

Quirky Quotes and Needles in the Haystack: Tracing Grammatical Change in Untagged Corpora

Muriel Norde (Berlin)

Abstract

This paper discusses pivotal theoretical and methodological problems of historical corpus linguistics. In two case studies from Swedish language history, the development of the epistemic adverb *kanske* and the group genitive respectively, it illustrates how the use of qualitative method in addition to corpus investigation can contribute to understanding grammatical change.

1 Introduction

Tracing grammatical change in historical corpora is a rewarding, albeit challenging task. In this paper I will discuss some of the theoretical traps and empirical pitfalls one is typically confronted with in historical corpus linguistics. First of all, we have to define what 'change' entails. This might seem rather straightforward, but 'change' is often confused with 'correspondence' (in Andersen's (2001) sense), and on closer inspection, a change may in fact comprise several smaller, primitive changes. One has also to bear in mind that change is gradual, which makes it difficult, if not impossible, to identify discrete stages in a change (or rather, in a chain of changes). And finally, change, particularly grammatical change, is not unidirectional, which makes it difficult to reconstruct it.

For the reasons given in the preceding paragraph, the only viable method for uncovering change in times long past is historical corpus linguistics, but this brings along its own share of issues. The historical material may be fragmented, chronologically discontinuous, or stylistically imbalanced, all of which seriously reduces the possibility of collecting random samples. Automated quantitative studies are difficult, if not impossible to carry out, particularly when the corpora lack part-of-speech (POS) tagging. In the absence of such tags, all one can do is search for particular strings ((parts of) words or larger units), using regular concordance software such as WordSmith lexical analysis software (Scott 2004). Especially in older, non-standardized texts, one furthermore has to consider variation in spelling, which may be substantial. In this paper, I will discuss these theoretical and data-related challenges, illustrated by two case studies from Swedish language history. Specifically, I will discuss two types of change, which I will term 'quirky quotes', and 'needles in the haystack', and the crucial role they may play in the qualitative approach to change.

The organization of the paper is as follows: in section 2, I will outline the theoretical preliminaries, and in section 3 the methodological issues. In section 4, I will present two case

studies from the history of Swedish, which represent the types of change alluded to in the title of this paper. Finally, I will summarize my findings in section 5.

2 Theoretical preliminaries

2.1 The correspondence trap

'Change' is an elusive concept, and often confused with what Andersen (2001: 228) terms 'diachronic correspondences', defined as the "relation between homologous elements [...] belonging to two chronologically separate synchronic states in a linguistic tradition [...]." In other words, if one considers form α in period A, which is etymologically related to form β in period B, one is tempted to conclude that α changed to β , when in fact all one has done is contrast two forms, in which form β is the **result** of change. Yet what constitutes the change(s) is what actually **happened** between period A and period B, or to quote Andersen (2001: 228) once more: "the historical events in a linguistic tradition by which practices of speaking vary over time." Lass (1997: 288) put it this way:

If I drive from Edinburgh to London, make some stops for petrol, and take a brief trip east on the way to visit a friend in Cambridge, I can still be said (from the point of view of 'the accomplishment', or juxtaposition of initial and final states) to 'have driven from Edinburgh to London'. How I got there is another (kind of) story. [...] We seem usually to be thinking of macro-stories when we talk about 'change'; but the micro-stories are of enormous theoretical importance as well.

One example of the 'correspondence trap' is the well-known example of the suffix *MENTE*¹, found in all Romance languages except Romanian (e.g. Karlsson 1981; Bauer 2003; Torner 2005). Contemporary examples are French *heureusement* 'fortunately', Portuguese *cruamente* 'cruelly', Spanish *distintamente* 'distinctly', Italian *raramente* 'rarely', Occitan and Catalan *bellamen(t)* 'beautifully', and Sardinian *finalmenti(s)* 'finally' (Bauer 2003: 440). Originally the ablative singular of a noun meaning 'mind', which developed into a derivational suffix, *MENTE* has often been quoted as a stock example of grammaticalization (e.g. in Hopper/Traugott 2003: 141–142; Lehmann 1995 [1982]: 87; Giacalone Ramat 1998: 120; Ramat 2011: 508).

The nominal origin of the suffix, the Latin noun *mens* (FEM) 'mind', is traditionally illustrated by the following passage from Ovid's *Metamorphoses* XIII:

- (1) *consolor* *socios* *ut* *longi* *taedia* *belli*
 encourage-1SG allies-ACC so.that long-GEN boredoms-ACC war-GEN

mente *ferant* *placida*
 mind-ABL bear-SUBJ.3.PL quiet-ABL

'I encourage our allies so that they may bear the boredom of the long war with a quiet mind'

In the course of time, *mens* went through a series of metonymic meaning changes illustrated in (2) (for details of this development see Detges 1998):

¹ The suffix has slightly different forms in different languages; hence the notation *MENTE* is used to refer to all suffixes collectively.

- (2) 'mental state of the participant in the event' > 'way in which the event is perceived' > 'manner in which the event takes place'

Once the meaning 'manner' had been established, *mente* collocations were no longer restricted to adjectives expressing states of mind, sparking an increase in type frequency to the extent that adverbials involving *mente* replaced a number of the Classical Latin adverbs ending in *-e* and *-iter*. Crucially however, the change from FEM.SG.ABL noun to derivational suffix was not a straightforward one. Thus Bauer (2003: 447), in a careful study of the *Vulgata* bible, has shown that *mente* adverbials were far less frequent than *animo* adverbials, yet the latter did not develop into a suffix. A second problem is that it is not exactly known whether *mente* adverbials originated in the spoken or in the written language (see Hummel 2000 for discussion). Other loose ends in the history of MENTE include its morphological status as well as distributional differences at different stages of development in different Romance varieties (Norde 2009: 44–46). Finally, the development from Latin *mens* to MENTE has been shown to be discontinuous in some varieties, e.g. Spanish, in which the form *-mente* was borrowed, presumably from Aragonese, Catalan or even French, to replace the native suffix *-mientr(e)* (Torner 2005: 139).

What this yet inexhaustive discussion of Spanish *-mente* shows is that the history of this suffix is far more complex than a simple correspondence might suggest. Nevertheless, correspondences play an important role in historical research, because the changes themselves often go unnoticed by members of a speech community. More often than not, a striking correspondence forms the point of departure for a historical investigation.

Another illustration of the peril of basing linguistic reconstruction on correspondences alone is the alleged diachronic link between the s-genitive in Norwegian (example (3)a), and the so-called possessor doubling construction (example (3)b), in which the possessor is marked by a following possessive pronoun:

- (3) a. *den gamle mannen med skjeggets hus*
 [the old man-def with beard-def]S house
- b. *den gamle mannen med skjegget sitt hus*
 the old man-def with beard-def poss house
- 'the house of the old man with the beard'

These two constructions are not only similar in that they express a possessive relationship; they both occur in 'group genitive' constructions (as in the examples above), and both are largely confined to animate possessors. On the basis of these similarities, it has been suggested by Fiva (1987) and Lødrup (1989) that the s-genitive is a reduced form of the (reflexive) possessive pronoun. At a superficial glance, this may seem a phonologically plausible change, and one corroborated by the functional similarities between the two constructions. However, there is abundant historical evidence to the contrary. First, the (enclitic) s-genitive has been shown to derive from the former (inflectional) genitive case (Norde 1997; Trosterud 2001), and secondly, the s-genitive is at least as old as the possessor doubling construction (Norde 2012).

2.2 The nature of grammatical change

Grammatical change is very complex, encompassing changes at several levels. For example, when a noun develops into a preposition, there can be said to have been 'a shift from noun to preposition', but that is just the categorical reanalysis involved. What Lass (1997) means by 'micro-stories' in the quote in the preceding section are the changes at different linguistic levels. For example, when English *to be going to* grammaticalized into the future auxiliary *gonna*, it did not only change category, but went through a series of changes, which I will term 'primitive changes'. Primitive changes in the case of *gonna* include phonological reduction, loss of inflectional properties, and semantic bleaching. These changes are not entirely independent of one another – for instance, semantic bleaching (from 'moving on foot towards a certain goal' to FUTURE) opens the door to an expansion of contexts and an increase in frequency which in turn may result in phonological reduction. Yet they have to be examined separately, if only because they do not always occur simultaneously.

Another important observation in the functional-typological approach to grammatical change (the prevalent approach in grammaticalization studies) is that change is gradual, which may be reflected by synchronic gradience (Traugott/Trousdale 2010). When a new structure arises, the old one does not disappear at that very same instant; it continues to co-exist with the new structure, sometimes for a considerable period of time. This gradualness can be represented as follows (Hopper/Traugott 2003: 49):

(4) A > {A / B} > (B)

When construction A changes into B, it coexists with B, and in fact need not disappear at all (hence, the last stage is put in parentheses).

A final point to be made about the nature of grammatical change concerns directionality. In the 1990s, which saw a revived interest in grammaticalization, the view that grammatical change was unidirectional was quite widespread. This unidirectionality implied that lexical items could change into grammatical items and go on to adopt more grammatical functions, but not vice versa; in other words, there could be no **d**egrammaticalization. Some early studies (Campbell 1991; Ramat 1992), however, provided evidence that counterdirectional change, though rare, is by no means impossible, and while the body of counterdirectional evidence grew (Norde 2009), it became increasingly recognized that unidirectionality of change is a statistical universal, not an absolute one (e.g. Haspelmath 2004: 23). This has serious implications for grammatical reconstruction, for in the absence of historical evidence, a 'less grammatical' form cannot be reconstructed as historically prior to a 'more grammatical one', at least not with absolute certainty (for discussion of this issue see Norde 2009: 36–41).

2.3 Implications for historical linguistics

Summing up this section, diachronic linguistic research proceeds in two steps:

- Step 1: identifying correspondences;
- Step 2: identifying changes in order to
 - establish whether synchronic correspondence reflects diachronic correspondence;
 - identify the micro-changes that resulted in the correspondence.

The second step is essential – as unidirectionality is not an exceptionless principle of change; changes cannot be reconstructed. Needless to say, reconstruction may be the only method available, for instance in languages that lack written historical records. But whenever such records **are** available, I think they simply cannot and should not be ignored, tiresome as historical corpus linguistics may be (cf. section 4). In addition, we need to bear in mind that change is gradual, which implies that intermediate stages will also show gradience. As a result, lots of texts will have to be scrutinized in order to detect all micro-changes involved. This raises a number of methodological issues, to which I now turn.

3 Methodological issues

3.1 Building a corpus

Historical linguists have the disadvantage of not having access to the competence of speakers of past stages of a language, and hence they have to rely on evidence from historical records and/or linguistic reconstruction, both of which bring along their own problems. One of those problems, raised by Janda/Joseph (2003), concerns the over-representation of high-prestige sources: "there is little we can do to change the circumstance that the texts which most often tend to be written and preserved are those which least reflect everyday speech" (Janda/Joseph 2003: 17). Citing Labov's famous study of Philadelphia English, they argue that speakers tend to be much more consistent in spontaneous speech (in Labov's study: in the realization of /æ/ in *sad* versus /æh/ in *bad*) than when reading word-lists aloud. This is probably because writing favors both conservatism and hypercorrection. In other words, the variation attested in older texts need not reflect variation in the spoken language. "Broken threads" in language history pose another challenge (Janda/Joseph 2003: 19). This is notoriously true for English, where most of the oldest records are written in the Wessex dialect spoken in the West-Saxon kingdom, which was both politically and culturally dominant at that time, whereas Modern English descends from Mercian, spoken in and around London, which became powerful in the Middle Ages. This means that there exists no uninterrupted timeline from "old" to "modern" English.

Both problems are also significant in historical texts from Sweden, on which the case studies in the next section are based, with the additional problem of two different systems of writing. The oldest Swedish texts (circa 800–1100) are runic inscriptions. Although they may be considered a very rich source of the language of that time (more than 3000 inscriptions have been preserved), they may be difficult to interpret for two reasons. Firstly, there were only 16 runes for some 30 phonemes, and secondly, subsequent identical sounds were usually not repeated. For example, the phrase *ok Guðs móðir* 'and God's mother' was often carved **ukusmuþir** – there was only one rune for both /k/ and /g/, and this was not repeated, even though *ok* and *Guðs* are two separate words (Palm 2004: 112–113). From the 12th century, there are no Swedish sources, neither runic nor written – the first texts written in the Latin alphabet are from the 13th century. This means that there is a crucial gap in the documentation of Swedish language history. Another problem concerns the nature of the sources: the oldest manuscript texts were provincial laws (with long oral history) and charters, written in a very different style. In the next centuries, most texts were translations (among them religious treatises, legends and courtly literature, all translated from Latin, French or German). In other

words, attested differences between different texts need not (only) be chronological, they may also be due to register, style, or foreign influence (or a combination of these).

3.2 Dealing with negative evidence

Another problem with diachronic textual evidence is observed by Traugott (1989: 34):

All claims about the order of development that are based [...] on written records and evidence from grammars and dictionaries, must be regarded with caution. As is well known, attestation is often a matter of accident. Furthermore, it does not necessarily reflect changes in the spoken language. What is significant is cumulative evidence from different but related semantic domains, and, wherever possible, from other languages, of the same order of attestation among exemplars, whatever the time lag.

Lehmann (2004: 172) similarly argues that the absence of a given form or construction does not necessarily imply that it did not exist at the time, a problem for which he coined the apt phrase "non-demonstrability of non-existence". Janda/Joseph (2003: 15) likewise note that there may be "accidental gaps in the historical record". They provide the example of Ancient Greek *éor* which does not appear in written records before the fifth century AD, but must be much older than that, since it refers to a female relative of some kind and derives from PIE **swés(o)r* by regular sound laws. The non-occurrence of this word in the massive body of texts from the preceding centuries is purely accidental. Unfortunately however, Janda/Joseph (2003) do not reveal how frequent this word was in documents from the fifth century (and onwards?), but it cannot have been too frequent, given that the exact meaning of the word is not even known. Hence, it may have been extremely marginal, possibly confined to a very small and non-prestigious part of the Ancient Greek speech community. Therefore I think that this is a problem which should not be overemphasized. Surely we must always be aware that non-occurrence is not tantamount to non-existence, but the relative infrequency of this phenomenon should not inhibit us from using historical data.

4 Case studies

In this section, I will briefly present two case studies posing two different kinds of problems one frequently encounters in historical corpus research: "quirky quotes", and "needles in the haystack". Quirky quotes are examples of constructions that are perfectly inconsistent with attested patterns of development. The question is what to do with them – dismiss them as plain errors, or try to account for them? This type of data will be discussed in section 4.1.

Needles in the haystack are of a very different kind – they are examples that are extremely difficult to trace in untagged corpora, simply because they are not orthographically salient; hence they cannot be detected by means of regular search queries. An example will be given in section 4.2.

4.1 Quirky quotes

Quirky quotes, as stated above, are data that do not conform to the expected or attested path of development. They are "the odd ones out" and can, in principle, be dealt with in two ways. We can either dismiss them as "slips of the feather", or try to explain them as changes in their own right (and if we fail, decide they were probably slips of the feather after all). In this

section, I will discuss a case in which the quirky quotes eventually turned out to be highly relevant, with some serious implications for the initial hypothesis.

This case concerns the development of epistemic adverbs in the history of Swedish. These are sentence adverbs, meaning 'maybe', that originate in the univerbation of a modal verb meaning 'can' or 'may', and a main verb meaning 'happen': *kanske*, *kanhända*, *måhända*, *törhända* (Norde/Rawoens/Beijering in prep.).² From the point of view of Swedish main clause syntax (Beijering 2010), these adverbs are very interesting because they may violate verb-second (V2), i.e. the syntactic rule that the finite verb always appears in second position. In the examples below, (5)a and (5)b are V2-clauses, but in (5)c, it is the sentence adverb that appears in second position, whereas in (5)d the subject appears in second position. Example (5)e, finally, illustrates the phenomenon of insubordination (Evans 2007), in which a subordinate clause is not bound by a full matrix clause, but by an adverbial phrase.

- | | | | | | | | |
|-----|----|-------------------|---------------|---------------|-------------|---------------|----------|
| (5) | a. | <i>Olle</i> | <i>har</i> | <i>kanske</i> | <i>läst</i> | <i>boken.</i> | [V2] |
| | | Olle | has | maybe | read | book-DEF | |
| | b. | <i>Kanske</i> | <i>har</i> | <i>Olle</i> | <i>läst</i> | <i>boken</i> | [V2] |
| | | Maybe | has | Olle | read | book-DEF | |
| | c. | <i>Olle</i> | <i>kanske</i> | <i>har</i> | <i>läst</i> | <i>boken</i> | [non-V2] |
| | | Olle | maybe | has | read | book-DEF | |
| | d. | <i>Kanske</i> | <i>Olle</i> | <i>har</i> | <i>läst</i> | <i>boken</i> | [non-V2] |
| | | Maybe | Olle | has | read | book-DEF | |
| | e. | <i>Kanske att</i> | <i>Olle</i> | <i>har</i> | <i>läst</i> | <i>boken</i> | |
| | | Maybe that | Olle | has | read | book-DEF | |

'Maybe Olle has read the book'

The etymology of epistemic adverbs as deriving from an epistemic verb phrase (EpVP) is fairly uncontroversial, but note that at this point this is merely a correspondence, not a change (cf. section 2.1). A possible path of development for the adverb *kanske* has however been suggested by Wessén (1967). In Wessén's scenario, the development comprises five stages, which are illustrated below by Modern Swedish equivalents.

Stage I: The EpVP forms part of a full matrix clause, which is followed by a subordinate clause:

- | | | | | | |
|-----|----------------|-------------|------------|------------|----------------|
| (6) | <i>Det kan</i> | <i>ske,</i> | <i>att</i> | <i>han</i> | <i>kommer.</i> |
| | It can/may | happen, | that | he | comes. |

Stage II: The expletive subject *det* 'it' is dropped,³ and the modal verb and main verb merge into a single word, but the subordinate clause remains:

- | | | | | | | |
|-----|---------------|------------|------------|---------------|--------------|---------------|
| (7) | <i>Kanske</i> | <i>att</i> | <i>han</i> | <i>kommer</i> | <i>redan</i> | <i>i dag.</i> |
| | Maybe | that | he | comes | already | today. |

² This pattern is known from other European languages as well (Ramat/Ricca 1998: 212–216), e.g. English *maybe*, French *peut-être*, Russian *možet (byt')* meaning 'may (be)'.
³ There is evidence that in Old Swedish expletive subjects were not obligatory; so the absence of such a subject in *kanske*-constructions does not necessarily imply a separate stage, but rather that is not really relevant to the argumentation in this section. See Norde/Rawoens/Beijering (in prep.) for a detailed account of Wessén's stages.

Stage III: The subordinator *att* is dropped, resulting in a non-V2 clause, with the subject in second position instead of the finite verb:

- (8) *Kanske han kommer redan i dag.*
 Maybe he comes already today.

Stage IV: *Kanske* is still in clause-initial position, but the subject and the finite verb are reversed so that the clause no longer violates V2:

- (9) *Kanske kommer han redan i dag.*
 Maybe comes he already today.

Stage V: *Kanske* can occur in other positions for sentence adverbs:

- (10) *Han kommer kanske redan i dag.*
 He comes maybe already today.

Wessén's scenario is plausible, because it is gradual and accounts for most of the syntactic variability attested in Modern Swedish (cf. the examples in (5)), including the non-V2 construction in (5)d and the semi-autonomous subordination construction in (5)e. However, this scenario cannot account for the non-V2 construction in (5)c, repeated below as (11)a, because in this construction, the adverb *kanske* cannot be replaced by a full matrix + subordinate clause construction, as shown in (11)b. In other words, it would be difficult to explain why a non-V2 construction should arise that does not derive from a matrix clause construction.⁴

- (11) a. *Olle kanske har läst boken.* [non-V2]
 Olle maybe has read book-DEF
 b. **Olle det kan ske att har läst boken.*
 Olle it can happen that has read book-DEF

In order to test whether Wessén's scenario is reflected in historical texts, we carried out a large-scale corpus investigation into epistemic adverbs and epistemic verb phrases in the history of Swedish (Norde/Rawoens/Beijering in prep.). The corpus was about 1,668,500 words in size, comprising texts from the late 14th century to the end of the 18th century. From this corpus, we selected all instances of the sentence adverb *kanske* (which is the most common epistemic adverb in Swedish) as well as the infinitive forms (including their spelling variants) of the verbs meaning 'happen': *ske* and *hända*. All instances in which these main verbs combined with the modal verbs *kunna* 'can', *må* 'may' or *tör* 'may' were further analysed. Table 1 summarizes the total number of relevant constructions.

⁴ Andréasson (2002: 43) suggests that *kanske* in older Swedish may have had two functions: one as a fully developed sentence adverb, and one as an EpVP (orthographically identical to adverbial *kanske*). This phrasal *kanske*, then, would have been moved to a position where it could no longer be replaced by a verb phrase. To my mind however, analysing *kanske* as a phrase rather than an adverb does not really solve the problem of why a new non-V2 construction should arise in the first place.

		MiSw	EMoSw ⁵
MAIN VERB	MODAL		
<i>hända</i>	<i>kunna</i>	2	46
	<i>må</i>	2	1
	<i>tör</i>	–	15
<i>ske</i>	<i>kunna</i>	5	34
	<i>må</i>	2	–
	<i>tör</i>	–	1
total EpVP		11	97
<hr/>			
<i>kanske</i> (sentence adverb)		1	159

Table 1: EpVPs and kanske in the corpus

Among the results of this corpus investigation were a few quirky quotes that at first seemed difficult to explain. Three of them are quoted in (12). In (12)a, *kan ske* is written as two words, yet it cannot be analysed as a (subjectless) matrix clause, because it is not possible to add a subject and a subordinator (compare (12)a', which is ungrammatical). It clearly functions as an adverbial, as in Wessén's Stage V above, even though the adverb was written as one word from Wessén's Stage II onwards. This might suggest that univerbation of the modal verb and the main verb had not been completed when the sentence adverbial uses arose. However, orthography was not yet standardized at that time, and it was not unusual for compounds to be written as two words. In other words, *kan ske* may have been a single (compound) word anyway, in spite of its spelling. A far more problematic example, however, is (12)b: in this example, univerbation cannot possibly have occurred, because the adverb *wäl* is inserted between the modal verb and the main verb. Example (12)c, finally, is quirky for yet another reason, because the modal verb is inflected for past tense, whereas it is assumed that the adverb *kan ske* arose from a construction in which the modal was in the present tense (and indeed, the vast majority of examples of epistemic VPs involve the present tense).

- (12) a. *så hadhe iagh nu kan ske vahrit en annan carl* [1700]
then had I now may be been an other man
'in that case I would have been another man now'
- a'. **så hadhe iagh nu det kan ske att vahrit en annan carl*
so had I now it can happen that been an other man
- b. *gwenel iärl hafwer ma wäl ske swikith them* [14th century]
Gwenel earl has may well happen betrayed them
'Earl Gwenel maybe betrayed them'
- c. *Så hade the och icke heller ähn nu sett någin fiende* [...]
So had they also not before than now seen some enemy [...]

⁵ These texts cover parts of two periods in the history of the Swedish language: Middle Swedish (MiSw) and Early Modern Swedish (EMoSw). The year 1526, when the New Testament was translated into Swedish, marks the beginning of the EMoSw period.

kunde skie icke heller finge see [1585]
 could happen not either got see

'So up till now they had not seen a single enemy, perhaps they would not even get to see one'

The examples above clearly do not fit with Wessén's scenario, but they are too frequent to be ignored and beg for an explanation. And the explanation Norde/Rawoens/Beijering (in prep.) suggest is that the matrix clause was not the only source construction for the sentence adverb *kanske*. The examples in (12), we propose, do not derive from a sentence initial matrix clause, but from a parenthetical clause inserted in a main clause. The two source constructions are illustrated in figure 1. A is Wessén's scenario with a main clause as the source of the adverb, B is the alternative construction with a parenthetical clause as the source of the adverb. The quirky quotes, then, do not turn out to be quirky at all – they are simply indicative of an alternative route to adverbhood, which might not have been discovered otherwise. Moreover, the second scenario can account for the occurrence of non-V2 constructions, such as (11)a.

A.	<i>Det kan ske att han kommer.</i> ↓ <i>Kanske han kommer.</i> 'Maybe he's coming.'	[main clause + subordinate clause]
B.	<i>Han, det kan (väl) ske, kommer.</i> ↓ <i>Han kanske kommer.</i> 'Maybe he's coming.'	[parenthetical in main clause]

Figure 1: Source constructions of the Swedish adverb *kanske* 'maybe'

4.2 Needles in the haystack

Untagged corpora can be extremely challenging for the historical linguist who wants to study morphological change. With regular software (e.g. WordSmith, Scott 2004) it is possible to search for words, strings of words or (using wildcards) parts of words, but this is usually not very useful for identifying changes in, say, inflectional morphology. Inflections are typically short, mostly monosyllabic, and often even monophonemic. Obsolescent morphology is even more problematic because it is evidently impossible to search for the absence of inflection. In this section, I will discuss a case study of changes in inflectional patterns, where these problems present themselves. The case study concerns a part of the intriguing history of the development of the s-genitive found in English, Danish, Norwegian and Swedish. Once an inflectional suffix to mark the genitive case of some masculine and neuter singular nouns, adjectives and pronouns, it is at present a once-only marker which is attached to full noun phrases.⁶ The most impressive reflections of this change are so-called 'group genitives', in which the s-genitive appears on the very right edge of an NP containing a postmodifying PP

⁶ The exact morphological status, clitic or phrase-final affix, has been the topic of some debate (see Norde 2009: 160–172 for a summary), but this has no bearing on the methodological issues discussed in this paper.

(examples (13) and (14)) or relative clause (examples (15) and (16)). Note that in group genitives, the word to which the s-genitive is attached is invariably the final one, irrespective of word class. Thus the s-genitive is attached to the object form of a personal pronoun in (14), to an adverb in (15), and to a tensed verb in (16).

- (13) *Det är egentligen den unga mannen med glasögonens tur.*
It is actually [the young man-DEF with glasses]S turn

'Actually, it is the turn of the young man with the glasses'

[source: <http://aventyrligaeventyr.blogspot.com/2008/06/lrdag-216-kl2334.html>, accessed October 18, 2013.]

- (14) [...] *så ser man inte personen bakom migs ansikte*
[...] then sees one not [person-DEF behind me]S face

'then one does not see the face of the person behind me'

[source: www.fragbite.se › Forumindex › Övrigt, accessed October 18, 2013.]

- (15) *det var skitkabeln som följde meds fel*
it was [bloody.cable-DEF that followed with]S fault

'it was the fault of that bloody cable that came with it'

[source: <http://www.minhembio.com/forum/index.php?showtopic=125062&st=7300>, accessed October 18, 2013.]

- (16) *de som jobbars hälsa*
[they who work]S health

'the health of those who work'

[source: stafrin.blogspot.se/post/362/13613, accessed October 18, 2013.]

Since the s-genitive is monophonemic and not orthographically marked (unlike the English s-genitive which is separated from its host by an apostrophe⁷), finding examples of it is a prototypical needle-in-the-haystack task. The examples above were found in Google searches using specific words or strings of words that might be the final part of a group genitive construction. For example, many present tense forms followed by *s*, such as *jobbars*, are not homonymous with any other Swedish word form. Nevertheless, searching *jobbars* in Swedish web pages yields many false positives, i.e. spelling errors.⁸

In the absence of annotated corpora, empirical studies of the rise of the Swedish group genitives have not been carried out yet. Individual examples have been noted – the oldest

⁷ An apostrophe is only used (optionally) when the possessor noun ends in a sibilant (Teleman et al. 1999a: 112). In other cases, apostrophes may be found in informal writing (probably due to English influence), but this usage is not accepted by the Swedish Language Council (cf. http://www.sprakradet.se/GetDoc?meta_id=1950, accessed October 18, 2013).

⁸ Google offers the possibility of searching pages written in a specific language, but this does not work perfectly. In order to further exclude pages in other languages I usually add the Swedish word *och* 'and' in the search field. Most of the time, this works well, because this particular string of letters appears to be rare in other languages (even if it does occur, e.g. Dutch *och* 'ah, well', it is not very frequent), whereas it is the most frequent word in written Swedish (cf. https://svn.spraakdata.gu.se/sb-arkiv/pub/frekvens/stats_all.txt, accessed October 18, 2013).

example attested so far being *Swen i Kleffs tomt* (1452) 'Swen of Kleff's property' (Delsing 1991: 28). In a paper on the history of the Swedish group genitive (Norde 2013), I used the following method. Point of departure was the observation that group genitives involving lexicalized semantic units, such as *mannen på gatan* 'the man in the street', or *kungen av Preussen* 'the king of Prussia' (Thorell 1977: 49; Teleman et al. 1999b: 131) are the only [[NP][PP]] type of group genitive constructions that is accepted in normative grammars. Moreover, in online documents group genitives appear to be preferred when the possessor is such a semantic unit. For instance, Google found 42 instances⁹ of the group genitive *drottningen av Englands* 'the queen of England's', as in (17)a, and only eight instances of *drottningens av England* 'the queen's of England', as in (17)b.¹⁰

- (17) a. *drottning-en av England=s krona*
 [queen-DEF of England]=GEN crown
 'the queen of England's crown'
- b. *drottning-en~sav England vapen*
 queen-DEF~GEN of England coat of arms
 'the queen of England's coat of arms'

Coupled with Delsing's (1991) observation that a [[NP][PP]] construction was the oldest group genitive he was aware of, I decided to focus on this particular type. Using sources from the late Middle Swedish and Early Modern Swedish period (covering the years 1380–1758), I generated concordances for the two prepositions that were most commonly used in this group genitive construction, to wit *i* 'in' (spelled <i>, <j> or <ij>), and *af* (<af>, <aff> or <av>), and excerpted all relevant constructions manually. These were complex NPs, consisting of a noun denoting some noble title (e.g. 'king'), optionally followed by a personal name, which forms a semantic unit with a PP consisting of a preposition plus a geographic name (e.g. 'of Denmark'). Some of these examples were group genitives, but this particular search method produced other types of [[NP][PP]] genitive constructions as well. These are exemplified in (18). In the abstract schemas, [NP] is the noble title, optionally followed by a personal name; [PP] is the prepositional phrase that modifies [NP]; X is the possessee, the head which [[NP][PP]]_{gen} is attributive of; and subscript _{gen} marks the position of the genitive marker(s).

- (18) a. *biscop Bryniolff~z fadher i Skara* [1530]
 [bishop Brynjolf]~GEN father in Skara
 'the father of bishop Brynjolf of Skara'
 [[NP]_{gen}X[PP]]
- b. *konung-en~s i Poland skip* [1640]
 king-DEF~GEN in Poland ships
 'the king of Poland's ships'
 [[NP]_{gen}[PP]X]

⁹ These are the results of a search performed on December 20th, 2009.

¹⁰ The notations <=s> and <=GEN> indicate that the genitive is enclitic; the notations <~s> and <~GEN> indicate that the genitive is a phrase marker. See Norde (2006) for discussion of these variable degrees of attachment.

- c. *konung-en~s i Pålånd~z skipp* [1640]
king-DEF~GEN in Poland~GEN ships
'the king of Poland's ships'
[[NP]_{gen}[PP]_{gen}X]
- d. *konungen i Danmarck=s krigzfolck* [1585]
[king-DEF in Denmark]=GEN forces
'the king of Denmark's armed forces'
[[[NP][PP]]_{gen}X]

It turned out that even these constructions were extremely rare in the period under consideration. As is shown in table 2, only 81 relevant constructions were attested in a corpus of 1,228,148 words. Excerpting these manually would have been extremely time-consuming, obviously, but the method outlined above has the disadvantage of finding two particular construction types only. Undoubtedly, there are other group genitive constructions out there, but since their particular form is unknown, they will remain unnoticed unless one reads one's way through all texts.

Total number of words in corpus	1,228,148
Tokens of preposition <i>i</i>	24,488
Genitive constructions involving <i>i</i>	22
Tokens of preposition <i>af</i>	11,797
Genitive constructions involving <i>af</i>	59
Total number of relevant constructions	81

Table 2: Genitive constructions (Norde 2013)

Furthermore, figure 2 shows that the number of occurrences is too small to draw firm conclusions about the chronological order of the four construction types. Type 1, [[NP]_{gen}X[PP]] (as in example (18)a), is clearly the oldest pattern, as it occurs in the oldest text in the corpus, dating from the end of the 14th century. The group genitive (type 4 as in example (18)d) is the youngest, and does not occur before 1585. But it occurs only five times in the entire corpus, in the works of three authors: Per Brahe (1585), Carl Gyllenhielm (1640), and Agneta Horn (1657). It is not attested in younger texts, unlike types 1–3, which predate the group genitive. Another striking observation is that most authors in the corpus use more than one construction, Per Brahe even uses all four of them.

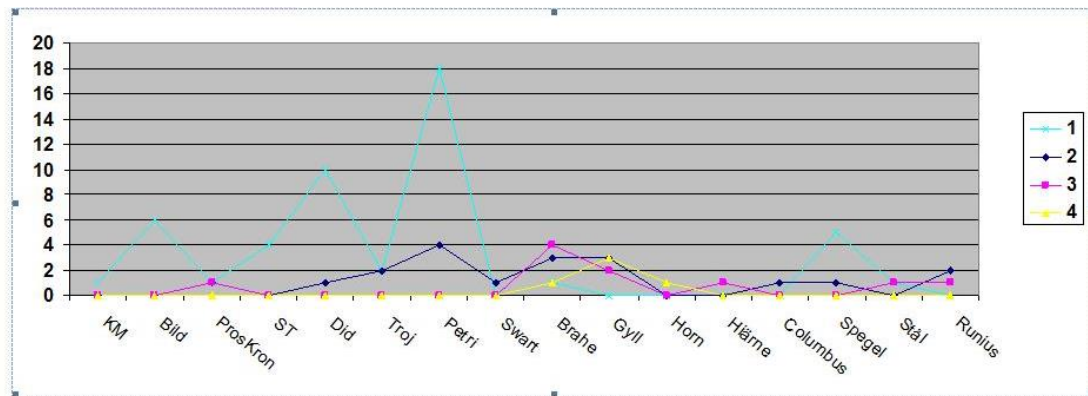


Figure 2: Relative frequency of construction types

Summing up thus far, it may seem as if this semi-automated method of collecting data does not generate much insight into the history of the s-genitive and competing genitive constructions. However, if we take a closer look, some interesting observations can be made nevertheless. First of all, most authors use more than one pattern. This strongly suggests that several constructions were in competition to succeed the original $[[NP]_{\text{gen}}X[PP]]$ pattern, and it is in accordance with the gradualness view on language change discussed in section 2.2. As authors may use both type 2 (as in example (18)b), where the genitive marker is best analysed as an (inflectional) phrase-marker, and 4, where it is more clitic-like, the morphological status of the genitive is clearly gradient. A second observation is that type 3, $[[NP]_{\text{gen}}[PP]_{\text{gen}}X]$ (as in example (18)c), mostly occurs in texts where at least one other genitive construction is being used. The most plausible explanation for this is that type 3 is a contamination of types 2 and 4. Authors who use both type 2 and type 4 use type 3 as well. Thus the very existence of type 3 reinforces the view that type 2 and type 4 were contemporaneous (for more extensive discussion of this point see Norde 2013).

To conclude this section, what I hope to have demonstrated with this case study is that needles in the haystack are definitely worth looking for. In spite of their relative infrequency, they may yield important information on complex changes such as the rise of the Swedish group genitive.

5 Conclusions

This paper started out with some notorious problems that historical linguists find themselves confronted with. Some are related to the sources themselves – the material available today is the result of "accidents of history", and native-speaker judgments are obviously not available. For these and other reasons, historical corpora do not really lend themselves to large-scale quantitative investigations. However, this is not necessarily a bad thing. The qualitative method illustrated in this paper has several advantages: it enables much more fine-grained analyses and may reveal the delicate interplay between changes at different levels (phonology, morphology, syntax, semantics, pragmatics). Finally, detailed qualitative analyses may disclose data that turn out to be crucial to a correct understanding of the changes involved: quirky quotes because they force the researcher to consider alternative pathways of

change, and needles in the haystack because they might just be the missing link. In my view, both are indispensable in historical corpus linguistics.

References

- Andersen, Henning (2001): "Actualization and the (uni)directionality of change". In: Andersen, Henning (ed.) (2001): *Actualization. Linguistic change in progress*. Amsterdam/Philadelphia, John Benjamins: 225–248.
- Andréasson, Maia (2002): *Kanske – en vilde i satsschemat*. Göteborg: Institutionen för Svenska Språket. (= *MISS* 41).
- Bauer, Brigitte L.M. (2003): "The adverbial formation in *mente* in vulgar and late Latin. A problem in grammaticalization". In: Solin, Heikki/Leiwo, Martti/Halla-aho, Hilla (eds.) (2003): *Latin vulgare – latin tardif VI*. Hildesheim/Zürich/New York, Olms-Weidmann: 439–457.
- Beijering, Karin (2010): "The grammaticalization of Mainland Scandinavian MAYBE". In: Bugge, Edit/Hareide, Lidun (eds.) (2010): *Seven Mountains; Seven Voices*. Bergen: University of Bergen. (= *Bergen Language and Linguistics Studies [BeLLS]* 1).
- Campbell, Lyle (1991): "Some grammaticalization changes in Estonian and their implications". In: Traugott, Elizabeth Closs/Heine, Bernd (eds.) (1991): *Approaches to grammaticalization I*. Amsterdam/Philadelphia, John Benjamins: 285–299.
- Delsing, Lars-Olof (1991): "Om genitivens utveckling i fornsvenskan". In: Malmgren, Sven Göran/Ralph, Bo (eds.) (1991): *Studier i svensk språkhistoria 2*. Göteborg, Institutionen för Nordiska Språk: 12–30.
- Detges, Ulrich (1998): "Echt die Wahrheit sagen. Überlegungen zur Grammatikalisierung von Adverbmarkern". *Philologie im Netz* 4: 1–29. <http://web.fu-berlin.de/phin/phin4/p4t1.htm>, accessed October 18, 2013.
- Evans, Nicholas (2007): "Insubordination and its uses". In: Nikolaeva, Irina (ed.) (2007): *Finiteness. Theoretical and empirical foundations*. Oxford, Oxford University Press: 366–431.
- Fiva, Toril (1987): *Possessor chains in Norwegian*. Oslo: Novus Forlag. (= *Tromsø studier i språkvitenskap* 9).
- Giacalone Ramat, Anna (1998): "Testing the boundaries of grammaticalization". In: Giacalone Ramat, Anna/Hopper, Paul J. (eds.) (1998): *The limits of grammaticalization*. Amsterdam/Philadelphia, John Benjamins: 107–127.
- Haspelmath, Martin (2004): "On directionality in language change with particular reference to grammaticalization". In: Fischer, Olga/Norde, Muriel/Perridon, Harry (eds.) (2004): *Up and down the Cline – The Nature of Grammaticalization*. Amsterdam/Philadelphia, John Benjamins: 17–44.
- Hopper, Paul. J./Traugott, Elizabeth Closs (2003): *Grammaticalization*. Cambridge: Cambridge University Press.
- Hummel, Martin (2000): *Adverbale und adverbialisierte Adjektive im Spanischen. Konstruktionen des Typs Los niños duermen tranquilos und María corre rápido*. Tübingen: Gunter Narr Verlag.
- Janda, Richard D./Joseph, Brian D. (2003): "On language, change, and language change – Or, of history, linguistics, and historical linguistics". In: Joseph, Brian D./Janda, Richard D. (eds.) (2003): *The handbook of historical linguistics*. Oxford, Blackwell: 3–180.

- Karlsson, Keith E. (1981): *Syntax and affixation. The evolution of MENTE in Latin and Romance*. Tübingen: Max Niemeyer Verlag.
- Lass, Roger (1997): *Historical linguistics and language change*. Cambridge: Cambridge University Press.
- Lehmann, Christian (1995 [1982]): *Thoughts on grammaticalization*. München/Newcastle: Lincom Europa.
- Lehmann, Christian (2004): "Theory and method in grammaticalization". *Zeitschrift für Germanistische Linguistik* 32/2: 152–187.
- Lødrup, Helge (1989): *Norske hypotagmer. En LFG-beskrivelse av ikke-verbale hypotagmer*. Oslo: Novus forlag. (= *Oslo-studier i språkvitenskap* 4).
- Norde, Muriel (1997): *The history of the genitive in Swedish. A case study in degrammaticalization*. PhD thesis. Amsterdam: University of Amsterdam.
- Norde, Muriel (2006): "Demarcating degrammaticalization: the Swedish s-genitive revisited". *Nordic Journal of Linguistics* 29/2: 201–238.
- Norde, Muriel (2009): *Degrammaticalization*. Oxford: Oxford University Press.
- Norde, Muriel (2012): "On the origin(s) of the possessor doubling construction in Norwegian". In: Van der Liet, Henk/Norde, Muriel (eds.) (2012): *Language for its own sake. Essays on Language and Literature offered to Harry Perridon*. Amsterdam, Scandinavisch Instituut: 327–358.
- Norde, Muriel (2013): "Tracing the origins of the Swedish group genitive". In: Carlier, Anne/Verstraete, Jean-Christophe (eds.) (2013): *The genitive*. Amsterdam/Philadelphia, John Benjamins: 299–332. (= *Case and Grammatical Relations Across Languages* 5).
- Norde, Muriel/Rawoens, Gudrun/Beijering, Karin (in prep.): "Från matrissats till satsadverb. En diakron studie av adverbet *kanske*".
- Palm, Rune (2004): *Vikingarnas språk 750–1100*. Stockholm: Norstedts.
- Ramat, Paolo (1992): "Thoughts on degrammaticalization". *Linguistics* 30: 549–560.
- Ramat, Paolo (2011): "Adverbial grammaticalization". In: Narrog, Heiko/Heine, Bernd (eds.) (2011): *The Oxford handbook of grammaticalization*. Oxford, Oxford University Press: 502–510.
- Ramat, Paolo/Ricca, Davide (1998): "Sentence adverbs in the languages of Europe". In: Van der Auwera, Johan/Ó Baoill, Dónall P. (eds.) (1998): *Adverbial constructions in the languages of Europe*. Berlin/New York, Mouton de Gruyter: 187–273.
- Scott, Mike (2004). *WordSmith Tools version 4*. Oxford: Oxford University Press.
- Teleman, Ulf/Hellberg, Staffan/Andersson, Erik (1999a): *Svenska Akademiens Grammatik II: Ord*. Stockholm: Norstedts.
- Teleman, Ulf/Hellberg/Staffan/Andersson, Erik (1999b): *Svenska Akademiens Grammatik III: Fraser*. Stockholm: Norstedts.
- Thorell, Olof (1977): *Svensk grammatik. Andra upplagan*. Stockholm: Esselte Studium.
- Torner, Sergi (2005): "Spanish adverbs in *-mente*". *Probus* 17/1: 115–144.
- Traugott, Elizabeth Closs (1989): "On the rise of epistemic meanings in English: an example of subjectification in semantic change". *Language* 65/1: 31–55.
- Traugott, Elizabeth Closs/Trousdale, Graeme (eds.) (2010): *Gradience, gradualness and grammaticalization*. Amsterdam/Philadelphia: John Benjamins.
- Trosterud, Trond (2001): "The changes in Scandinavian morphology from 1100 to 1500". *Arkiv för Nordisk Filologi* 116: 153–191.

Wessén, Elias (1967): "Ett fornsvenskt vardagsord Fsv. *maxan* – da. *måske* – sv. *kanske*". *Nysvenska Studier* 47: 5–16.