>
<td>Hansard 1956/K</td>
</tr>
<tr>
<td>c. kinship:</td>
<td>“I only wish <em>some good live Canadian farmer’s daughter</em> would get hold of the minister and wake him up…”</td>
<td>Hansard 1906/A</td>
</tr>
<tr>
<td>d. whole-part relationship:</td>
<td>“In regard to <em>the first four lines of this subsection</em>…”</td>
<td>Hansard 1906/E</td>
</tr>
<tr>
<td>e. whole-part relationship:</td>
<td>“…put it in <em>the trunk of their car</em> and took it to a friend.”</td>
<td>Maclean’s 1956/r</td>
</tr>
<tr>
<td>f. prototypical with inanimate:</td>
<td>“at least up to the time they got to <em>the ship’s side</em>…”</td>
<td>Hansard 1906/F</td>
</tr>
</tbody>
</table>

Wolk et al. (to appear:14) suggest that prototypical possessors, such as those in (8), should be expected to favour *s*-genitive. The examples in (8) could therefore be considered misleading in this regard, since only two of the six are examples of the *of*-genitive (8c, 8f). This is because in the *Maclean’s* and *Hansard* data, prototypical possessors are more likely to be found with the *of*-genitive.\(^\text{11}\)

---

\(^{11}\) The total number of prototypical tokens in the data (*Maclean’s* N = 188, *Hansard* N = 118) was insufficient to justify separation of the different subtypes of prototypical genitives in multivariate analysis, especially once they are further subdivided into animate and inanimate (Figure 4-12 and Figure 4-13).
The explanation for this apparent issue is likely due to a strong interaction with animacy. Figure 4-12 (Maclean’s) and Figure 4-13 (Hansard) show the distribution of the data by both the semantics (prototypical or non-) and animacy of the possessor. Non-prototypical possession makes up the bulk of the data: in Maclean’s, it represents 84 percent of animates and 87 percent of inanimates, while in Hansard, non-prototypical possession represents 89.5 percent for animates and 92 percent of inanimates. Notice that in Hansard, prototypical inanimates (N = 62, as in example (8f)) are so rarely found with s-genitive as to make them near-categorically of-genitive.

The final internal constraint on genitive choice tested is the presence or absence of a sibilant as the final phoneme of the possessor. The presence of the sibilant would be expected to discourage an immediately adjacent sibilant, due to restrictions of the Obligatory Contour Principle (Goldsmith, 1976; Leben, 1973; McCarthy, 1986) and what has been referred to as the principle of horror aequi, and so discourage use of the s-genitive (Hinrichs & Szmrecsanyi, 2007:452–3; Rohdenburg, 2003).

Figure 4-14 and Figure 4-15 show the distribution of the s-genitive in Maclean’s and Hansard by whether the final phoneme of the possessor is a sibilant or not, demonstrating that this constraint

---

12 According to Rohdenburg (2003:236) and Rudanko (2003:662), this is “the universal tendency to avoid the (near-)adjacency of identical grammatical structures”
is in operation in both corpora. Possessors ending in a sibilant are always less likely to take the s-genitive. The strength of the constraint, though, is much stronger in the final time period.

From the comparative analysis presented here, there are five factors that will be subjected to multivariate analysis in the next section: animacy of the possessor, length of the possessor, length of the possessum, prototypical vs. non-prototypical possession, and presence or absence of a final sibilant in the possessor.

### 4.3 Multivariate Analysis of grammatical factors: animate possessors

For the purposes of multivariate analysis, animate and inanimate possessors have been run in separate analyses, as motivated by the distributional analysis present in section 4.2.1. Table 4-1 (*Maclean’s*) and Table 4-2 (*Hansard*) show the results of independent multivariate analyses (according to the methodology detailed in Chapter 2) of the internal factors discussed above, one for each corpus and time period. The interaction factor group comprised of animacy and

---

13 As an additional check, multivariate analyses were also run for both corpora with all time periods together (not shown), such that the factor groups of time and register were included in the analysis. The factor of register was found to be significant, and the factor of time was found to be significant within both corpora.
Table 4-1:

<table>
<thead>
<tr>
<th>Contribution of grammatical factors to the probability of s-genitive with animate possessors: Maclean’s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>POSSESSOR TYPE</td>
</tr>
<tr>
<td>1-wd. Human Singular</td>
</tr>
<tr>
<td>2+wd. Human Singular</td>
</tr>
<tr>
<td>1-wd. Plurals + Collectives</td>
</tr>
<tr>
<td>2+wd. Plurals + Collectives</td>
</tr>
<tr>
<td><strong>range</strong></td>
</tr>
<tr>
<td>FINAL SIBILANT IN POSSESSOR</td>
</tr>
<tr>
<td>No sibilant</td>
</tr>
<tr>
<td>Sibilant</td>
</tr>
<tr>
<td><strong>range</strong></td>
</tr>
<tr>
<td>POSSESSION TYPE</td>
</tr>
<tr>
<td>Non-prototypical</td>
</tr>
<tr>
<td>Prototypical</td>
</tr>
<tr>
<td><strong>range</strong></td>
</tr>
<tr>
<td>POSSESSUM LENGTH</td>
</tr>
<tr>
<td>2 words</td>
</tr>
<tr>
<td>3+ words</td>
</tr>
<tr>
<td><strong>range</strong></td>
</tr>
</tbody>
</table>

Not significant: Possessum Length, Final Sibilant

Possessor type always has the strongest effect in Maclean’s, where it always has the highest range (i.e. 62/44/89). In Hansard, there is a shift in the magnitude of the effect in 1956 and 2006, when final sibilant becomes near-equal (in 1956, range of 31) and then the strongest significant effect (in 2006, range of 63). The constraint hierarchy for significant factors is stable across time in both corpora, though, with human possessors, no final sibilant and non-prototypical possession always favouring the s-genitive. The constraint hierarchy is not the same between the

---

14 The Ns for 1-word singular humans in Hansard are very low, and so were excluded from the multivariate analysis in all time periods. Similarly, the 1-word plurals and collectives were excluded in 1906 and 1956. Although the Ns for 3+ word possessum lengths were also very low in both Table 4-1 and 4-2, the factor group was never significant in the analysis of animates. They are left in here for the purpose of illustration.
registers, however, as 1-word plurals and collectives favour s-genitive in 2006 *Hansard*, but disfavour it in *Maclean’s*. Finally, note the 1-word singular human possessors in *Maclean’s* — by 1956, these have shifted to be near categorical with the s-genitive.  

**Table 4-2:**

<table>
<thead>
<tr>
<th>POSSESSOR TYPE</th>
<th>1906</th>
<th></th>
<th>1956</th>
<th></th>
<th>2006</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FW</td>
<td>%</td>
<td>N</td>
<td>FW</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>1-wd. Human Singular</td>
<td></td>
<td>50%</td>
<td>4</td>
<td></td>
<td>50%</td>
<td>3</td>
</tr>
<tr>
<td>2+wd. Human Singular</td>
<td>.75</td>
<td>32%</td>
<td>123</td>
<td>.66</td>
<td>22%</td>
<td>73</td>
</tr>
<tr>
<td>1-wd. Plurals + Collectives</td>
<td>*</td>
<td>3%</td>
<td>3</td>
<td>*</td>
<td>30%</td>
<td>9</td>
</tr>
<tr>
<td>2+wd. Plurals + Collectives</td>
<td>.15</td>
<td>3%</td>
<td>77</td>
<td>.33</td>
<td>6%</td>
<td>71</td>
</tr>
<tr>
<td><strong>range</strong></td>
<td></td>
<td>50</td>
<td></td>
<td>33</td>
<td></td>
<td>42</td>
</tr>
<tr>
<td>FINAL SIBILANT IN POSSESSOR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No sibilant</td>
<td>[.50]</td>
<td>22%</td>
<td>166</td>
<td>.61</td>
<td>21%</td>
<td>98</td>
</tr>
<tr>
<td>Sibilant</td>
<td>[.50]</td>
<td>15%</td>
<td>41</td>
<td>.33</td>
<td>7%</td>
<td>58</td>
</tr>
<tr>
<td><strong>range</strong></td>
<td></td>
<td>31</td>
<td></td>
<td></td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>POSSESSION TYPE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-prototypical</td>
<td>[.51]</td>
<td>21%</td>
<td>191</td>
<td>[.49]</td>
<td>15%</td>
<td>141</td>
</tr>
<tr>
<td>Prototypical</td>
<td>[.34]</td>
<td>13%</td>
<td>16</td>
<td>[.58]</td>
<td>20%</td>
<td>15</td>
</tr>
<tr>
<td><strong>range</strong></td>
<td></td>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POSSESSUM LENGTH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 word</td>
<td>[.49]</td>
<td>21%</td>
<td>178</td>
<td>[.49]</td>
<td>16%</td>
<td>129</td>
</tr>
<tr>
<td>2 words</td>
<td>[.55]</td>
<td>14%</td>
<td>22</td>
<td>[.55]</td>
<td>15%</td>
<td>27</td>
</tr>
<tr>
<td>3+ words</td>
<td>[.64]</td>
<td>29%</td>
<td>7</td>
<td>--</td>
<td>0%</td>
<td>0</td>
</tr>
<tr>
<td><strong>range</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* = these tokens were excluded for the multivariate analysis of this factor group only due to low Ns

---

15 Examples of a 1-word animate possessors include things like “Leopold’s connection” (1-word singular human, 1906 *Hansard*); “the person’s reasons” (1-word singular “person”, 2006 *Hansard*); “people’s business” (1-word plural, 2006 *Maclean’s*); “man’s long battle” (1-word collective, 1956 *Maclean’s*).

16 The few tokens (N = 4) of 1-word singular humans realized with the of-genitive in 2006 *Maclean’s* are those where the possessor is longer than the possessum, sometimes by quite a bit: “the son of an Algerian wife-beater”, “the likelihood of any girl born today being able”, “the voice of the federal politician responsible for the CBC”. This is evidence that the principle of end weight can override the animacy constraint on occasion.
Possessum length is never statistically significant for animate possessors. This is contrary to expectations in a written corpus according to Szmrecsanyi & Hinrichs (2008:304):

…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, [...] 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.

Given the above statement, we might expect possessum end weight to be active in written registers, where on-line processing constraints are not active, but not active for spoken registers — and indeed, Tagliamonte & Jankowski (forthcoming) find that it was never significant in 20th century vernacular Toronto English. It is also never significant in Hansard, which could be an indicator of its ‘written to be spoken’ status. Maclean’s, despite being a clearly “written” register, exhibits the opposite tendency to that expected from a written register — a hint that it is indeed a more “colloquial” written register (cf. Hundt & Mair, 1999; Leech, Hundt, Mair & Smith, 2009: 239), and possibly more speech-like than the press registers used in previous studies (from the BROWN family of corpora; Hinrichs & Szmrecsanyi, 2007:442; Hundt & Mair, 1999:265–66). 17

The phonological factor of final sibilant in the possessor is expected to be more likely to play a role in a more oral register like Hansard than in a purely written register like Maclean’s, as it does here. This factor is never significant in Maclean’s. 18 It was originally not significant in Hansard, but not only becomes so by 1956, it is on par with the strongest effect (a range of 31 vs. 33 for possessor type in 1956). By 2006, it has become the strongest effect in Hansard (a range of 63 vs. 42 for possessor type).

17 This tendency is visible only in the reversed (and non-significant) constraint hierarchy for the factor of final sibilant in 1956 Maclean’s, making this observation cautious at best.

18 The shift of constraint hierarchy for final sibilant in 1956 in Table 4-1 is puzzling, even though the factor is not significant at any time for animates in Maclean’s. Closer examination of the individual tokens here could provide a partial answer. Because the 1-word singular humans had to be excluded from the multivariate analysis due to categoricity, the plural possessors exert a stronger effect on the remainder of the data than they would normally. These include things like, “the strikers’ purposes, the listeners’ week, their pupils’ parents” etc., with the plural-s morpheme, which could be overriding the constraint of avoiding the s-genitive when the possessor ends in a final-sibilant.
The factor group of possession type starts out as not significant in both corpora, and becomes significant by in 1956 Maclean’s and 2006 Hansard, with prototypical possessors strongly disfavouring the s-genitive. As discussed above in the distributional analysis for possession type (section 4.2.2, see also Figure 4-12 and Figure 4-13), Wolk et al. (to appear:14) suggest that prototypical possessors should be expected to favour s-genitive. I suggest that what we see in the above data is the strong combined effect of register variation and the animacy/length constraint. Prototypical possessors are sparse in the data, and more often inanimate rather than animate. Those that are animate are more likely to be multi-word plurals and collectives (i.e. “the parents of her pupils”, from 1956 Maclean’s; “the hands of the people”, from 1956 Hansard) rather than 1-word singular humans (the main favouring factor for the s-genitive).

Overall, we can say that while the constraint hierarchies of significant factors are consistent across time, the grammatical constraints are generally strengthening over time with regard to the use of s-genitive with animate possessors, and there is shifting of the magnitude of the significant effects in Hansard. These are indications of grammatical change as the s–genitive takes on new functions in the grammar over time in both corpora. In particular, there is some evidence of grammaticalization (recall the discussion in section 2.1, pp. 22-23): we see layering, with new functions coming into play for the form (for example, the significance of possession type in 1956 and 2006 Maclean’s as well as in 2006 Hansard), while older functions, visible in the constraint hierarchies for factors such as animacy (combined with possessor length), show persistence and remain intact (Hopper & Traugott, 1993:22–31).

4.4 Multivariate Analysis of grammatical factors: inanimate possessors

Moving on to the inanimate possessors, Table 4-3 (Maclean’s) and Table 4-4 (Hansard) show the results of parallel multivariate analyses of the same internal factors as those run for animate possessors in Table 4-1 and Table 4-2. Again, the animacy and possessor length factor group has the strongest effect in both registers, as it always has the highest range within each analysis (63/50/69 in Maclean’s and 64 in 2006 Hansard). The constraint hierarchy for this factor group is again relatively stable across time in Maclean’s. Looking more closely at the factor weights in Hansard, though, there is evidence of grammatical change over time that is not progressing the
same way in the two corpora—a shift in the constraint hierarchy of one-word place possessors, which go from never being found with the s-genitive in 1906 *Hansard* to being highly favoured in this context in 2006 (factor weight of .80, N = 80).

Table 4-3:

<table>
<thead>
<tr>
<th>Contribution of grammatical factors to the probability of s-genitive with inanimate possessors: <em>Maclean’s</em></th>
<th>1906</th>
<th>1956</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>POSSESSOR TYPE</strong></td>
<td><strong>FW</strong></td>
<td><strong>%</strong></td>
<td><strong>N</strong></td>
</tr>
<tr>
<td>1-wd Places &amp; Organizations</td>
<td>.87</td>
<td>46%</td>
<td>55</td>
</tr>
<tr>
<td>2+wd Places &amp; Organizations</td>
<td>.42</td>
<td>6%</td>
<td>186</td>
</tr>
<tr>
<td>Concrete Inanimates</td>
<td>.24</td>
<td>3%</td>
<td>39</td>
</tr>
<tr>
<td><strong>FINAL SIBILANT IN POSSESSOR</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not sibilant</td>
<td>[.56]</td>
<td>13%</td>
<td>226</td>
</tr>
<tr>
<td>Sibilant</td>
<td>[.26]</td>
<td>4%</td>
<td>51</td>
</tr>
<tr>
<td><strong>POSSSESSION TYPE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-prototypical</td>
<td>[.48]</td>
<td>12%</td>
<td>226</td>
</tr>
<tr>
<td>Prototypical</td>
<td>[.58]</td>
<td>9%</td>
<td>54</td>
</tr>
<tr>
<td><strong>POSSESSUM LENGTH</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 word</td>
<td>[.52]</td>
<td>10%</td>
<td>166</td>
</tr>
<tr>
<td>2 words</td>
<td>[.56]</td>
<td>16%</td>
<td>77</td>
</tr>
<tr>
<td>3+ words</td>
<td>[.30]</td>
<td>8%</td>
<td>37</td>
</tr>
</tbody>
</table>

---

19 Examples of 1-word place possessors include the proper noun *Canada*, as in “the nature of Canada” and “Canada’s tax system” (both from 2006 *Maclean’s*). It is worth noting in particular that the proper nouns *Canada* and *Ontario* as a possessor head are found with a higher frequency in *Hansard*, due to the topics of discourse. They are only ever found with the of-genitive until 2006, when we start to see examples like “Canada’s new government” in addition to “the government of Canada.”
Table 4-4: Contribution of grammatical factors to the probability of $s$-genitive with inanimate possessors: *Hansard*

<table>
<thead>
<tr>
<th>POSSESSOR TYPE</th>
<th>1906</th>
<th>1956</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Input: 0.03</td>
<td>Input: 0.04</td>
<td>Input: 0.45</td>
</tr>
<tr>
<td></td>
<td>N: 222</td>
<td>N: 284</td>
<td>N: 209</td>
</tr>
<tr>
<td>POSSESSOR TYPE</td>
<td>FW</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>1-word Places &amp; Organizations</td>
<td>--</td>
<td>0%</td>
<td>25</td>
</tr>
<tr>
<td>2+word Places &amp; Organizations</td>
<td>[.43]</td>
<td>3%</td>
<td>147</td>
</tr>
<tr>
<td>Concrete Inanimates</td>
<td>[.71]</td>
<td>6%</td>
<td>50</td>
</tr>
<tr>
<td>FINAL SIBILANT IN POSSESSOR</td>
<td>range</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not sibilant</td>
<td>[.46]</td>
<td>3%</td>
<td>184</td>
</tr>
<tr>
<td>Sibilant</td>
<td>[.71]</td>
<td>5%</td>
<td>37</td>
</tr>
<tr>
<td>POSESSUM LENGTH</td>
<td>range</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 word</td>
<td>[.52]</td>
<td>4%</td>
<td>170</td>
</tr>
<tr>
<td>2+ words</td>
<td>[.44]</td>
<td>2%</td>
<td>52</td>
</tr>
</tbody>
</table>

The effect of the factor group of final sibilant also changes over time in both corpora. In *Maclean’s*, this factor reaches significance (with the effect in the expected direction, such that $s$-genitives are disfavoured when a final sibilant is present) in 1956 and remains so in 2006. In *Hansard*, the factor does not reach significance until 2006.21 Possessum length is also significant in *Maclean’s* here, though not until 1956, where the predicted direction of the effect is found (longer possessums favour the $s$-genitive). The magnitude of this effect changes such that it is the second-strongest effect in 2006 (range of 42, vs. 30 for final sibilant). Finally, possession type

---

20 Recall that for inanimate possessors in *Hansard*, the N = 62 tokens of non-prototypical possession were realized 99 percent with the of-genitive (Figure 13, see also example (8e)). This factor group is therefore not run in the multivariate analysis of inanimate possessors in *Hansard* shown in Table 4-4.

21 It is interesting that the OCP appears to be operative in the later time periods of both corpora, as shown by the factor group of final sibilant. I take this to be further evidence of potential ongoing colloquialization in these registers such that they are becoming more ‘speech-like’ (cf. Hundt & Mair, 1999; Leech, Hundt, Mair & Smith, 2009: 239).
reaches significance in 1956 *Maclean’s*, only to no longer be significant in 2006. The constraint hierarchy, though, is preserved.\textsuperscript{22}

Overall, again, there is a general trend for more internal linguistic factors to become significant over time, indicating ongoing grammatical change of the *s*-genitive into the contexts of 1-word place and organization possessors. However, unlike the animate possessors in Table 4-1 and Table 4-2, *Hansard* is lagging behind *Maclean’s* for this change, with no internal grammatical factors reaching significance until 2006. With the animate possessors, the changes are occurring in tandem across time in the two registers.

To sum up so far, according to the methodology laid out in section 2.4.1, there is evidence of changes to the conditioning of the internal factors across time within each register. There is a tendency in both registers for the *s*-genitive to take on more, and stronger, grammatical conditioning over time, with additional factors becoming significant in each time period, and there is a shift in the magnitude of the significant factor groups — final sibilant shifts to become the most significant factor group for animates in *Hansard*, while possessum length surpasses final sibilant to become the second-strongest effect for inanimates in *Maclean’s*. This is all evidence of grammatical change, according to the three lines of evidence laid out in Chapter 2. Register change is also indicated for *Maclean’s*: Figure 4-7 showed an overall increase in 1-word organization and place-possessors in both corpora, with a steeper increase in *Maclean’s*. Finally, the input probabilities and all proportions over the three time periods in Tables 4-1–4-4 show an overall increase in the *s*-genitive in both registers.

### 4.5 Type-token ratio (lexical density) and Information status (New vs. Given)

I now turn to the analysis of lexical density (type-token ratio) and information status, which have been shown to tap into the structure of the register and provide diagnostics for register change. In addition, the genitive is potentially susceptible to the effects of economy, or the need to encode

\textsuperscript{22} As with the animate possessors, when the effect of possession type is significant, it is the opposite to that predicted by Wolk et al. (to appear: 14). Again, I point to the interaction of animacy, length and possession type as the mitigating factor here.
as much information into as little space (or time) as possible, since the \textit{s}-genitive is appreciably more compact than the \textit{of}-genitive. Recall that this effect is believed to be particularly active in journalistic prose in the later 20\textsuperscript{th} century (Biber, 2003:170–1, 180–1). Lexical density is an appropriate diagnostic to test for this effect as well. A higher lexical density will encourage the use of more compact forms, where the option exists.

Lexical density is calculated for this study in the same manner as it was for the relative pronoun study in Chapter 3 — by counting how many different parts of speech (to a maximum of 16) were present within 50 words on either side of an instance of an interchangeable genitive.\footnote{See 2.4.2 and Appendix B for details on this methodology, with an example text.} In general, the higher the type-token ratio, the more likely the need should be to use the more “economical” \textit{s}-genitive. If the type-token ratio measurement is changing over time (becoming higher or lower), this indicates a change in the information structure of the register: a register change. In this data, the number of unique parts of speech fell within a range from 8–15.\footnote{As discussed in Chapter 3, the final grouping of this factor group into the three categories (here, 8–10, 11, and 12–15) was an arbitrary artifact of how the factor group was coded, since a considerable portion of the data clustered at the mid-point value of 11, and the categories above and below that had considerably fewer data points each than the data with the value of 11. The reader is referred to fn 19, pg. 61 for further discussion of this and speculation for future work, and to fn 27, pg. 39, section 2.4.2 for discussion of how the factor group was coded.}
result was that within a window of 100 words, 50 words on either side of the target token, a number was assigned ranging from 8–15 different parts of speech. The results break down as shown in Figure 4-16 (Maclean’s) and Figure 4-17 (Hansard). These results show that there is generally higher lexical density around s-genitive tokens in the more recent time periods, particularly for Maclean’s (Figure 4-16), indicating that not only could economy be in play in this register during those years, particularly in 2006 where the gap between lower and higher lexical density is quite pronounced (as opposed to 1906, where all three values are very close to one another). The effect of lexical density appears to be changing over time. In other words, not only does this seem to be an active force in Maclean’s (the trend is less so in Hansard, Figure 4-17), the strength of the effect seems to have increased over time. This will be further tested in the multivariate analysis, with animate and inanimate possessors treated separately.

The second factor that can test for changes in the information structure of the register is information status, or the givenness, or not, of the head noun of the possessor NP within the previous 100 words of discourse. As described in 2.4.2, new material introduced into the discourse is more likely to appear as a noun than a pronoun (Fox & Thompson, 1990:304), and more nouns indicate more informational load. If the register uses more nominal constructions

Figure 4-18:

Information Status of s-genitive possessor head noun (Maclean’s) in previous 100 words

![Graph showing the percentage of s-genitive use for Maclean’s in previous 100 words over the years 1906, 1956, and 2006.]

Figure 4-19:

Information Status of s-genitive possessor head noun (Hansard) in previous 100 words

![Graph showing the percentage of s-genitive use for Hansard in previous 100 words over the years 1906, 1956, and 2006.]

---

25 Again, see 2.4.2 and Appendix A for more details and an example of how this was counted.
that are New information as opposed to Given (or vice-versa) over time, this is evidence of register change. Results for this factor are plotted over time for the s-genitive in Figure 4-18 (Maclean’s) and Figure 4-19 (Hansard). In both registers, the lines for New and Given information are parallel, with little difference between the two at any point in time, meaning that this factor is less likely to exert an influence than lexical density. This observation will also need to be tested in the multivariate analysis, with animate and inanimate possessors treated separately, in order to provide statistical verification.

4.6 Multivariate analysis of the register-internal factors

As with the internal linguistic factors, separate multivariate analyses of the register-internal factors have been run for animate and inanimate possessors.\(^\text{26}\) For animate possessors in Maclean’s (Table 4-5), the register-internal situation appears to be stable. Information status is significant in all time periods with the same constraint hierarchy, while lexical density is not

<table>
<thead>
<tr>
<th>Table 4-5: Contribution of register factors to the probability of s-genitive with animate possessors: Maclean’s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>LEXICAL DENSITY</td>
</tr>
<tr>
<td>7–10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12–15</td>
</tr>
<tr>
<td>INFORMATION STATUS</td>
</tr>
<tr>
<td>Given</td>
</tr>
<tr>
<td>New</td>
</tr>
<tr>
<td>range</td>
</tr>
</tbody>
</table>

Not significant: Lexical Density

\(^{26}\) For all multivariate analyses, internal linguistic factors (such as in Tables 4-1–4-4) and register-internal factors (such as in Tables 4-5–4-8) were also run together and separately in the multivariate analysis, with no difference in log likelihood, and so can be considered part of a single analysis.
significant. The situation for animate possessors in *Hansard* (Table 4-6) is not as stable. Information status loses significance over time, while lexical density becomes significant in 2006. I interpret this to mean that although the information status does play a role in the variation for animates in both registers — something we could already have predicted simply from the differences in distributions between the two registers — its role is not changing in *Maclean’s*, and so is not playing a role in the increase of *s*-genitive there (see Figure 4-18). Contrast *Hansard*, where there is evidence of register change influencing the increase in *s*-genitive with animate possessors (see Figure 4-19). Information status was once significant and is no longer, while lexical density was formerly not significant and becomes so.\(^{27}\)

**Table 4-6:**

<table>
<thead>
<tr>
<th>Information Status</th>
<th>1906</th>
<th>1956</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> 0.20</td>
<td>0.20</td>
<td>0.15</td>
<td>0.36</td>
</tr>
<tr>
<td><strong>N:</strong> 208</td>
<td>156</td>
<td>167</td>
<td></td>
</tr>
<tr>
<td><strong>LEXICAL DENSITY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7–10</td>
<td>.47</td>
<td>.52</td>
<td>.50</td>
</tr>
<tr>
<td>11</td>
<td>.62</td>
<td>.56</td>
<td>.61</td>
</tr>
<tr>
<td>12–15</td>
<td>.41</td>
<td>.42</td>
<td>.33</td>
</tr>
<tr>
<td><strong>N:</strong> 56</td>
<td>59</td>
<td>52</td>
<td>37</td>
</tr>
<tr>
<td><strong>INFORMATION STATUS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Given</td>
<td>.65</td>
<td>.66</td>
<td>.53</td>
</tr>
<tr>
<td>New</td>
<td>.44</td>
<td>.46</td>
<td>.49</td>
</tr>
<tr>
<td><strong>N:</strong> 57</td>
<td>31</td>
<td>124</td>
<td>44</td>
</tr>
</tbody>
</table>

Hinrichs & Szmrecsanyi (2007:460–4) conclude that the factor of lexical density contributes significant explanatory power to the model of variation for their data, but not more than the grammatical factors. Similarly, note here that the ranges of the significant register-internal factor

---

\(^{27}\) While it may seem counter-intuitive that *Hansard* would be susceptible to changes in lexical density, recall that it is not truly a spoken register. *Hansard* is edited, sometimes to the extent that entire clauses are moved around in order to improve readability and clarity. The *Hansard* style guide states that it: “is not a verbatim transcript of the debates of the House. […] it is acceptable to make changes so that anyone reading *Hansard* will get the meaning of what was said. Those who edit *Hansard* have an obligation to make a sentence more readable since there is a difference between the spoken and the written word.” (Hansard, n.d.:48)
groups in Table 4-5 (lexical density has ranges between 27/22/24) and Table 4-6 (information status has a range of 21 in 1906, lexical density a range of 28 in 2006) are almost always higher than the ranges of the significant animacy and possessor length constraints within the same year\textsuperscript{28} in Table 4-1 for *Maclean's* (ranges generally between 38 and 89) and in Table 4-2 for *Hansard* (ranges between 31 and 63).\textsuperscript{29}

The results of parallel multivariate analyses for register-internal factors with inanimate possessors are given in Table 4-7 (*Maclean's*) and Table 4-8 (*Hansard*). Once again, the results across the two registers could not be more different, nor could they be more different between animate and inanimate possessors. For inanimate possessors in *Maclean's*, only lexical density is significant, and only in 2006 (with a range of 26) — compare this to Table 4-5 (*Maclean's* animate possessors) where lexical density is never significant and the effect of information status is stable over time. This indicates that there is register change happening, but it is only relevant with regard to the increased use of s-genitive with inanimate possessors.

Table 4-7:

<table>
<thead>
<tr>
<th></th>
<th>1906</th>
<th>1956</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Input: 0.11</td>
<td>Input: 0.38</td>
<td>Input: 0.57</td>
</tr>
<tr>
<td></td>
<td>N: 280</td>
<td>N: 248</td>
<td>N: 244</td>
</tr>
<tr>
<td>LEXICAL DENSITY</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7–10</td>
<td>[.58]</td>
<td>[.51]</td>
<td>.32</td>
</tr>
<tr>
<td></td>
<td>14%</td>
<td>39%</td>
<td>37%</td>
</tr>
<tr>
<td></td>
<td>102</td>
<td>51</td>
<td>44</td>
</tr>
<tr>
<td>11</td>
<td>[.43]</td>
<td>[.45]</td>
<td>.58</td>
</tr>
<tr>
<td></td>
<td>8%</td>
<td>34%</td>
<td>65%</td>
</tr>
<tr>
<td></td>
<td>97</td>
<td>89</td>
<td>80</td>
</tr>
<tr>
<td>12–15</td>
<td>[.48]</td>
<td>[.55]</td>
<td>.52</td>
</tr>
<tr>
<td></td>
<td>9%</td>
<td>42%</td>
<td>57%</td>
</tr>
<tr>
<td></td>
<td>71</td>
<td>99</td>
<td>116</td>
</tr>
<tr>
<td>INFORMATION STATUS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Given</td>
<td>[.47]</td>
<td>[.56]</td>
<td>[.61]</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>44%</td>
<td>67%</td>
</tr>
<tr>
<td></td>
<td>68</td>
<td>57</td>
<td>54</td>
</tr>
<tr>
<td>New</td>
<td>[.51]</td>
<td>[.48]</td>
<td>[.47]</td>
</tr>
<tr>
<td></td>
<td>12%</td>
<td>37%</td>
<td>54%</td>
</tr>
<tr>
<td></td>
<td>208</td>
<td>189</td>
<td>190</td>
</tr>
<tr>
<td>range</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not significant: Information Status</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Similarly, in Table 4-8, we see that only information status is significant, and again, only in 2006 (with a range of 19) — compare this result to Table 4-6 (Hansard animate possessors) where information status is not significant outside of 1906, and in 2006 only lexical density is significant. Again, I interpret this as register change, but again, no significant register-based factor in Tables 4-7 and 4-8 have a magnitude of effect higher than a significant grammatical factor within the corresponding analysis in Table 4-3 (Maclean’s inanimate possessors, with the many significant effects having ranges from 23–69) or Table 4-4 (Hansard inanimate possessors, with significant effects only in 2006, ranges from 54–64). When these factors are run together in the same analysis, the internal grammatical factors always exert a stronger influence on the variant choice than the register-internal factors.  

Table 4-8:

<table>
<thead>
<tr>
<th></th>
<th>Contribution of register factors to the probability of s-genitive with inanimate possessors: Hansard</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEXICAL DENSITY</td>
<td>1906</td>
</tr>
<tr>
<td></td>
<td>Input: 0.03</td>
</tr>
<tr>
<td></td>
<td>N: 222</td>
</tr>
<tr>
<td>FW % N</td>
<td>FW % N</td>
</tr>
<tr>
<td>7–10</td>
<td>[.40] 3% 58</td>
</tr>
<tr>
<td>11</td>
<td>[.57] 7% 77</td>
</tr>
<tr>
<td>12–15</td>
<td>-- 0% 80</td>
</tr>
<tr>
<td>INFORMATION STATUS</td>
<td>Given</td>
</tr>
<tr>
<td>range</td>
<td>[.33] 2% 61</td>
</tr>
<tr>
<td>range</td>
<td>[.57] 4% 160</td>
</tr>
<tr>
<td>Not significant: Lexical Density</td>
<td></td>
</tr>
</tbody>
</table>

The consistency of the significant constraints in Maclean’s animates (Table 4-5) and shifting significant effects for inanimates (Table 4-7), as well as the shifts in the significant effects in Hansard (Table 4-6 and Table 4-8) are further evidence that there is register change happening in both registers. The effect of this change in Maclean’s is limited to inanimate constructions,

---

30 Cross-tabulations of the register-internal factor groups against the grammatical factors did not reveal any obvious interactions.
and is probably due to the pressures of economy, as shown via the diagnostic of lexical density. In *Hansard*, this effect is present for both animate and inanimate possessors.31

The additional evidence for register change from the results of the register-internal factors in this section bolsters the evidence of change discussed in sections 4.3 – 4.5, where evidence of both register change and grammatical change for the *s*-genitive was found. Evidence for register change there came from the overall increase in 1-word organization and place possessors in both corpora in Figure 4-7, with a steeper increase in *Maclean’s*. Grammatical change was visible in Tables 4-1 – 4-4, with increasing input probabilities and proportions for the *s*-genitive across the three time periods in both registers, and the tendency in both registers for the *s*-genitive to take on more, and stronger, grammatical conditioning, with additional factors becoming significant over time. Further evidence of grammatical change came from the shifting magnitude of effects (particularly in *Hansard*) of internal factors across time within each single register.

Recall from Chapter 2 that *Hansard* was described as a spoken register that also shares some qualities of a written register, falling at the midpoint of the *Involved vs. Informational* scale, *Maclean’s*, on the other hand, is highly *informational*. In order to fully interpret these results, we must obtain a further baseline for these factors in a register that falls at the *involved* end of the scale, such as vernacular speech, at which point we can see how similar or different *Hansard* really is with regards to this variable.

### 4.7 Genitive variation in vernacular speech

Figure 4-20 shows the proportion of *s*-genitive use with human (singular, plural and collective) antecedents in the *Maclean’s* and *Hansard* real-time data from 1906, 1956 and 2006. Added to this are the results from Tagliamonte & Jankowski (forthcoming), for vernacular speech data of Ontario-born speakers, ages 30–59 and over 60 when they were recorded between 2002–2004. The over-60 year-old Ontario speakers are therefore roughly comparable in apparent-time to the *Maclean’s* authors and *Hansard* speakers (all 30-60 years old in the sample year) from 1956,

31 Again, it bears repeating, that while it may seem that a “spoken” register would not be susceptible to the pressures of economy, *Hansard* is edited specifically to make the speaker transcripts “more readable”.
while the 30–59 year-old Toronto speakers are comparable to the *Maclean’s* and *Hansard* authors and speakers from 2006.  

**Figure 4-20:**

![Diagram showing the percentage of s-genitive use with human and collective possessors over time across different years and registers.](image)

First, we can see that the rates of *s*-genitive use with human possessors differ quite a bit across the three registers. Vernacular speech is now (only just) the most advanced in using the *s*-genitive with human antecedents, though *Maclean’s* was ahead in the mid-century. Within the written registers, *Maclean’s* is far more advanced than *Hansard* in this usage. Use of the *s*-genitive appears to be on the increase more rapidly in *Maclean’s* and vernacular speech than in *Hansard*, which is stable until the later part of the century.

---

32 Tagliamonte & Jankowski (forthcoming) uses data from both The Toronto Corpus (214 speakers) as well as the Directions of Change corpus, containing 68 speakers from the smaller city of Belleville and the towns of Lakefield and Burnt River. As was the case in Chapter 3 for the vernacular speech study of relative clauses (D’Arcy & Tagliamonte, 2010), not all speakers from the corpora were used in the vernacular speech genitive study (Tagliamonte & Jankowski, forthcoming). Additionally, the vernacular speech data includes female speakers, while the *Maclean’s* and *Hansard* data do not (see section 2.2.2). The reader is referred to that work for specifics on the sample size and speakers used. As was noted in Ch. 3 (fn 5, pg. 52), the inclusion of women could be expected to result in higher levels of a standard or prestige variant (Labov 2001: 266), such as the *of*-genitive. Despite this, there is still a noticeable increase in the *s*-genitive visible in the vernacular data. Further, women are often implicated as the leaders of linguistic change, especially in change from above (2001: 274). However, discussion of the gender dimension is beyond the scope of this study — the reader is referred to Tagliamonte & Jankowski (forthcoming).

33 But recall from the evidence provided in the multivariate analysis of animate possessors (Table 4-1 and Table 4-2) that the changing and layering of the effects of significant factors was proceeding in tandem across time in the two registers.
Tagliamonte & Jankowski (forthcoming) originally grouped their 200 tokens of collective and organization possessors with inanimates rather than animates. Figure 4-20 (above) and 4-21 (below) have been re-calculated from their data. When only human possessors are plotted in the vernacular speech data (without collectives or organizations), the trajectory over time is stable. Figures 4-20 and 4-21 show the stylistic differences between the two spoken registers, and reinforce that the s-genitive has long been favoured with human possessors, while at the same time showing how journalistic prose is aligned more closely with the vernacular than it is with formal spoken language. While this is partially consistent with the hypothesis that changes begin in the vernacular and spread to writing (Pintzuk, 2003:525), it does not support the hypothesis that Hansard, as a quasi-spoken register, would pick up the change first and from there disseminate it to other written registers.

**Figure 4-21:**

![Graph](image)

Figure 4-22 is the data for inanimate possessors, charted in the same manner as Figure 4-20 and 4-21. Recall that the spread of s-genitive into inanimate possessors is claimed to be a later stage in the ongoing changes in this variable (Rosenbach, 2005:615, 2003:386). Here we see a different picture than for the animate possessors in Figure 4-20: in fact, it seems that we have caught an early point in the changing animacy constraint. Use of the s-genitive with inanimates appears in Maclean’s and Hansard at the same time, but the change progresses more quickly in Maclean’s. This is consistent with the observation that press registers are a potential locus for change in written registers (Hundt & Mair, 1999; Westin 2002).
When all inanimates are grouped together in vernacular speech, there is no evidence of an increase of *s*-genitive there. Rather the trajectory of the vernacular in Figure 4-22 is stable, remaining under 10%.

**Figure 4-23: Proportion of *s*-genitive by animacy in Ontario English (adapted from Tagliamonte & Jankowski, forthcoming)**

However, Figure 4-23 shows the distribution over time of the *s*-genitive for each of the separate inanimate categories from Tagliamonte & Jankowski (forthcoming). Here we can see that the *s*-genitive is, in fact, increasing over time in vernacular speech, but only with the collectives and
organizations (these two categories were grouped together in this analysis, reported in Figure 4-23 as “collectives”, the dashed line with the open triangle markers). For this group of inanimates, the difference between the proportion of 6 percent s-genitive use for the oldest speakers (a confidence interval calculated at $p < 0.05$ gives an upper limit of 11 percent) is significantly different than the 29 percent s-genitive use for the youngest speakers (confidence interval at $p < 0.05$ gives a lower limit at 15 percent).\(^{34}\)

If the Ontario Vernacular s-genitive inanimates in Figure 4-22 are replaced by only the Ontario Vernacular s-genitive collectives and organizations from Figure 4-23, we get Figure 4-24:

**Figure 4-24: s-genitive with inanimate possessors over time + Ontario Vernacular collectives (and organizations) (adapted from Figure 4-22 and 4-23)**

![Diagram showing s-genitive usage with inanimate possessors over time](image)

This, then, is the confirmation that *Maclean’s* and *Hansard* are increasing their use of s-genitive with inanimates earlier than vernacular speech. Use of the s-genitive with inanimates in vernacular speech is rare overall (Figure 4-23), but it does seem to have begun more recently than in *Maclean’s*, about the same time as *Hansard*, and is still lagging behind both. This would

---

\(^{34}\) Since the lower limit of the confidence interval for the younger speakers (15 percent) and the upper limit of the confidence interval for the older speakers (11 percent) do not overlap, the two values can be interpreted as being significantly different at the $p < 0.05$ level.
therefore appear to be evidence against the hypothesis that change always spreads from the vernacular to writing (Pintzuk, 2003:525).

Rather, as in the case of the WH-forms of the relative pronouns described by Romaine (1982) and Tottie (1997a), this instead appears to be a “change from above” (Labov, 1994:78) that travels from writing into speech, starting in the press register and moving first into formal speech such as the parliamentary speech of Hansard before the vernacular. This is particularly interesting, as the overall trajectory of genitives in English from the 10th–16th centuries was toward the use of the Norman of-genitive (Rosenbach, 2002:186). The Saxon s-genitive was not historically a prestige form, making it an unlikely candidate for change from above, yet the variation it exhibits across the register dimension, with change toward increased use of s-genitive with inanimates, indicates that it is operating above the conscious level. The higher self-monitoring (or editorial scrutiny) of language in registers such as Maclean’s, implies a level of conscious awareness when there is variation and change. However, the s-genitive also acts as a vernacular form in Canadian English today (Tagliamonte & Jankowski, forthcoming). We must conclude that factors other than prestige, such as register-internal change, can play a strong motivating role in change from above.

We now have an idea as to the direction of the change between the written registers and vernacular speech. As further evidence for the analysis of Maclean’s leading this change for inanimate possessors, I return for a moment to Table 4-3 and Table 4-4 above. The main change in the internal factors was shown to be the change in the animacy constraint toward use of one-word place and organization inanimates with the s-genitive (i.e. Toronto’s waterfront vs. the history/pride/problems of Toronto). This is happening in both Maclean’s and Hansard. The effect of animacy is always the strongest internal constraint in Maclean’s, while in Hansard, it exerts the strongest effect in the only time period where grammatical factors are significant at all (2006). In Hansard, one-word places (i.e. Canada’s immigration system) and organization possessors (i.e. L.A.M.P.'s homelessness programs) are never found with s-genitive in 1906, but are a highly favouring context by 2006 with a factor weight of .80).

Tagliamonte & Jankowski (forthcoming) also test for the register-internal factor of information status, and find it to never be significant in any analysis of the vernacular speech data, for either animates or inanimates. This provides the final bit of evidence that register change is not a factor
in the vernacular speech data. I would suggest, then, that the *register* change is the necessary component in *Maclean’s* and *Hansard* for driving forward the grammatical change in the inanimate possessors. Without this component, vernacular speech is remarkably stable for this variable. The only increase in *s*-genitive to be found is in collectives. These are becoming more like other human possessors, though this change lags behind the written data. This makes it one of the “relatively infrequent” cases of a change from above that is moving from writing into speech (Tottie, 1997a:84), caught here in its incipient stage.
Chapter 5
Recap, Implications, and Conclusions

5 Recap of major findings

This work has presented two morphosyntactic variables that are currently undergoing change in speech and writing, with the goals of further elucidating the locus and spread of such changes, differentiating grammatical from register changes, and tracking the paths along which changes spread from one register to another. The methodological innovation (following from work by Hinrichs & Szmrecsanyi, 2007; Szmrecsanyi & Hinrichs, 2008 and further advocated by Szmrecsanyi, 2013) of using register-internal constraints which rely on measures of frequency provides evidence toward the study of the constraints problem (Weinreich et al., 1968:183–4, see discussion in Ch. 1, p. 6), viz. “the goal […] to determine the set of possible changes and possible conditions for change.” Such constraints are typically left to the purview of corpus linguists. However, when used alongside the proportional and multivariate analysis of internal grammatical factors in a variationist framework, such factors can provide important evidence toward the actuation problem. That is, they can provide further insight into the nature of the “stimuli and constraints both from society and from the structure of language” that trigger the beginning of a change (Weinreich et al., 1968:186–7).

These register-internal constraints tap into potentially changing stylistic notions, including editorial conventions and what Görlach calls the “communicational intentions” (2002: 18). Factors such as the information status of a register (measured by looking at the givenness of head nouns in construction like relative clauses and genitives) and economy (measured by the type-token ratio, or lexical density) provide further evidence for why a change might take place in a language at a particular time. If there are changes within a register — such as a shift toward higher lexical density, or a change in the proportion for new vs. given information status, or changing pressures of economy, this causes the linguistic makeup of a register to change, and so then certain variants become more preferred.

There are several parallel findings between the two empirical studies in Chapter 3 (relative pronoun variation) and Chapter 4 (genitive variation). I will summarize these in turn, and then discuss what larger, overall conclusions can be drawn from those observations.
5.1 Lexical replacement vs. grammatical change

In Chapter 3, the press register Maclean’s was found to be leading the change in the loss of which as a restrictive subject function relative pronoun. Restrictive which in standard Canadian English has been completely replaced by that for this function in Maclean’s, due to a changing stylistic convention. This stylistic convention appears to have been brought on by the prescription against restrictive which in press writing, which began to be applied in earnest by the late-20th century.

One of the effects of this change was visible in the shifting significance and constraint hierarchy of the register-internal factor of information status (Table 3-10). Relative clause antecedents that are previously given in the discourse were once predisposed to which as their relative pronoun. This effect had shifted by 1956 such that restrictive which became favoured for antecedents that were new information in the discourse. By 2006, restrictive which had disappeared from the register completely. No such evidence of ongoing register change was found in Hansard, where restrictive which still exists, albeit minimally (Table 3-11). In addition, no evidence of grammatical change was found for restrictive which in either Maclean’s or Hansard (Tables 3-8 and 3-9). Rather, the grammatical system of relative which — in fact of the relative pronoun system in general — was stable well before the loss of which. This is supported by previous research on the relative pronouns (e.g. D’Arcy & Tagliamonte, 2010). The loss of which was therefore argued to be an instance of lexical replacement rather than ongoing grammatical change.

In Chapter 4, Maclean’s was once again found to be leading change. In particular, evidence was found for both ongoing grammatical change and register change. Grammatical change was visible for both animate and inanimate possessors, with the s-genitive increasing in both domains. For animate possessors, this was visible with the changing significance and strength of grammatical constraints over time (Table 4-1 and 4-2), indicating that the s–genitive was taking on new functions in the grammar in both corpora. These included the developing tendency to be conditioned by the type of possession, and in Hansard, a developing tendency to be conditioned by the phonological factor of final sibilant in the possessor, which also caused a shift in the magnitude of the significant effects, with final sibilant taking the place of the combined animacy/length constraint as the strongest effect there by 2006. For inanimate possessors (see
Table 4-3 and 4-4), the effect of animacy was stable over time in *Maclean’s*, but showed a dramatic shift in *Hansard*. There, animacy was not a conditioning factor until 2006, when the frequency of *s*-genitive with inanmites rises, and the animacy constraint then matches that of inanimate possessors in *Maclean’s*. There is increasing significance of other grammatical factors as well, including for final sibilant in the possessor in both registers, and for possessum length in *Maclean’s*. Overall, there was a general trend for more internal linguistic factors to become significant over time, indicating ongoing grammaticalization of the *s*-genitive into the contexts of 1-word place and organization possessors, with *Hansard* lagging behind *Maclean’s* across time. Recall that evidence of grammaticalization (Hopper & Traugott, 1993:22–31) was previously discussed in section 2.1 (pp. 22-23) as being when a form takes on new functions in the grammar, as the *s*-genitive is shown to do here. As mentioned in section 4.3 (pg. 89), there is evidence of layering, with new factors becoming significant over time in both corpora, hinting at new functions for the form in the grammar (i.e., the significance of possession type in 1956 and 2006 *Maclean’s* as well as in 2006 *Hansard*). There is also persistence, in that the constraint hierarchies for factors such as animacy remain intact even as other factors come into play and the relative strength of significant factors shifts (as in *Hansard*).

Evidence for register change in the genitives was seen for animate possessors in *Hansard* (Table 4-6) and inanimate possessors in *Maclean’s* (Table 4-7), where the effect of lexical density became significant in 2006, and for inanimate possessors in *Hansard* (Table 4-8), where the effect of information status gained significance in 2006. It was noted, though, that in all cases, the strength of the significant grammatical factors outweighs the effect of the significant register factors. Additional evidence for register change came from the overall increase in 1-word organization and place possessors in both corpora (Figure 4-7), with a steeper increase in *Maclean’s*.

### 5.2 Some implications of the major findings

In Chapter 2, *Hansard* was described as a spoken register that also shares some qualities of a written register, falling at the midpoint of the *Involved vs. Informational* scale. *Maclean’s*, on the other hand, is highly *informational* (Biber, 1995:142, 146). Press reportage has been implicated previously as the locus for change in the *s*- and *of*- genitive by Hinrichs & Szmrecsanyi
From this, it was hypothesized that press reportage could be facilitating changes in other written registers (see, e.g. Hundt & Mair, 1999; Westin, 2002), such that forms entering the written registers from speech will first appear in press reportage before other written registers. In this case, the overall “chronology of change” of new forms would be that they first enter speech (e.g. Pintzuk, 2003), then spread to the press registers, like *Maclean’s*, and finally to other written registers, such as the written-to-be-spoken *Hansard*.

Indeed, the results from both empirical studies demonstrate that *Hansard*, though it is a partially spoken register, is more conservative than *Maclean’s*. As we saw in Chapter 3, it was in *Maclean’s* that restrictive *which* was completely replaced by *that* in 2006, although it was still present — barely — in *Hansard* and vernacular speech. In Chapter 4, the change in use of the *s*-genitive found with inanimates progresses more quickly in *Maclean’s*, even though it appears in *Maclean’s* and *Hansard* at the same time. Further, the comparison of the results from written registers to vernacular speech showed that *Maclean’s* and *Hansard* are ahead of the vernacular speech in adopting the *s*-genitive with inanimate possessors.

What this means is that, far from being “innovative” only when compared to other written registers, the press register can, in fact, be more innovative than speech. While the vernacular is extremely regular, what Labov calls “the most systematic data for our analysis of linguistic structure” (1972a:208), this does not mean that it will always be the most innovative, or the locus of change. The significance of the factor of lexical density for the genitives (Tables 4-6, 4-7 and 4-8) points to *economy* being a motivating factor for change in registers where there is a need to encode as much information into as little space as possible, such as press reportage. Hinrichs & Szmrecsanyi (2007:441, citing Biber, 2003 and Hundt & Mair, 1999) argue that economy should be taken into account whenever possible in studies that use press reportage (such as newspaper prose) as a genre.¹

When we are fortunate enough to capture the starting point of a grammatical change, as we did here for the inanimate genitives in Chapter 4, the cross-register perspective becomes a useful

---

¹ Hundt & Mair (1999:236) recommend caution, though, that “one should be careful not to generalize the findings from a newspaper corpus to other genres, considering this special place of press language in the spectrum of genres.” Also, as will be further discussed below, *economy* is not a potential diagnostic for every linguistic variable, only for those where one variant is appreciably “more economical” than another.
tool, beyond simply giving us the ability to determine what register the change originates in. By extension, a cross-register approach can also provide evidence as to whether a change is taking place above or below the level of consciousness for the speakers. In Chapter 1, it was noted that grammatical changes found to be happening in written registers that are the likely result of prescriptivism or literacy (i.e. the loss of which), represent a changes from above. This is because variables that are associated with particular styles and registers are necessarily operating above the level of social awareness (Bell, 1984).

In Chapter 3, we saw a change from above that involved the loss of a form, namely the lexical replacement (by that) of restrictive which, an historically prestigious form, for subject function in restrictive relative clauses. In Chapter 4, it was noted that the Saxon s-genitive was never an historically prestigious form. It has varied across the register dimension since it first competed with the of-genitive (Rosenbach, 2002:186). Over the 20th century, however, it has steadily undergone a process of grammatical change to rise again with inanimate possessors. It seems, then, that register-internal change (with register-internal factors as diagnostics) can join the ranks of factors like prestige in motivating changes from above.

5.3 Conclusions

Recall that Mair (2002:186) calls for “a model of how changing stylistic conventions and changing discourse traditions ultimately lead to changes in the underlying system of grammatical choices.” My intention with this work was to provide a model of language change that integrates change across both speech and writing. The overall goal has therefore been to further our understanding of the relationship between speech and written language.

From a purely methodological perspective, the empirical studies of Chapter 3 and Chapter 4 were intended to facilitate familiarization with and an example of frequency-based constraints, and to demonstrate a methodology for treating them as categorical (rather than continuous) independent variables for analysis using logistic regression. This methodology enables us to directly test for the effects of factors such as lexical density on variation, which, as discussed in section 2.4 and above, provides additional evidence to consider when written registers are under investigation, and to further disambiguate register change from grammatical change. As the
overall goal of the dissertation has been to track language change across spoken and written registers, this was imperative. Certainly, this is not intended to be a solution to the age-old problem of not having old speech data, since we cannot keep register constant as we move back to the days before recording technology. However, it provides a necessary way to test for the effects of register change within a synchronic analysis from this point forward, when speech and written data are both available.

Although vernacular speech remains “the style from which every other style must be calibrated” (Tagliamonte, 2006:8), I take the increasing range of written registers now available to us in the modern age to be a sign that we should expect innovations from such registers to occasionally affect spoken language. For example, studies of language variation in the Instant Messaging register reveal a medium that is intertwined with vernacular speech in surprising ways:

Contrary to expectation, speech has a more innovative profile in comparison to IM. [...] At the same time, speech tends to be more vernacular, containing higher rates of going to than will. In contrast, IM language is consistently more varied, exhibiting a wider range of variants than speech and, in particular, containing a higher proportion of standard forms than speech. (Tagliamonte & Denis, 2008:23)

As with any work of this nature, in the end we are often left with more questions than answers. One set of questions that has come up time and time again when this work is presented, is “How do you decide what element of a variable’s frequency is relevant?” In other words, when operationalizing a new register-internal constraint, how do we determine what to count? Is lexical density always the relevant factor? The answer is dependent upon the variable under investigation. Lexical density is an appropriate factor for some variables, but not necessarily for others. In the case of the relative pronouns vs. the genitives, lexical density was relevant because both were nominal constructions, and density of nouns is one of the features relevant to registers that exhibit more informational production, such as the press. Lexical density was also relevant for genitives because it taps into the notion of economy, and in the case of the genitives, one variant is more economical than the other.

This is not always the case. Not every variable will have an option that can be considered more economical — the possibility modals, for example, where the variants can, may, might and could are arguably of equal compactness; or the subject relative pronouns, where although a Ø variant exists in some dialects and styles, it is not necessarily a variant implicated in ongoing change. Yet economy could be implicated for variation within the modal paradigm when, for
example, one of the variants under investigation is a periphrastic modal, such as have to or have got to in variation with the core modal must. Economy is also only relevant when tested with registers in which “economy” can be motivated as a viable factor for register change, as it has been for press reportage. The same cannot be said of vernacular speech.

Previous research is often a source for potential predictors: Information status (new vs. given information) has been tested for many variables, including for previous research on the genitives (Hinrichs & Szmrecsanyi, 2007; Tagliamonte & Jankowski, forthcoming). However, when used in conjunction with analysis and the probabilistic conditioning of internal linguistic constraints to test for grammatical change, frequency-based constraints such as those described here can provide further explanations in the larger picture of language change.

The methodology demonstrated here is not intended to be comprehensive and applicable to all possible linguistic variables. As with any factor group to be operationalized, the researcher must make an educated hypothesis with justification, based on the grammar and construction under investigation. Lexical density, in particular, is only relevant when a reasonable hypothesis can be devised that relates it to the variable under study. If economy is not relevant for a variable because there is no variant that is more economical than another, then perhaps other linguistic features that cluster with it at one end of a register dimension will provide predictors for study. I leave such investigations to future research, and many future researchers.
References


Leech, Geoffrey, & Smith, Nicholas. (2005). Recent grammatical change in written English: some preliminary findings of a comparison of American and British English. In Antoinette Renouf & Andrew Kehoe (Eds.), *The changing face of corpus linguistics: papers from the 24th International Conference on English Language Research on Computerized Corpora (ICAME 24).* Amsterdam: Rodopi. 185–204.


http://languagelog.ldc.upenn.edu/nll/?p=1696#more-1696 [Accessed August 4, 2013]


Linguistic Changes in Canada Entering in the 21st Century. The Social Sciences and Humanities Research Council of Canada (SSHRC) for research grant #410-2003-005.


Appendices

Appendix A: Coding procedure for information status factor group

Operationalizing information status:
Possessor head noun of genitive or head noun of relative clause antecedent NP occurs within the last 100 words of the speaker

\[
\begin{align*}
N &= \text{New information, head noun does not occur within last 100 words} \\
O &= \text{Old ("given") information, head noun does occur within last 100 words}
\end{align*}
\]

To operationalize this factor group, text files of speaker transcripts were annotated by bracketing the possessor head noun of each interchangeable genitive token, or the antecedent head noun of each restrictive relative clause token. The following 299-word excerpt uses the example of the possessor head nouns of the genitives. For reference, the full genitive tokens are enclosed in [square brackets], while the possessor head nouns are enclosed in {curly brackets} (underlining here is only to facilitate clarity in the example). Note that for plural nouns, like offices, the plural –s morpheme is not bracketed, under the assumption that the root of the form should count as a “hit” for given information. For irregular plurals such as people, the search was performed on both the singular and plural forms.

Sample text (from 1956 Hansard, John Borden Hamilton, MP):

Mr. Speaker, I shall speak for only a few minutes in connection with [second reading of this {bill}], because I feel that at the resolution stage I expressed my views on as wide a basis as possible. There are, however, a couple of observations that I should like to make and I shall avoid, if at all possible, repeating what I said on the previous occasion.

I believe there are a few matters which might be considered by the committee upon [reference of this {bill}] to them. One of the things that it seems to me will be required for [the effective broadening of {the {scope} of {the {work} of this {institution}}}] is [the opening of additional {office}s of the {bank}] in more centres in Canada. We referred to that matter previously. Apparently there are only four regional offices. If we are going to make this type of money available to smaller business institutions, we must get down to their level. I would hope that the committee might give serious consideration to a recommendation of that kind when it is considering this bill.

The second thing I want to mention is this. The additional funds that are being made available, it appears, are to be limited to those institutions which require loans of $200,000 or more. An institution of that type can ordinarily go into the bond market, through the usual financial institutions, and obtain the kind of money it needs. I would think that what the country needs is an expansion of funds for those institutions which require help between the help that can be given to them by ordinary banking methods and the help they can obtain in the bond market. I would suggest that more consideration might be given to this phase of [the {bank}’s activities].
A Perl script (written by Daniel Jankowski) is then run on the annotated text file. The script counts back 100 words from each open-bracket, searching on the bracketed content. It then generates an output like the following:

```
1:[bill] 0
2:[bill] 1 bill
3:[scope] 0
4:[work] 0
5:[institution] 0
6:[office] 0
7:[bank] 1 banking
8:[bank] 1 bank
```

This output identifies each notated head noun. The number in the second column indicates how many “hits” were found on the bracketed content. If a hit is found, the third column contains the word that was counted as a hit. This allows for double-checking to avoid false hits. For example, a head noun *us* would return a hit on many words, such as *usual, peruse*, etc. Tokens were then assigned a code of N if no hits were returned, or G for a confirmed hit.
Appendix B: Coding procedure for *type-token ratio (lexical density)*
factor group, including sample part-of-speech tagged output.

**Operationalizing Type-token ratio/lexical density:**
Number of unique lexical categories within 50 words on either side of a genitive or relative clause token. Adapted from Hinrichs & Szmrecsanyi (2007:27-29), Szmrecsanyi & Hinrichs (2008:11), but counting lexical *categories* instead of lexical *items* (see section 2.4.2, pg. 39, fn 27 for discussion). *Theoretical max is 16, actual range found in the data was from 5 – 15 unique lexical categories.*

\[ f = 5, 6 = 6, 7 = 7, 8 = 8, 9 = 9, 10 = 10, 11 = 1, 12 = 12, 13 = 13, 14 = 14, 15 = 15 \]

To operationalize this factor group, text files of speaker transcripts were annotated by bracketing each token of an interchangeable genitive, or restrictive relative clause. The following 299-word excerpt uses the relative clauses as an example. For reference, the full relative clause tokens are enclosed in square brackets. To facilitate clarity in the example here, the relative pronouns are underlined, and the Ø variant is indicated:

Mr. Speaker, I shall speak for only a few minutes in connection with second reading of this bill, because I feel that at the resolution stage I expressed my views on as wide a basis as possible. There are, however, [a couple of observations that I should like to make] and I shall avoid, if at all possible, repeating what I said on the previous occasion.

I believe there are [a few matters which might be considered by the committee] upon reference of this bill to them. [One of the things that it seems to me will be required] for the effective broadening of the scope of the work of this institution is the opening of additional offices of the bank in more centres in Canada. We referred to that matter previously. Apparently there are only four regional offices. If we are going to make this type of money available to smaller business institutions, we must get down to their level. I would hope that the committee might give serious consideration to a recommendation of that kind when it is considering this bill.

[The second thing Ø I want to mention] is this. [The additional funds that are being made available], it appears, are to be limited to [those institutions which require loans of $200,000 or more]. An institution of that type can ordinarily go into the bond market, through the usual financial institutions, and obtain [the kind of money Ø it needs]. I would think that what the country needs is an expansion of funds for [those institutions which require help] between [the help that can be given to them] by ordinary banking methods and [the help Ø they can obtain] in the bond market. I would suggest that more consideration might be given to this phase of the bank's activities.

The annotated files are then tagged for part of speech using the “Tree-tagger” part-of-speech tagging software (Schmid, 1994a, 1994b, 1995). For the sake of simplifying the output to facilitate coding and running as a discrete variable in GoldVarb, the initial Tree-Tagger output of 36 different parts of speech was further reduced into coarser distinctions of a maximum of 16 unique lexical categories. This meant, for instance, collapsing number distinctions for nouns,
tense distinctions for the verbs, etc. The lexical categories used are listed at the end of this appendix, immediately after the sample part-of-speech tagged output below.

Though there are more sophisticated part-of-speech taggers, Tree-tagger has three great benefits. It is freely downloadable, Macintosh-compliant, and simple to run. Further, the coarseness of the 36 lexical categories that it distinguishes facilitates the coding of a discrete variable in GoldVarb. Software documentation (Schmid, 1994b, 1995) claims an accuracy rate of about 95 percent. Files were lightly post-edited and corrected by hand. Categories that were most commonly tagged incorrectly by the software were identified, and these were checked by hand (see the list of lexical categories at the end of this appendix). Every text file was run twice — once annotated for relative clause tokens, and a second time annotated for genitive tokens.

Once the text transcripts were marked up, with each viable token set off in brackets, a Perl script (written by Daniel Jankowski) is then run on the annotated text file. The script tallies the number of unique lexical categories within a 50-words window before and after each bracketed item. It then generates an output like the following:

0  [a couple of observations I should like to make]  12
1  [a few matters which might be considered by the committee]  12
2  [One of the things that it seems to me will be required]  12
3  [The second thing I want to mention]  12
4  [The additional funds that are being made available]  10
5  [those institutions which require loans of $200,000 or more]  10
6  [the kind of money it needs]  10
7  [those institutions which require help]  12
8  [the help that can be given to them]  12
9  [the help they can obtain]  12

This output identifies each bracketed token and the number in the rightmost column is the tally. Although the theoretical maximum number of unique lexical categories is 16, actual range found in the data was from 5 – 15 unique lexical categories.
Sample PoS-tagged text:

Mr. NP Mr. avoid VB avoid
Speaker NP Speaker if SC if
I PP I at IN at
shall MD shall all DT all
speak VB speak possible JJ possible
for IN for repeating VBG repeat
only RB only what WP what
a DT a I PP I
few JJ few said VBD say
minutes NNS minute on IN on
in IN in the DT the
connection NN connection previous JJ previous
with IN with occasion NN occasion
second JJ second . SENT .
reading NN reading I PP I
of IN of believe VBP believe
this DT this there EX there
bill NN bill are VBP be
because SC because [ SYM []
I PP I a DT a
feel VBP feel few JJ few
that SC that matters NNS matter
at IN at which PP which
the DT the might MD might
resolution NN resolution be VB be
stage NN stage considered VBN consider
I PP I by IN by
expressed VBD express the DT the
my PP$ my committee NN committee
views NNS view } SYM ]
on IN on upon IN upon
as RB as reference NN reference
wide JJ wide of IN of
a DT a this DT this
basis NN basis bill NN bill
as IN as to TO to
possible JJ possible them PP them
. SENT .
There EX there [ SYM [
are VBP be One CD One
however RB however of IN of
[ SYM [ the DT the
a DT a things NNS thing
couple NN couple that PP that
of IN of it PP it
observations NNS observation seems VBZ seem
that SC that to TO to
I PP I me PP me
should MD should will MD will
like VB like be VB be
to TO to required VBN require
make VB make } SYM ]
] SYM ] for IN for
and CC and the DT the
I PP I effective JJ effective
shall MD shall broadening VBG broaden
of the scope of the work of this institution is the opening of additional offices of the bank in more centres in Canada.

We referred to that matter previously.

Apparently there are only four regional offices.

If we are going to make this type of money available to smaller business institutions we must get down to their level.

I would hope that the committee might give serious consideration to a recommendation of that kind when it is considering this bill.

The second thing I want to mention is this.

The additional funds that are being made available it appears are to be limited to those...
An institution of that type can ordinarily go into the bond market through the usual financial institutions and obtain the kind of money it needs.

I would think that what the country needs is a n expansion of funds for those institutions which require help between the help that can be given to them by ordinary banking methods and the help they can obtain in the bond market.

I would suggest that more consideration might be given to this phase of the bank's activities.
Key to abbreviations for Tree-tagger output:

CC = Coordinating conjunction
CD = Cardinal number
   LS = List item marker (recoded as number)
DT = Determiner
   WDT = Wh-determiner (confirmed by hand, recoded as determiner)
EX = Existential there
FW = Foreign word
IN = Preposition (confirmed by hand)
SC = subordinating conjunction (confirmed by hand, initially tagged as ‘IN’)
JJ = Adjective
   JJR = Adjective, comparative (recoded as Adjective)
   JJS = Adjective, superlative (recoded as Adjective)
MD = Modal
NN = Noun, singular or mass
   NNS = Noun, plural (recoded as Noun)
   NP = Proper noun, singular (recoded as Noun)
   NPS = Proper noun, plural (recoded as Noun)
PDT = Predeterminer (quantifier)
POS = Possessive ending
PP = Personal pronoun
   PP$ = Possessive pronoun (recoded as pronoun)
   WP = Wh-pronoun (recoded as pronoun)
   WPS = Possessive wh-pronoun (recoded as pronoun)
   Relative pronouns confirmed by hand as pronouns, were often mis-tagged as DT
RB = Adverb
   RBR = Adverb, comparative (recoded as Adverb)
   RBS = Adverb, superlative (recoded as Adverb)
   WRB = Wh-adverb (recoded as Adverb)
RP = Particle (preposition or adverb; confirmed by hand and recoded)
TO = to (confirmed by hand, recoded if preposition)
UH = Interjection (none found)
VB = Verb, base form [TO BE]
   VBD = Verb, past tense (recoded as Verb)
   VBG = Verb, gerund or present participle (recoded as Verb)
   VBN = Verb, past participle (recoded as Verb)
   VBP = Verb, non-3rd person singular present (recoded as Verb)
   VBZ = Verb, 3rd person singular present (recoded as Verb)
   VH = Verb, base form [TO HAVE] (recoded as Verb)
   VHD = Verb, past tense (recoded as Verb)
   VHG = Verb, gerund or present participle (recoded as Verb)
   VHN = Verb, past participle (recoded as Verb)
   VHP = Verb, non-3rd person singular present (recoded as Verb)
   VHZ = Verb, 3rd person singular present (recoded as Verb)
   VV = Verb, base form [TO ANYTHING ELSE] (recoded as Verb)
   VVD = Verb, past tense (recoded as Verb)
   VVG = Verb, gerund or present participle (recoded as Verb)
   VVN = Verb, past participle (recoded as Verb)
   VVP = Verb, non-3rd person singular present (recoded as Verb)
   VVZ = Verb, 3rd person singular present (recoded as Verb)
Appendix C: Further details of Maclean’s & Hansard corpora
(Speakers, pertinent biographical information and word counts.)

Sample Design for Maclean’s and Hansard corpora

<table>
<thead>
<tr>
<th>Maclean’s magazine (100,000+ words)</th>
<th>Hansard transcripts (200,000+ words)</th>
</tr>
</thead>
<tbody>
<tr>
<td># authors</td>
<td># speakers</td>
</tr>
<tr>
<td>1906–12: 11 (34,000 words)</td>
<td>1906 8 (67,000 words)</td>
</tr>
<tr>
<td>1956–59: 7 (34,000 words)</td>
<td>1956 7 (67,000 words)</td>
</tr>
<tr>
<td>2006–07: 7 (34,000 words)</td>
<td>2006 9 (67,000 words)</td>
</tr>
</tbody>
</table>

Criteria
Men only, all between 30–65 at the time of sample, born and raised in Ontario.

1906 cutoff DOB for 60 y.o.: 1846
1912 cutoff DOB for 60 y.o.: 1852
1956 cutoff DOB for 60 y.o.: 1896
2006 cutoff DOB for 60 y.o.: 1946

Speakers — 1906 Hansard: Federal House of Commons

Armstrong, Joseph Elijah (b.1866.11.09, York County, Canada West, rep. Lambton East)
Bristol, The Hon. Edmund James, P.C., (b.1861.09.04, Napanee, Canada West, btw. Belleville & Kingston)
MacDonell, The Hon. Angus Claude (b.1861.06.23, Toronto, Ontario, Canada)
Gordon, David Alexander (b.1858.01.18, Wallaceburg, Canada West, N. of Chatam, S. of Sarnia)
Pardee, The Hon. Frederick, Forsyth (b. 1866.12.29, Sarnia, Canada West)
Hall, Robert Richard (b.1865.12.10, Fenelon Township, Canada West (now Kawartha Lakes, ON)
Bennett, The Hon. William Humphrey (b.1859.12.23, Barrie, Canada West)

Speakers — 1956 Hansard: Federal House of Commons

Wallace Bickford Nesbitt (b. August 7, 1918, Woodstock, Ontario)
John Borden Hamilton (b. May 16, 1913, Barrie, Ontario)
The Hon. Paul Theodore Hellyer (b. August 6, 1923, Waterford, Ontario)
Frank A. Enfield (b. June 16, 1920, Toronto, Ontario)
Allan Henry Hollingworth (b. August 29, 1918, Brockville, Ontario)
John Cameron Pallett (b. February 15, 1921, Dixie, Ontario)
Frederick Greystock Robertson (b. March 7, 1909, Belleville, Ontario)
Speakers — 2006 Hansard: Federal House of Commons

Dean Allison (b. February 18, 1965, London, Ontario)
Gordon Brown (b. August 31, 1960, Toronto, Ontario)
Mark Holland (b. October 16, 1974, Pickering, Ontario)
Brian A. Masse (b. July 9, 1968, Windsor, Ontario)
David J. McGuinty (b. February 25, 1960, Ottawa, Ontario)
Larry Miller (b. July 21, 1956, Wiarton, Ontario)
David Ven Kestern (b. October 7, 1955, Chatham, Ontario)
Bruce Stanton (b. December 20, 1957, Orillia, Ontario)
Michael Ignatieff (b. May 12, 1947, Toronto, Ontario)

Authors — 1905–6, 1910–12 Maclean’s/Busy Man’s Magazine

H. F. Gadsby (b. 5 April 1869, ST. Catherine's, Ontario, d. 1981)
C. M. (Clement Melville) Keyes (b. 1876 Chatsworth, Ontario)
Henry Horton Miller (b. 1861.01.10 Owen Sound, Ontario)
Ira Stratton (b. 1865 Trenton, Ontario, moved to Manitoba in 1889, age 24)
William Arnot Craick (b. 1880, Port Hope, Ontario)
Archie (Archibald) P. McKinshie (b. 20 June 1873 d.1939, Kent District, Harwich Twp., Ontario)
Frank Yeigh (b. 21 July 1861, Burford, Ontario, d.1935)
John Bruce Cowan (b. 1882 Chelsey, Ontario, d. 1970)
Henry John Pettypiece (b. 11 Nov. 1855, Anderson Twp., Lambton Co. Ontario)
J. B. Tyrell (b. 1858, Weston, Ontario)
Harry W. Anderson (b. 26 March 1876, Reading, England, sent to Chatam, Ontario at age 11)
M. O. Hammond (b. 17 July 1876, Clarkson, Ontario)
Henry Gerald Wade (b. 1875, Port Hope Ontario)

Authors — 1956–59 Maclean’s

Robert Fulford (b. Feb 13, 1932, Ottawa, Ontario)
Sidney Katz (b. approx. 1914–16, Ottawa, Ontario)
Robert Thomas Allen (b. 1911, grew up near Danforth & Coxwell, Toronto)
Fred Bodsworth (b. 1918 Port Burwell, Ontario)
Leslie Bell (b. 1906 Toronto, Ontario)
John McNaught (psed. “James Bannerman”) (b. 1902 Toronto, d. 1970)
Hugh Garner (b. 1913 Yorkshire, moved to Toronto, Ontario 1919)
Peter Gzowski (b. July 13, 1934, Toronto, Ontario)
Authors — 2006 Maclean’s

Michael Friscolanti (b. approx. 1976, raised Hamilton, Ontario)
Mark Steyn (b. December 8, 1959, Toronto, Ontario)
John Geddes (b. 1961 Shawville, Quebec, grew up in Cochenour, Ontario)
Paul Wells (b. 1966, Sarnia, Ontario)
Steve Maich (b. approx. 1974-75. Grew up in Markham and Thornhill, Ontario)
Scott Feschuk (b. approx. 1967–68, Grew up in southwest Ontario)
Michael Ignatieff (b. May 12, 1947, Toronto, Ontario, see above under 2006 Hansard)

Goals for Corpora Size
Hansard Goal: min. 200,000 words (min. 66,667 words each time period, max. 90-100,000 = 300,000)
Maclean's Goal: 100,000 words (min. 33,333 words each time period, max 60-65,000 = 180,000)

Word Counts
1906 Hansard: 89,634 words
Joseph Elijah Armstrong: 19,205
The Hon. William Humphrey Bennett: 21,167
The Hon. Edmund James Bristol, P.C.: 9,982
David Alexander Gordon: 2,460
Robert Richard Hall: 8,647
The Hon. Angus Claude MacDonell: 9,399
The Hon. Leighton Goldie McCarthy, P.C.: 9,709
The Hon. Frederick Forsyth Pardee: 10,065

1956 Hansard: 107,343 words
Wallace Bickford Nesbitt: 21,528
John Borden Hamilton: 20,013
The Hon. Paul Theodore Hellyer: 4,511
Frank A. Enfield: 17,040
Allan Henry Hollingworth: 16,189
John Cameron Pallett: 17,430
Frederick Greystock Robertson: 10,632

2006 Hansard: 96,078 words
Dean Allison: 6,017
Gordon Brown: 8,535
Mark Holland: 15,048
Brian A. Masse: 19,301
David J. McGuinty: 16,055
Larry Miller: 5,795
David Ven Kesteren: 8,113
Bruce Stanton: 12,579
Michael Ignatieff: 4,635
1906–12 Busy Man’s/Maclean’s Magazine: 67,320 words
Augustus Bridle: 5,837
H. Franklin Gadsby: 6,998
C. M. (Clement Melville) Keyes: 2,548
Henry Horton Miller: 3,406
Ira Stratton: 924
William Arnot Craick: 10,547
Archie (Archibald) P. McKinshie: 4,806
Frank Yeigh: 9,428
John Bruce Cowan: 1,700
Henry John Pettypiece: 3,758
J. B. Tyrell: 4,558
Harry W. Anderson: 3,996
M. O. Hammond: 4,505
Henry Gerald Wade: 1,942

1956–59 Maclean’s: 66,695 words
Robert Fulford: 4,084
Sidney Katz: 9,764
Robert Thomas Allen: 9,704
Fred Bodsworth: 10,762
Leslie Bell: 5,050
John McNaught (pseud. "James Bannerman"): 9,081
Hugh Garner: 9,112
Peter Gzowski: 9,182

2006 Maclean’s: 68,707 words
Michael Friscolanti: 9,160
Mark Steyn: 8,690
John Geddes: 9,964
Paul Wells: 10,228
Steve Maich: 10,060
Joseph Boyden: 7,784
Scott Feschuk: 9,813
Michael Ignatieff: 3,008
Appendix D: Common Coding instructions for Chapter 3 (Relatives) and Chapter 4 (Genitives)

**Corpus and Year:**

- 0 = Busy Man’s Magazine 1906-10
- 1 = Busy Man’s/Maclean’s (as of March 1911) Magazine 1911-16
- 5 = Maclean’s 1956-59
- Ø = Maclean’s 2006-2009
- A = Hansard 1906
- B = Hansard 1956
- C = Hansard 2006

**Writer/speaker:**

**Maclean’s 1906-1912**

- a = Augustus Bridle
- b = H. F. Gadsby
- c = C. M. (Clement Melville) Keyes
- d = Henry Horton Miller
- e = Ira Stratton
- f = William Arnot Craick
- g = Archie (Archibald) P. McKinshie
- h = Frank Yeigh
- i = John Bruce Cowan
- j = Henry John Pettypiece
- k = J. B. Tyrell
- l = Harry W. Anderson
- m = M. O. Hammond
- n = Henry Gerald Wade

**Maclean’s 1955-1958**

- o = Robert Fulford
- p = Sidney Katz
- q = Robert Thomas Allen
- r = Fred Bodsworth
- s = Leslie Bell
- t = "James Bannerman"/John McNaught
- u = Hugh Garner
- v = Peter Gzowski

**Maclean’s 2006**

- w = Michael Friscolanti
- x = Mark Steyn
- y = John Geddes
- z = Paul Wells
- Ø = Michael Ignatieff (see also speaker Ø in Hansard 2006)
- Y = Joseph Boyden
- Z = Steve Maich
- X = Scott Feschuk

**Hansard 1906**

- A = Joseph E. Armstrong
- B = The Hon. Edmund James Bristol
- C = The Hon. Angus Claude MacDonell
- D = The Hon. William Humphrey Bennett
- E = The Hon. Leighton Goldie McCarthy, P.C.
- F = The Hon. Frederick Forsyth Pardee
- G = David Alexander Gordon
- H = Robert Richard Hall

**Hansard 1956**

- I = John Cameron Pallett
- J = John Borden Hamilton
- K = The Hon. Paul Theodore Hellyer
- L = Frank A. Enfield
- M = Allan Henry Hollingworth
- N = Wallace Bickford Nesbitt
- O = Frederick Greystock Robertson
Animacy of Antecedent of Relative Clause OR Possessor Head Noun
(adapted and expanded from Tagliamonte et al., 2005, after Tagliamonte & Jankowski, forthcoming)

- **h** = human sg.
- **H** = human pl.
- **p** = "people(s)", "person(s)"
- **P** = person
- **n** = proper name sg.
- **N** = proper name pl.
- **C** = collection of humans e.g. family, band, group, committee, party, police (men), the community organization, Southerners EXCEPT PEOPLE
- **a** = animate non-human (animal) sg.
- **A** = animate non-human (animal) pl.
- **i** = inanimate concrete object sg.
- **l** = inanimate concrete object pl. (includes mass count nouns, like "gold", "coal", "money")
- **t** = non-concrete inanimates (which are not activities, events, states, conditions, time, etc., e.g. 'the tariff', 'the bill', 'the field that lay before him' (not an actual place), 'the only word', 'a secret')
- **L** = place (not a specific country)
- **E** = specific country, town, street, province (e.g. Canada, the colonies, the Mother Country)
- **v** = activity/event
- **x** = state/condition, property (e.g. 'the gift of discreet statesmanship')
- **b** = body part
- **t** = time/temporal noun
- **o** = organization (e.g. this House, Chicago Post, Dalhousie Law School, school district, provincial riding, government department, "the police (force)", "the press"

**Type-token ratio/lexical density:**
Number of different types of parts of speech (i.e. unique lexical categories) present within 50 words either side of s-genitive or relative clause token. (See appendix B for description of part-of-speech tagging and coding procedure)
Theoretical max is 16, actual range is from 5 - 15

- **f** = 5, **g** = 6, **w** = 7, **s** = 8, **9** = 9, **o** = 10, **1** = 11, **2** = 12, **3** = 13, **4** = 14, **5** = 15

**Information status:**
Possessor head noun of genitive or head noun of relative clause antecedent NP occurs within the last 100 words of the speaker. (See appendix A for description of coding procedure)

- **N** = New information, head noun does not occur within last 100 words
- **o** = Old ("given") information, head noun does occur within last 100 words
Appendix E: Coding instructions specific to Relative Pronouns (Chapter 3)

Relative pronoun:
- $T =$ THAT, $O =$ WHO, $Z =$ ZERO, $A =$ WHAT, $C =$ WHICH, $S =$ WHOSE, $W =$ WHOM, $E =$ AS, $W =$ WHERE

Grammatical function of the relative in the relative clause:
(adapted from Tagliamonte et al., 2005; D'Arcy & Tagliamonte, 2010)
- $S =$ subject, e.g. It's him THAT works it.
- $O =$ direct object, e.g. This is [the doll] that he gave [Ø] to the girl ...
- $I =$ indirect object, e.g. This is [the girl] that he gave the doll [Ø] ...
- $G =$ genitive
- $P =$ object of preposition {stranded}, if the verb is DITRANSITIVE, e.g. give, tell, show, etc.; This is [the girl] that he gave the doll {to} [Ø] ...
- $D =$ object of preposition (pied-piped) e.g. think about, think of, worry about,
- $V =$ direct object of a phrasal verb; e.g. Well, it was one Ø I'd BUILT UP myself. Also FIND OUT X, THROW OUT X, LOOK AFTER X etc.

Type of NP of the antecedent:
(as per Tagliamonte et al., 2005:98)
- $D =$ NP, definite
- $d =$ Pronoun, definite
- $A =$ NP, indefinite
- $a =$ Pronoun, indefinite

Other properties of antecedent: (said to influence that/which choice)
- $Q =$ quantifier/comparative/superlative NP modification (for recode purposes, said to prefer 'that' over 'which')
- $I =$ Strong indefinite quantifiers (all, both, etc.)
- $q =$ indefinite quantifiers (each, every, either, no, few, more, some, such, plenty, a lot, any, many, most etc.)

Determiners:
- $x =$ other quantifiers (certain, various, others, almost, nearly)
- $y =$ only
- $s =$ superlatives (the best/hardest/biggest/worst thing/most common)
- $c =$ comparatives (the more, better, harder, bigger, worse, X THAT came in...)
- $n =$ cardinal quantifier (one, two, etc.)

NP modification:
- $D =$ demonstrative determiners (this, that, these, those)
  (said to prefer 'which' over 'that')
- $- =$ NP with no quantifier or demonstrative
- $w =$ question words (which, what, etc.)
- $P =$ possessive nouns (Joe's, the priest's, my mother's)
- $N =$ bare NP (but could include descriptive adjectives)
- $t =$ definite determiner, but no other quantifiers (the)
a = indefinite determiner, but no other quantifiers (article a(n))
P = Pronoun
d = demonstrative pronouns (this, that, these, those)
E = Strong indefinite pronouns (all, both, etc.)
i = indefinite pronouns (each, every, either, no, few, more, some, such, plenty, a lot, any, many, much, etc.)
p = possessive pronouns, (his, your, their, whose, etc.)
2 = cardinal pronoun
M = bare pronoun
o = ordinal (first, second, etc.)

**Length and Complexity of relative clause:**
(Interaction FG as per Tagliamonte et al., 2005:97-8)

**Simple, Verb and non-clausal arguments:**
- **s** = Short (1-3 words) Simple (Verb and non-clausal arguments)
- **S** = Long (4+ words) Simple (Verb and non-clausal arguments)

- **m** = Short (1-3 words) middle (Verb and non-clausal arguments, including prepositional or adjectival phrases)
- **M** = Long (4+ words) middle (Verb and non-clausal arguments, including prepositional or adjectival phrases)

**Complex, has a clausal argument:**
- **c** = Short (1-3 words) Complex (has a clausal argument)
- **C** = Long (4+ words) Complex (including verb and clausal arguments)

**Adjacency:**
(as per Tagliamonte et al., 2005:97)
- **A** = Adjacent
- **N** = non-Adjacent, i.e. There was a man [called J. B.] Ø used to live over there.

**Plurality of antecedent:**
(as per D'Arcy & Tagliamonte 2010)
- **P** = plural ('people' is plural, 'persons' is plural), (NOTE: coded mass count nouns as plural)
- **S** = Singular ('person' is singular)
Appendix F: Coding instructions specific to Genitives (Chapter 4)

Genitive marker:
S = 's; s = s'; F = of

Length of possessor: (thing possessing something)
1 through 8; 9 = 9+

Possessum end weight: (thing being possessed)
1 through 8; 9 = 9+

Final sibilant in possessor: (as per Hinrichs & Szmrecsanyi, 2007:20-22)
s, z, c, j = final sibilant [s], [z], [sh/tsh], [zh/dzh]
- = no final sibilant

Animacy by possessor length: (cross-tabulated factor group)
a = human, sg., 1-word
b = human, sg., 2-word
c = human, pl., 1-word
d = human, pl., 2-word
e = organization, 1-word
f = organization, 2+word
x = collective, 1-word
y = collective, 2-word
g = place, 1-word
h = place, 2+word
i = inanimate objects, 1-word
j = inanimate objects, 2+word
k = other inanimates (state, activity time, body parts), 1-word
l = other inanimates (state, activity, time, body parts), 2+word
m = animate non-human, 1-word
n = animate non-human, 2+word

Possessive Relation: (adapted from Tagliamonte & Jankowski, forthcoming)
I = Inalienable/prototypical (body parts, kin terms, permanent/legal ownership, part/whole); e.g. my mother’s sisters
A = Alienable/non-prototypical (states, abstract possession, non-part/whole); e.g. somebody’s boat
? = Ambiguous/unclear tokens; e.g. And it had boards around the edge of it

Lexical Class of Possessor: (adapted from Tagliamonte & Jankowski, forthcoming)
P = proper noun (e.g. JOHN’s expressions)
I = indefinite pronoun (e.g. SOMEONE’s yard)
c = common noun, generic (e.g. the pace OF LIFE over there)
f = Indefinite NP (e.g. Guys checking out girls’ butts)
t = Definite NP (e.g. the name OF THE PLACE)
i = Personal pronoun (e.g. So at the core OF IT-)
D = Definite pronoun (e.g. What’s THE OTHER ONE’S name?)
d = Demonstrative NP (e.g. THAT KID’S hair)
m = Possessive NP (e.g. MY MOTHER’S child)
? = Unclear token (e.g. we were in THIS GUY’S house)