Historical and comparative perspectives on the subjectification of causal connectives.

Jacqueline Evers-Vermeul, Liesbeth Degand, Benjamin Fagard, Liesbeth Mortier

Historical and comparative perspectives on the subjectification of causal connectives
Jacqueline Evers-Vermeul, Liesbeth Degand, Benjamin Fagard, & Liesbeth Mortier
5 maart 2009

Jacqueline Evers-Vermeul
Utrecht Institute of Linguistics OTS
Utrecht University
Trans 10
NL-3512 JK Utrecht, The Netherlands
J.Evers@uu.nl

Liesbeth Degand
Institute for Language and Communication (IL&C)
Université catholique de Louvain
Place B. Pascal, 1
B-1348 Louvain-la-Neuve, Belgium
liesbeth.degand@uclouvain.be

Benjamin Fagard
Laboratoire Lattice
CNRS – ENS & Université Paris 3
1 rue Maurice Arnoux
F-92120 Montrouge, France
benjamin.fagard@ens.fr

Liesbeth Mortier
Institute for Language and Communication (IL&C)
Université catholique de Louvain
Place B. Pascal, 1
B-1348 Louvain-la-Neuve, Belgium
liesbeth.mortier@uclouvain.be

Acknowledgements
This research is supported by IUAP-grant P6/44 Grammaticalization and (Inter)Subjectification financed by the Belgian Federal Government. The second author is research associate at the Belgian Science Foundation FRS-FNRS.

Historical and comparative perspectives on the subjectification of causal connectives

Abstract
In this paper, we focus on the diachronic development of causal connectives and investigate whether subjectification occurs. We present the results of ongoing and previous analyses of the diachronic development of Dutch *want* and *omdat*, and French *car* and *parce que*, all four causal connectives roughly meaning 'because'. In addition, we try to show that "grammaticalization studies can gain from the systematic and principled use of large computerized corpora and the methods which have been developed within corpus linguistics" (Lindquist & Mair 2004: x). That's why we have combined two historical and two comparative corpus methods to chart the diachronic development of these four causals. Our study reveals that subjectification is not an integral part of the diachronic development of these causals: subjectification does occur in the rise of these connectives, but in the later stages of their development only *parce que* undergoes subjectification. Our analyses show that the four methods all have their own merits and limitations, but they are most effective when combined.

1. Introduction
Research on grammaticalization has established itself as a major area in linguistic studies. A noteworthy development in the past decade is the growing interest of grammaticalization theorists in the use of corpora, and hence in the techniques developed in the field of corpus linguistics (cf. Rissanen, Kytö & Heikkonen 1997; Lindquist & Mair 2004). Mair (2004) points out several commonalities between the two fields, and argues in favor of a closer collaboration between corpus linguists and grammaticalization specialists. In our paper, we follow Mair's recommendation and confront current ideas within grammaticalization theory with the results of various corpus studies. More specifically, we will show the advantages of taking a corpus-based approach to the study of causal connectives, a subclass of discourse markers. "Discourse markers are ideal for observing variation and change: they originate in different grammatical categories, they often compete with many other forms, and they are sensitive to trends regarding language use" (cf. Vincent 2005: 191).

The diachronic development of discourse markers often involves a process of '(inter)subjectification' (cf. Traugott & Dasher 2002; Athanasiado, Canakis & Cornillie 2006), a shift from meanings pertaining to the characterization of the objective world first to meanings involving the expression of personal attitudes of the speaker (subjectification) and then to meanings linked to speaker-hearer interactions (i.e., intersubjectivity). A famous example in the area of connectives concerns the diachronic development of English *while* (cf. Traugott 1995; González-Cruz 2007). The first step in the direction of its subjectification is the change of the adverbial phrase *þa hwile þe* ('at the time that') into the temporal connective *while*. Instead of profiling a specific time in the real world, the connective *while* profiles the ordering of events within the discourse structure, an ordering provided by the speaker. Its second step of subjectification is the development of temporal *while* into concessive *while*. This new use construes a relation between events that has no reference in the described situation, but only in the speaker's belief about coherence.

Traugott (1995: 39) put forward a strong claim regarding subjectification in the area of connectives. According to her, "historically almost all grammatical markers of clause combining have developed out of a more 'objective' function" (see also Dasher 1995). And indeed, many temporal, causal and conditional connectives have grown out of adverbial constructions (Genetti 1991). Examples in various languages include the subjectification of German *weil* 'because' (Keller

1995; Günthner 1996), Japanese *na* elements (Onodera 2000, 2004), English *because* and its Japanese counterpart *kara* (Higashiizumi 2006), Dutch *dus* 'so' (cf. Evers-Vermeul 2005, Evers-Vermeul & Stukker 2003), French *car* and *parce que* 'because' (Degand & Fagard in press), to mention just a few of many connectives from a variety of languages.

With the growing availability of digital diachronic corpora, the number of diachronic analyses of connectives in terms of subjectification has shown a tremendous increase, thus providing evidence in favor of Traugott's (1995) claim. However, a few critical remarks can be made on the methodology of various studies in this area. A first point concerns studies in which claims about subjectification or grammaticalization are based solely on synchronic data. For instance, Günthner and Metz (2004) analyze the variation of *obwohl* 'although' and *wobei* 'whereby' in contemporary spoken German. They conclude on the basis of these synchronic data that *obwohl* and *wobei* "have developed discourse-pragmatic functions and have become, or are on their way to becoming discourse/pragmatic markers" (Günthner & Metz 2004: 98). Vincent (2005) is another case in point (see König & Van der Auwera (1988) and Erman & Kotsinas (1993) for more examples). Although she supports some of her claims about the diachrony of *par exemple* by reference to ancient dictionaries, she predominantly bases her claims on synchronic data.[1] In our view, subjectification studies as well as other studies with diachronic implications may of course take synchronic data as a starting point to build diachronic hypotheses, but can be validated only through its testing against diachronic data. For instance, synchronic variation between speech and writing (see section 6 below) can serve as an indication that a specific type of language evolution is developing and thus lead to a given hypothesis; but then, this hypothesis should be tested against diachronic data to validate it.

A second point is that studies which do meet this diachronic criterion and base themselves on authentic diachronic data have been predominantly qualitative in nature. Qualitative discourse studies typically take a small data set, a single text or a relatively small sample of texts, and examine it in depth. The majority of this type of research only provides anecdotic examples, and lacks quantitative underpinning (but see e.g. Prévost 1999, 2003, 2007, who systematically analyzes quantified data in the area of discourse markers). Although detailed qualitative analyses based on manual extraction are useful in themselves (cf. Traugott 1995 on *while*; Molencki 2007 on *since*), they can and should be fruitfully complemented by corpus-based methods. Stefanowitsch (2006: 12) formulates this urge for quantification in the area of research into metaphorical mappings: "many of the results are provisional, awaiting more stringent quantification and statistical evaluation." As Partington (2006: 268) puts it: "Complementing the qualitative with a more quantitative approach, as embodied in Corpus Linguistics, not only allows a greater distance to be preserved between observer and data but also enables a far greater amount of data to be contemplated. In addition, it can identify promising areas for qualitative forms of analysis to investigate." Moreover, corpora enable researchers to meet the criterion of total 'accountability' (cf. Johansson 1985: 208; Labov 1994: 550), which demands that a linguistic description should account for all the data in a body of texts, and not just for particular instances.

Given this characterization of previous studies on subjectification in the use of connectives, and given the new perspective corpus-based approaches seem to offer, the following two research questions can be formulated: 1) Does subjectification occur in the diachronic development of causal connectives? and 2) What does a corpus-based approach add to the study of this diachronic

[1] This is very striking, since the sociolinguistic interviews she analyzes stem from three different years: 1971, 1984, 1995, and therefore would have allowed for a minimally/short-term diachronic analysis according to the real time method. Unfortunately, however, Vincent (2005) has chosen to group all the data together.

development? In the remainder of this paper, we will combine two historical and two comparative corpus methods to chart the diachronic development of four causal connectives. In section 2 we give the rationale behind this method of using converging evidence. In sections 3 to 6 we present the results of our four corpus studies. Not all four of our studies give rise to in-depth analyses here, our focus being on the type of results each of the four studies brings about, and on the methodological need to combine them. In section 7, we will answer our research questions and put forward some points for discussion.

2. Method

We focus on the diachronic development of Dutch and French causal connectives and investigate whether subjectification occurs during these developments. To this end, we present the results of ongoing and previous analyses of the diachronic development of Dutch *want* and *omdat*, and French *car* and *parce que*, all four causal connectives roughly meaning 'because'.

The initial premise of this article is that the techniques of corpus linguistics can assist the diachronic study of connectives. They can help reveal recurrent patterns of connective usage which reflect the systematic behavior and attitudes of the users. The choice of the corpus, as Hoffmann (2004: 197) points out, is fundamental: the "reliability and meaningfulness of empirical data is heavily dependent on the assumption that language corpora constitute suitable mirrors of actual language use, either in its totality or at least in a wider functional domain. The choices made by the compilers of a corpus with respect to the selection and proportional representation of different text domains consequently have direct influence on the relevance of the linguistic results." This is why we will analyze the degree of subjectivity of causal connectives in four different corpora (see Table 1). Confronting the results of these different corpora applied to the same set of connectives should help us determine the contribution of each of the corpora when trying to trace subjectification processes.

Table 1. Four corpora used in our subjectification research

	Same texts	Different texts
Same period	Corpus 1 Present-day translation corpora	Corpus 4 Present-day spoken and written language
Different periods	Corpus 2 Bible translations from different periods	Corpus 3 Various written texts from different periods

Corpus 1 is a so-called parallel corpus, or more specifically, a translation corpus.[2] It is a collection of present-day original texts and their translations in another language. In our case, Corpus 1 is compiled of original Dutch texts and their French translations, and of original French texts and their translations into Dutch (see section 3 for more details). Corpus 2 comprises translated texts from different periods; in our study we selected Dutch Bible translations from four periods. Corpus 3 consists of a variety of original texts, both from present-day and from ancient written sources. Ideally, this is a comparable corpus, in the sense that it contains texts matched by such criteria as domain, genre, intended audience, etc. (cf. Johansson 1998: 5). This is the type of corpus that is most commonly used in diachronic research. Corpus 4 is only compiled of present-day data; it contains one subcorpus of written data and another of spoken data, thus enabling a comparison of

[2] See Noël (2003: 278) for some remarks on this terminology.

the two modalities. In sections 3 to 6, we will discuss for each corpus a) its potential advantages and disadvantages, and b) the results it leads to in terms of the subjectification of causal connectives.

In order to measure the degree of subjectivity of each connective fragment in a reliable and quantifiable way, we took Sweetser's (1990) domains of use as an analytical instrument (cf. Pander Maat & Degand 2001; Pander Maat & Sanders 2001; Evers-Vermeul & Stukker 2003; Pit 2003). We therefore distinguish between *content, epistemic,* and *speech act* relations, as illustrated by the constructed examples in (1)-(4) below.

(1) Non-volitional content
The temperature rose quickly because the sun was shining.
(2) Volitional content
We went out in the garden because the sun was shining.
(3) Epistemic
The temperature is probably going to rise, because the sun is shining.
(4) Speech act
Let's have dinner in the garden, because the sun is shining.

If subjectivity "implies some degree of integration of the perceiver in the description of an object or a process" (Cuenca 1997: 5), then it can be argued that the different domains can be used as a way to measure the degree of subjectivity of causal (and other) connective fragments. In a content relation, the speaker provides a description of facts that can be established objectively in reality. Content relations like the ones in (1) and (2) describe relations that can be objectively observed in the real world. Characteristic of non-volitional relations is that the causal relation occurs without human intervention. Hence, this relation type is more objective than the volitional kind of content relations, in which human activity is involved. More subjective are epistemic relations (see (3)) in which the consequence is not a state of affairs in reality, but a mental state of the protagonist. The causal relation as a whole – which often involves argumentation – is not objectively observable and the speaker who presents the relation has to adopt the perspective of the protagonist in order to interpret the sentence. Maximally subjective are speech-act relations like (4), since they do not concern a reality outside the speech event, but the structure of the ongoing discourse (cf. Pander Maat & Degand 2001: 216-228). The relation type of each connective fragment can be established using a paraphrase test (cf. Evers-Vermeul 2005 for a detailed discussion of this test).

3. Results Corpus 1: analysis of present-day translation corpora

Our aim in using Corpus 1 for diachronic research is twofold.[3] Our first purpose is not to track actual diachronic changes, but to gain insight into the precise meaning of the linguistic items under study. Translation corpora contain texts that are intended to express the same meanings and have identical or at least very similar discourse functions in the relevant languages. Successively using the source and target language as a starting-point, we can establish paradigms of correspondences: the translations can be arranged as a paradigm where each target item corresponds to a different meaning of the source item. The use of translation corpora for this purpose is relatively new.[4] Traditionally, linguists have asked native informants to make judgments about meanings. Native speakers can distinguish different uses of the same polysemous item and say how they are related. However, at times too many uses may be distinguished, or too few (cf. Aijmer 2004: 58; Bybee,

[3] Cf. Aijmer & Altenberg (1996: 12) for additional advantages of translation corpora.
[4] See Noël (2003) for an overview.

Perkins & Pagliuca 1994: 44). Translations are more reliable as sources of meanings and uses than native informants, because they are produced by trained translators without any theoretical concern in mind. Dyvik (1998, 2004) was one of the first to argue in favor of the use of translation corpora to establish the precise semantics of words. According to her, "the activity of translation is one of the very few cases where speakers evaluate meaning relations between expressions without doing so as part of some kind of metalinguistic, philosophical or theoretical reflection, but as a normal kind of linguistic activity. This inspires confidence in the intersubjectivity of such evaluations" (Dyvik 1998: 51). As such, the advantages of a translator-based approach to semantics and pragmatics are clear: "by taking the translator's profile as a starting point, one is likely to acquire some information on the original propositional content of the message and on the potentially accompanying pragmatic implicatures" (Mortier 2007: 144). This type of analysis can thus also be used to place linguistic alternatives relative to each other on a subjectivity scale, specifying the semantic profile of closely related connectives.

Our second purpose in using translation corpora for diachronic research is that they reveal alternative markers expressing similar meanings in a specific genre. By performing back and forth translations, resulting in what we will call a *mirror analysis*, it becomes possible not only to track the most important synchronic translation equivalents, but also to reveal a field of competing markers for comparable meanings in one language (cf. Dyvik 2004; Lewis 2005; Mortier 2007). This is useful for subsequent diachronic analyses of these linguistic competitors. Although we acknowledge that diachronically there is no "need to see a new or alternative marker as contingent on the loss or dysfunction of another marker (cf. Bybee et al. 1994: 21)" (Aijmer 2004: 70), we do believe that changes in the system as a whole may have repercussions on the use of specific linguistic items. For example, it might be the case that certain causal connectives take over the function of their competitors or that the competitors take over one or more functions of the connective under investigation.

There are also some disadvantages associated with translation corpora. First of all, translations only provide synchronic insights and do not reveal diachronic changes. Second, although there is a growing body of translation corpora, especially for translation from and to the English language, the availability of translation corpora is still restricted. Also, as far as genre is concerned, the range of translated texts is restricted as compared with the range of original texts (cf. Johansson 1998: 4). This may have repercussions on the generalizability of the conclusions drawn from these data. A third disadvantage is that the data may be infected by *translationese* (cf. Gellerstam 1996), i.e. translation-based deviations from target language conventions. Translated texts may differ from original texts because of source language influence. Finally, is it well-known that linguistic choices often depend on the individual translator's particular style and skill, and that there may be outright mistakes in translations. It is conceivable that "somewhere along the interpretation process, a mismatch occurs between the speaker's intentions and the hearer's interpretation" (Mortier 2007: 144). This, however, should not prevent us from using translations as linguistic evidence, because "it is exactly the translators' performance, not so much as good translators but as language users, which is of interest" and because "consistency in syntactic and lexical discrepancies between source and target texts is precisely what this kind of evidence hinges on" (Noël 2003: 779-780).

Our present-day translation corpus consists of a corpus of original Dutch texts and their translations into French, and of original French texts and their translations into Dutch. The size of Corpus 1 is approximately 550,000 words. Two main types of text are represented: fiction (literature) and non-fiction (newspaper texts). We selected all occurrences of the causal connectives

want, *omdat*, *car* and *parce que* and their translation. Table 2 and Table 3 present the resulting lists of markers that were used as translations, as well as their respective frequencies.

Table 2. Frequencies of Dutch translations of *car* and *parce que*

Dutch translation	Translations of *car* (N = 45)	Translations of *parce que* (N = 53)
Want	25 (55.6%)	5 (9.4%)
Omdat	12 (26.7%)	41 (77.4%)
Doordat	-	1 (1.9%)
Aangezien	3 (6.7%)	-
No translation / punctuation	4 (8.9%)	5 (9.4%)
Reformulation	1 (2.2%)	1 (1.9%)

As Table 2 shows, *want* and *omdat* are the most common translations of *car* and *parce que* (cf. examples (5) and (6)). Other linguistic alternatives (such as *doordat*, *aangezien*, and *immers*) are very infrequent, and are outnumbered by fragments in which the connective is replaced by punctuation (as in (7)) or is deleted altogether.

(5) F: Je lui ai dit que j'avais adopté sans peine la tactique du sourire **car** je suis convaincue qu'en effet cette histoire ne compte pas tant pour Maurice. (S. de Beauvoir, *La Femme Rompue*)
D: Ik heb haar verteld dat het me weinig moeite kostte om op de tactiek van de glimlach over te gaan, **want** ik ben ervan overtuigd dat deze affaire inderdaad niet zo veel voor Maurice betekent.
E: 'I have told her that it didn't take me much effort to proceed to the tactics of the smile, car/want/because I am convinced that this affaire does not mean much for Maurice.'

(6) F: Je croyais aux couples, **parce que** je croyais au nôtre. (S. de Beauvoir, *La Femme Rompue*)
D: Ik geloofde in paren, **omdat** ik in onszelf als paar geloofde.
E: 'I believed in couples, parce que/omdat/because I believed in us (as a couple).'

(7) F: Kerry n'a pas oublié l'échec subi par Bill Clinton, en 1995, **parce que** sa réforme faisait trop appel au budget et à l'intervention de l'Etat fédéral. (*Le Monde - De Morgen*)
D: Kerry is de mislukking van Bill Clinton in 1995 niet vergeten; zijn hervorming kostte een te grote hap uit het budget en deed een beroep op de inmenging van de federale staat.
E: 'Kerry did not forget Bill Clinton's failure in 1995, parce que/;/(because) his reform took too large a slice from the budget and appealed to the interference of the federal state.'

Table 3. Frequencies of French translations of *want* and *omdat*

French translation	Translations of *want* (N = 127)	Translations of *omdat* (N = 153)
Car	76 (59.8%)	7 (4.6%)
Parce que	8 (6.3%)	79 (51.6%)
Comme	1 (0.8%)	17 (11.1%)
Puisque	3 (2.4%)	3 (2.0%)
No translation / punctuation	25 (19.7%)	15 (9.8%)
Gerund / (pour) infinitive	3 (2.4%)	12 (7.8%)
Reformulation	11 (8.7%)	20 (13.1%)

Table 3 indicates that *want* is most often translated by *car* (see (8)), and *omdat* by *parce que* (see (9)). In addition, *comme* is a frequent equivalent of Dutch *omdat* (cf. (11)). The causal connectives are frequently omitted in the translations, or translated by non-connective linguistic alternatives (e.g. syntactic alternatives such as the gerund, or the *pour*-infinitive, or lexical alternatives like *à cause du fait que* 'because of the fact that' or *en raison de* 'for the reason that').

(8) D: Gauw vrijlaten die fascist, **want** wij zijn geen fascisten, wij houden onze handen schoon.
F: Dépêchons-nous de le libérer, **car** nous ne sommes pas des fascistes, nous, nous gardons les mains propres.
E: 'Let us quickly release that fascist, want/car/because we are not fascists, we keep our hands clean.'

(9) D: Dat weet ik **omdat** je oom hier kort na de bevrijding is geweest.
F: Je le sais **parce que** ton oncle est venu ici juste après la libération.
E: 'I know that omdat/parce que/because your uncle has been here shortly after the liberation.'

(10) D: **Omdat** hij voorlopig toch ook in de weekends ergens heen moest, kocht hij nog een kleine boerderij in Gelderland (…)
F: **Comme** il lui fallait bien, pour le moment, aller quelque part en week-end, il acheta en Gueldre une petite ferme (…)
E: 'Omdat/Comme/Since he temporarily had to go places in the weekends, he bought a small farm in Ghelderland'

For all four connectives, we also translated the translation marker back into the original language. For example, in the case of Dutch *want*, we listed all the linguistic items that were used as translations of *want*, and translated these linguistic items back into Dutch to find out which Dutch counterparts could be regarded as the linguistic competitors of *want*.[5] From the results of this mirror analysis, we derived a semantic map, within which the different markers are organized according to their respective importance within the field. We used three criteria to determine this relative importance, see (11).

(11) Criteria to determine the importance of a linguistic markers within the semantic map
a. The overall frequency of the marker
b. The number of relations
c. The strength of relations

Criterion (11)a looks at the overall frequency of the marker: the primary markers have a high frequency in the corpus data (*omdat* occurs 4.3 times per 10,000 words, *want* 3.5, *car* 2.4, and *parce que* 2.8 per 10,000 words), when compared to alternatives such as *doordat* (0.02) or *puisque* (0.3). This also works the other way around: if markers occur extremely infrequently, they are probably less relevant for the semantic field, and hence are not placed at the core of the map.

Criterion (11)b takes into account the number of relations: the more relations a marker entertains with other (primary) markers, the more it is at the core of the semantic field, especially when these relations are bidirectional (e.g. when *parce que* is translated by *omdat* and *omdat* by *parce que* in a significant amount of cases). Again, this argumentation can also be inverted: if L2 translations from L1 markers are entered into a back and forth analysis and do not provide L1

[5] This was not possible for *pour*-infinitives and gerunds.

output with a meaning that is at least partially related to the original L1 markers, then they are less likely to be relevant for the semantic field. They probably belong to a different semantic field, they have a very general, non-specific meaning, or they are instances of lexical reformulations that are most likely the result of translator interference.

Criterion (11)c investigates the strength of relations: a high frequency of translation pairs suggests a strong correlation between markers which thus have comparable semantic 'weight' in the given field. For example, *parce que* is translated by *omdat* in 41 out of the 54 cases (77.4%), whereas its translation by *doordat* does not exceed 1.9% (1 case). A hierarchy of equivalents is thus established, with three main categories: primary, secondary, and tertiary equivalents.

Our analysis resulted in the following semantic map (see Figure 1). The white fields represent the primary equivalents, the core of the semantic field with the maximum of relations between the four constitutive connectives. The dotted fields represent the secondary equivalents associated with *parce que* and *omdat*; secondary, because they still have a two-way relationship with those markers. The grey fields represent the tertiary equivalents associated with *car* and *want*, on the one hand, and with *omdat* and *want*, on the other hand. They are tertiary because of their one-way relation with these causals.

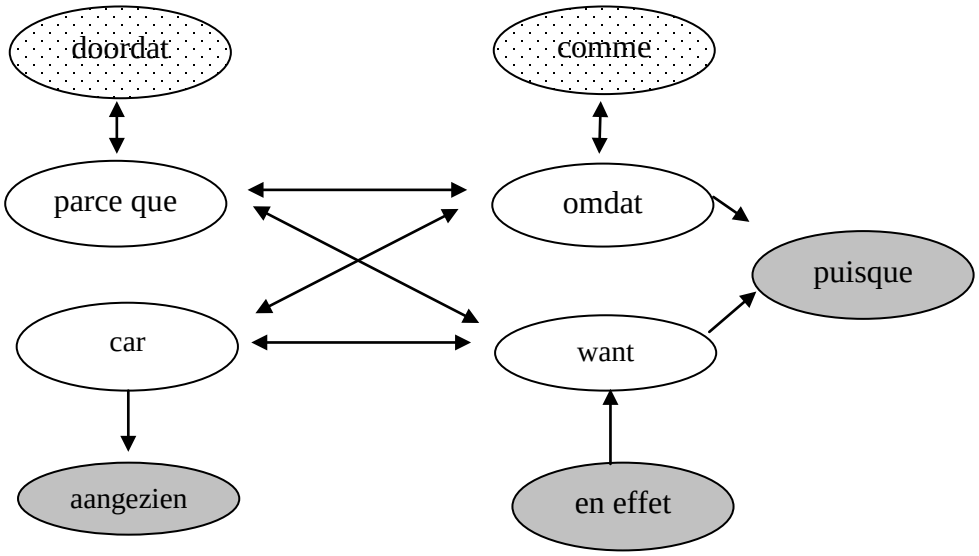

Figure 1. Semantic map of *want, omdat, car* and *parce que*

From Figure 1 it appears that *comme* and *en effet* are serious competitors for the two French causals. For Dutch, *doordat* and *aangezien* are the most important linguistic alternatives to be looked at when starting diachronic analyses.

An additional analysis we could have performed on the translation data in Corpus 1 concerns an analysis in terms of subjectivity. By establishing the domain of each source fragment, we could find out whether connective use in specific domains results in different translations. Prior research on the "translation pair" *puisque* and *aangezien* (Degand 2004) has indeed shown that translators tend to respect the level of subjectivity expressed by the connective fragments. It would thus be conceivable that e.g. non-volitional *doordat* would be translated into *parce que*, but not into *car*. Similarly, it is likely (given previous synchronic analyses of *want* and *omdat*, cf. Pit 2003, among others) that the French *car*-fragments that are translated into *omdat* are more objective than the fragments translated with *want*. Such a subjectivity analysis could result in a subjectivity scale

of the various linguistic markers. However, given limitations of time and space, we have not performed such a subjectivity analysis on the present data.

4. Results Corpus 3: analysis of various texts from different periods

We will now turn to Corpus 3, leaving Corpus 2 for discussion in the next section. Analyzing various texts from different periods is the most common way to investigate the historical development of linguistic phenomena. This allows researchers to perform both qualitative and quantitative analyses. An additional advantage is that an increasing number of digital diachronic data becomes available for research so that researchers can base their claims on actual diachronic data in context. Such information about contexts of use is lacking for diachronic data from dictionaries, which do not reveal frequency patterns either.

Disadvantages of Corpus 3 are that it restricts the researcher to the analysis of written data and that the results may suffer from a possible confounding with genre-effects: historical "developments" may be the result of studying different text types in different periods. For example, corpora with ancient texts often contain charts, moralistic texts and rhyming literature, whereas corpora with modern texts are often compiled of journalistic texts and non-rhyming novels. Ideally, Corpus 3 should be a comparable corpus, in the sense that it should contain texts matched by such criteria as domain, genre, intended audience, etc. (cf. Johansson 1998: 5). It is not always possible to compile such a comparable corpus, however, because of the restricted availability of ancient texts.

We analyzed *want*, *omdat*, *car* and *parce que* in various texts from different periods. We included various spellings, which is an important point for Old French and Middle Dutch in particular. Table 4 shows more details on our selection of periods, texts, and number of connective fragments. We selected arbitrarily 50 to 150 occurrences per period for each marker, and then registered formal and functional aspects of each occurrence. On the formal side, we looked at the categorical status of the connective, and at the positioning of the connective clause as a whole, distinguishing between pre- and postpositioning. On the functional side, we analyzed the relation type (causal, temporal, concessive or other) and the domain type (non-volitional content, volitional content, epistemic, and speech act).[6]

Table 4. Corpus 3: data on the Dutch and French diachronic corpora containing various texts

	Dutch	French
Periods	13th, 16th and 20th century	Old French (OF, 11th-13th), Middle French (MF, 13th-15th), Classical French (CF, 16th-17th), and Present-day French (PDF, 18th-20th century)
Data	Equal distribution of rhyming, literary texts and non-rhyming, non-literary texts, mostly from CD-roms (*Middelnederlands* 'Middle Dutch' and *Klassieke literatuur* 'Classical	Mostly literary texts, from the BFM Database (Base du Français Médiéval), the DMF Database (Dictionnaire du Moyen Français), and the Frantext database

[6] It should be noted that the Dutch and the French studies differ slightly in their operationalization of the epistemic relation: mental states such as 'being sad' were classified as epistemic in the French studies, but as non-volitional content in the Dutch studies. In the Dutch studies, the epistemic category only includes argumentative relations.

	literature') and the Internet (e.g. the project *Laurens Jansz. Coster*)[7]	
Fragments	50 per connective per period	50-150 per connective per period
References	Evers-Vermeul (2005)	Degand & Fagard (in press), Fagard (2008), Fagard & Degand (2008)

On the basis of our French analyses (see for more details Degand & Fagard in press; Fagard 2008; Fagard & Degand 2008), we can state that the evolutions of *car* and *parce que* are clear cases of grammaticalization. The rise of the connective *car* shows a series of parameters associated with grammaticalization: phonological reduction and internal bonding (from Latin *qua re* 'for which/what reason' to Middle French *quar/quer*, to PDF *car*), and semantic bleaching (the original meaning of *res* 'object, cause' progressively fades to the point that the presence of the noun is completely hidden not only by phonetics (*re* > *r*) but also by semantics). In addition, *car* changes from a complex subordinating conjunction to a simple coordinating conjunction, which can be derived from the loss of *car*'s ability to occur in preposed connective clauses. The grammaticalization of *parce que* is shown by loss of variation, phonetic attrition and internal bonding (OF *par/por ce que* 'for this that' > MF *parce (…) que* 'because' > PDF *parce que/paske*), and semantic bleaching (*ce* 'this' is anaphoric in OF *por ce que*, but not in PDF *parce que*).

The corpus approach enabled us to quantify the evolution of *car* and *parce que*, as we can see in Figure 2. Our subjectification study of the French part of Corpus 3 reveals that *parce que* shows a clear subjectification line in diachronic data, while *car* remains relatively stable throughout the centuries (Degand & Fagard in press).

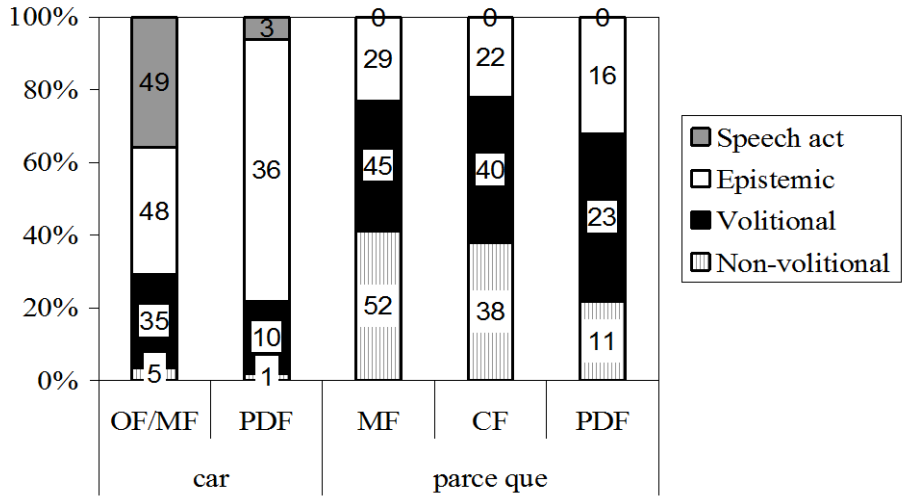

Figure 2. Distribution of *car* (left) and *parce que* over the domains of use in three periods[8]

For *car*, both speech act and epistemic uses are already present in Old French (cf. examples (12) and (13)). The speech act use of *car* decreases, but the amount of content and epistemic use remains relatively stable. *Parce que*, which remains a subordinator throughout the ages, is predominantly used in the content domain (see (14)). The development of this connective shows subjectification:

[7] See Chapter 4 in Evers-Vermeul 2005 for more details on the exact compilation of the Dutch corpus.
[8] For *car*, the Old & Middle French period were grouped together, because the date of several source texts could not be established exactly. *Parce que* does not occur before the Middle French period. The total of OF/MF *car* and *parce que* was 150; *car* showed non-connective usage in 13 cases, *parce que* in 24 instances.

there is a decrease of use in the objective non-volitional domain in favor of the more subjective volitional content and epistemic usage.

(12) Speech act connective use of *car* (12th century) (Le roman de Thèbes, c. 1150, v. 865)
Di va! fet il, pour coi me celes? **Car** assez set on les noveles, que Edyppus fist de son pere quant il l'ocist et prist sa mere.
'Tell me! he exclaimed, why do you keep it a secret? **For** we know very well what Oedipus did to his father, killing him and taking his own mother.'

(13) Epistemic connective use of *car* (12th century) (*Floire et Blancheflor*, 1150-1160, v. 2568)
Par foi, ci le cuidai trover, sire, **car** ains de moi leva.
By faith, here him think.past.1Sg find, sir, for before of me get-up.3Sg
'Truly, I thought I would find him here, sir, **because** he got up before me.'

(14) Content connective use of *parce que* (13th century) (*Roman de Renart*, early 13th c.)
Li anfes ploroit de grant fin **por ce que** n'avoit que mengier.
'The child cried of hunger, **because** he had nothing to eat.'

The rise of the Dutch connectives in Corpus 3 also reveals changes at the grammatical level (see Evers-Vermeul 2005 for a more detailed analysis). In the 13th century, *want* could be used both as a subordinator (see example (15)) and as a coordinator. Over time, *want* changed into a pure coordinator (as in (16)): it lost the ability to appear in preposed connective clauses or in clauses with a finite verb in final position (which is typical of Dutch subordinating clauses). *Omdat* shows internal bonding and reanalysis of the preposition *om* and the relativum *dat*; it became a fixed combination, the subordinator *omdat*, in the 16th century. Its positioning properties remain stable over time (cf. (17).

(15) Subordinating content use of *want* (*Historie van Malegijs*, 1556)
Doen dit ghedaen was, soo ghingen si ter maeltijt (**want**et byder noenen was)
'When this was done, they went to have a meal (because it was almost noon)'

(16) Coordinating epistemic use of *want* (*In de schaduw bloeien de rozen*, 1994)
Er moet hem een behoorlijk bedrag nagelaten zijn, **want** zijn ouders waren redelijk welgesteld.
'A substantial amount must have been left to him, because his parents were fairly well-off.'

(17) Subordinating content use of *omdat* (*Het verbroken zegel*, 1991)
Zij (= Célestine) *knikte hem met een fijn glimlachje toe, en gaf niet meer uitleg,* ***omdat*** *Sarah al teruggelopen was en naast haar stond.*
'She (= Célestine) nodded to him with a subtle smile, and did not explain anymore, because Sarah had returned already and stood next to her.'

At the semantic level, Corpus 3 shows a fairly stable profile for both *want* and *omdat*, not straightforwardly supporting the subjectification hypothesis of discourse markers. Figure 3 charts the distrubution of *want* and *omdat* over time. It reveals that *want* is mainly used as a marker of epistemic causal relations, whereas *omdat* mainly occurs in content relations. Statistical analysis indicates that the domains profile of *omdat* is stable across ages (χ^2 (6) = 10.1; p < .25). The only change in its use is that, after the 16th century, *omdat* has lost its ability to mark finalistic 'so that' causal relations. The connective *want* has hardly changed during the selected time span of 800 years; only in the 16th century was a significant increase in speech-act use found (χ^2 (1) = 12.2; p <

.001). This increase seems to point to subjectification. However, this subjectification was not a lasting phenomenon, since the number of speech-act fragments decreased again in the 20th century.

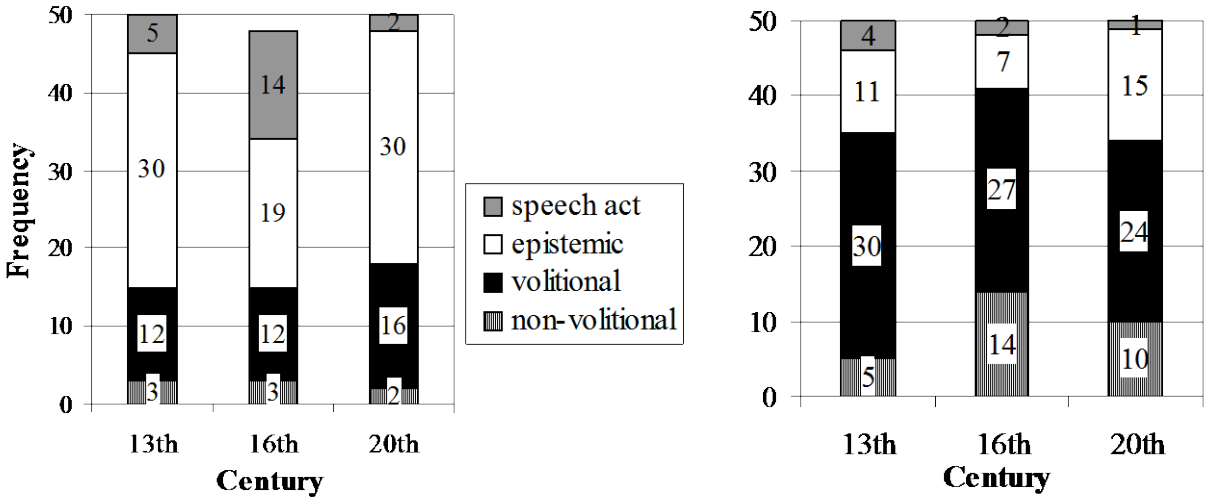

Figure 3. Distribution of *want* (left) and *omdat* over the domains of use in three periods

Our analysis of the speech-act fragments from the 16th century revealed that seven of the fourteen *want*-fragments came from the same moralistic source – *Devoot ende profitelyck boecxken* 'Devout and profitable book' – in the sample of rhyme texts. In this text from 1539, advice and orders like (18) are presented for a 'good life-style'. This advice and these orders are frequently supported with arguments, which results in the high number of speech-act relations. The temporary increase in speech acts, then, should not be seen as a case of subjectification, but as a genre-effect.

(18) God kent sijn schapen ende sij hooren na sijn stem
Nyemant en machse trecken wt sijnder hant / Sijn woert aenhoert
want God ghetuycht van hem
'God knows his sheep and they listen to his voice
No one can draw them from his hands / Listen to His (Jesus') word
because God testifies to him'

All in all, our study of Corpus 3 does show clear grammaticalization of three of the four connectives under investigation (*omdat, car, parce que*). This grammaticalization involves a transition from use of linguistic items at the lexical level to use of these items (albeit in a condensed form) at the discourse level. Once the items serve as connectives, subjectification is not a frequent phenomenon: only one of the four connectives – *parce que* – shows subjectification within the connective function.

5. Results Corpus 2: analysis of same texts in different periods

The previous section revealed that researchers who study a variety of texts in different periods run the risk of interpreting genre effects as a change of the linguistic phenomenon under investigation. Only texts from certain genres tend to be preserved from any historical period, making it difficult to separate out the effects of diachrony from the effects of genre (cf. Herring, Van Reenen & Schøsler 2000, and the references there). Of course there are other factors that affect the homogeneity of the

corpus. For example, if texts from different periods vary in register, dialect, and/or subject matter (cf. Biber, Conrad & Reppen 1998: 248), this may also result in a diachronic difference being unjustedly interpreted as a change of the linguistic phenemon under investigation. We will focus on genre effects here, because – as Condamines (2008:115) puts it: "Within corpus linguistics, genre is one of the most crucial but also one of the most difficult problems to be tackled."

There are at least two ways out of the problem of genre diversity. Gries (2006) proposes a rather sophisticated statistical solution, which enables researchers to calculate effect sizes of variability within and between corpora. For researchers who are less statistically equipped, a second solution may be more attractive: take genre into account during the analyses (cf. the suggestions in Condamines 2008). This is also useful when trying to identify regularities in the use of linguistic items which depend on the purpose of the text.

One way to take genre into account is to keep the genre constant throughout the ages and study a number of texts of the same genre or a limited set of genres in each period. Although this restricts the generalizability of the conclusions, it avoids the confounding problem by and large. A disadvantage of this method is that the characteristics of the genres themselves may change over time and that it introduces a new danger: confusing linguistic changes with genre changes (see, for example, Claridge & Wilson 2002 on the evolution of the genre 'sermons'). A second way of taking genre into account is to use different translations of the same source text. Because the same source text forms the basis of the translation in each period, the effect of changing genre conventions is diminished. That's why this approach enables the researcher to separate out the effects of diachrony from the effects of genre.

Our Corpus 2 contains the same text in different periods: four Dutch translations of the bible (see Schoonenboom 2000 and Vogl 2007 for a similar approach). Just like Corpus 3, this corpus may reveal actual semantic and/or structural diachronic changes. In addition, this corpus gives insight into the linguistic competitors of the connectives under investigation. Disadvantage of this method is that it concerns a very specific genre, namely religious texts, and that a translation effect may occur (cf. Gellerstam 1996). For example, Degand (2004) found that *aangezien* 'because' is more subjective in texts translated from French into Dutch than in original Dutch texts, because of transfer of the subjectivity profile of French *puisque*. Observations based on such a translation corpus, then, need to be checked against a control corpus consisting of comparable original texts in the same language.

The selection of bible translations has its own merits, which are less likely to be found when using ancient translations of other source texts. First, the presence of chapter and verse numbers facilitates a comparison of the translations from different periods. Second, the bible does not represent just one specific genre, but it contains a variety of text types, including stories, proverbs, songs, and argumentative texts. Third – although this certainly does not hold for all languages – for Dutch a relatively large amount of ancient bible translations is available (thanks to the efforts of Nicoline van der Sijs, see Van der Sijs 2008a, 2008b). Fourth, the translations are often constructed by a team of translators, rather than by one individual. This diminishes effects of personal preferences of individual authors. Finally, it is possible to compare the Dutch findings with translations in other languages.

The use of bible translations has some disadvantages as well. First, there may be effects of the religious nature of the texts: the language use may be more formal, or lag behind compared to the language used in secular documents. Second, it is not possible to cover the complete history of a language using only bible translations. Furthermore, it is hard to incorporate other factors causing

diversity in the corpus. For example, it would be very difficult to take dialect variation into account. All these disadvantages affect the generalizability of the conclusions.

Details on the Dutch corpus of bible translations can be found in Table 5. We focussed on the book of Genesis, which consists of 50 chapters and 1533 verses.

Table 5. Corpus 2: data of the Dutch corpus of bible translations

Translation	**Year of publication**	**Number of words in Genesis**
Delftse Bijbel (DB, see Van der Sijs 2008a)	1477	35,659
Statenvertaling (SV, see Van der Sijs 2008b)	1637	37,754
Translation of the Dutch Bible Association (NBG)	1951	36,482
New Bible Translation (NBV)	2004/2007	35,319

We analyzed all fragments that were marked with *want* or *omdat* in at least one of the translations (cf. Evers-Vermeul 2008). In total, 248 fragments were selected. For each fragment, we checked whether it was marked with *want*, *omdat*, or with some other or no marker. Table 6 shows the percentages of use of these markers in the selected fragments.

Table 6. Corpus 2: results on the diachronic corpus containing Dutch bible translations (N=248)

Translation	**Want (%)**	**Omdat (%)**	**Other or no marker (%)**
DB 1477	45	23	32
SV 1637	52	10	38
NBG 1951	41	11	48
NBV 2004	24	15	61

Statistical analysis reveals that the distribution over the connectives is not stable over time (χ^2 (6) = 28.4; $p < .001$).The percentage of *want*-fragments is relatively stable in the first three translations. Only in the translation of NBV 2004, there is a significant decrease in the number of *want*-fragments ($z = -2.6$).[9] This can be ascribed to a preference of the 2004-translators to leave the causal coherence relation implict, which also results in a significant increase in other uses in that translation ($z = 2.4$). For example, Genesis 3:19 contains *want* in the first three translations (see (19)a for the version of SV 1637), but it lacks a causal marker in the NBV 2004 (marked with **ø** in (19)b). Translations from other languages do insert a causal connective (German *denn*, English *for* and French *car*), suggesting the source text contains a causal marker as well.

(19) a. SV 1637 In ’t sweet uwes aenschijns sult ghy broot eten, tot dat ghy tot d’aerde wederkeert, dewijle ghy daer uyt genomen zijt:
want ghy zijt stof, ende ghy sult tot stof wederkeeren.
‘In the sweat of thy face shalt thou eat bread, till thou return unto the ground; **for** out of it wast thou taken: for dust thou [art], and unto dust shalt thou return.’ (King James 1611)

b. NBV 2004 Zweten zul je voor je brood, totdat je terugkeert tot de aarde, waaruit je bent genomen: **ø** stof ben je, tot stof keer je terug.
‘You will have to sweat for your bread, until you return to the ground, from which you were taken: **ø** dust you are, and to dust you will return.’

[9] A z-score is significant when $z < -1.96$ or $z > 1.96$.

The percentage of *omdat* drops after the first translation (in which it occurs more frequently than in the other three translations; z = 2.1), but remains stable after that. This drop is probably due to the disappearance of the finalistic use of *omdat*. 25 Fragments that would be marked with a *to*-infinitive in English, a *pour*-infinitive in French or (*auf*) *daß* or *zu*-infinitive in German, contain *omdat* in the translation of DB 1477, but *opdat* 'so that' or an *om*-infinitive in more recent translations (compare the a- and b-examples in (20) and (21)). In addition, the DB 1477 contains several *omdat*-clauses that are left out in the other translations.

(20) Genesis 21:30

a. DB 1477 Du sulste die seuen oyen ontfaen. van mijnre hant.
om dat si mi een oercontscap sullen sijn: dat ic desen put groef.

b. NBG 1951 Voorzeker moet gij de zeven lammeren uit mijn hand aannemen,
opdat het mij tot een getuigenis zij, dat ik deze put gegraven heb.
'These seven ewe lambs shalt thou take of my hand,
that it may be a witness unto me, that I have digged this well.'

(21) Genesis 15:7

a. DB 1477 Ic bin die here di v wtleide van hur van chaldea:
om dat ic di geuen soude dat lant ende dattu dat besitten souste

b. NBG 1951 Ik ben de HERE, die u uit Ur der Chaldeeën heb geleid
om u dit land in bezit te geven.
'I am the LORD that brought thee out of Ur of the Chaldees,
to give thee this land to inherit it.'

What do these data tell about changes in the degree of subjectivity of *want* and *omdat*? Firstly, they confirm previous findings that *want* is the more subjective of the two: 28 fragments that contain *want* in three of the four translations, are fragments that contain the more subjective markers in other languages: German *denn*, English *for*, and French *car*. Fragments containing *omdat* are often more objective and equal fragments with German *darum daß*, English *because*, and French *parce que*.

The previous section showed that the connective *want* lost the ability to be used as a subordinator. This finding is confirmed by this analysis (compare the three translations of Genesis 30:18 in (22)). However, this loss does not seem to affect the overall subjectivity profile of *want*, because the more objective subordinating use was far less frequent than the coordinating use of *want*.

(22) Genesis 30:18

a. DB 1477 God heuet mi desen loen gegeuen: **want** ic mijn ioncwijf minen man gaf
'God has given me my hire, **because** I gave my handmaid to my husband'

b. SV 1637 Godt heeft mijnen loon gegeven;**na dat** ick mijne dientmaecht mijnen man gegeven hebbe
'God has given me my hire, **after** I gave my handmaid to my husband'

c. NBV 2004 God heeft mij beloond **omdat** ik mijn slavin aan mijn man heb gegeven
'God has rewarded me, **because** I gave my bondwoman to my husband'

In the most recent translations, *want* appears to loose ground to markers such as *namelijk* 'namely', *immers* 'indeed', *tenslotte* 'after all', and *toch* – which would occur as a question tag in English. These markers indicate that certain information is already given, or accessible to the reader. For example, *want* in the SV 1637 translation of Genesis 4:25b in (23) shows up as *immers* 'indeed' in the NBG 1951, and as a relative clause in the NBV 2004. Because these fragments take the knowledge of the reader into account, they can be labeled intersubjective (in the sense of Traugott & Dasher 2002: 22). This suggests that *want* becomes more restricted in its more subjective use.

(23) Genesis 4:25b

a. SV 1637 Godt heeft my een ander zaet geset voor Habel; **want** Kaïn hem dootgeslagen heeft.
'God has appointed me another seed instead of Abel, **for** Cain slew him.

b. NBG 1951 God heeft mij een andere zoon gegeven in plaats van Abel; hem **immers** heeft Kaïn gedood.
'God has given me another son instead of Abel; **after all** he was killed by Cain.'

c. NBV 2004 God heeft mij in de plaats van Abel, **die** door Kaïn is gedood, een ander kind gegeven.'
'God has given me another child to take the place of Abel, who was killed by Cain.'

Modern *omdat* replaces other causal markers such as *dewijl, overmits (dat), naardien (dat),* and *daarin dat*, which are not used in modern Dutch anymore. Because these archaic connectives are comparable in terms of subjectivity, this does not affect the subjectivity profile of *omdat.*

Much more can be said about the results of Corpus 2. The analysis so far, however, already shows the usefulness of analyzing ancient bible translations for diachronic research. Because bible translations are very convenient for tracing linguistic competitors, Corpus 2 appears to be especially suitable for an onomasiological ('function-to-form') approach, which may supplement studies with a semasiological ('form-to-function') approach. In addition, Corpus 2 can be used as a control corpus for corpora with different texts from different periods, in order to be able to distinguish real changes from possible genre effects.

6. Results Corpus 4: comparison of present-day spoken and written language

The previous section showed the need for taking genre differences into account. This confirms the idea that the range of text categories (or *registers*) that samples are selected from is one of the two major issues in corpus linguistics.[10] Biber (1988, 1993, and many other works) has argued repeatedly that register variation is inherent to natural language and that diversified corpora representing a broad range of register variation are required as the basis for general language studies, especially for the external validity of the corpus study (i.e. the extent to which it is possible to generalize from a sample to a larger target population).

Although we acknowledge the need for using multigenre corpora, we will focus here on the importance of distinguishing between the two primary modalities of language: speech and writing.[11] First of all, we think this distinction is the most basic one of all the distinctions that can be made in order to incorporate register variation. Second, this distinction is relatively easy to operationalize, for synchronic and even for diachronic data. For example, Biber (1988) lists five dimensions to

[10] The second major issue is the size of the corpus.

[11] Of course we are aware of the fact that there are a number of 'in between' realizations of language, such as 'written-to-be-spoken', chat, sms, etc., but these finer distinctions do not matter to us here.

define similarities and differences among spoken and written registers. In addition, Koch & Oesterreicher (2001) provide criteria differentiating between the more "speech-like" data (e.g. drama, conversations in literature) and real "written" data (cf. also Chafe & Danielewicz 1987; Jacobs & Jucker 1995: 8; Traugott & Dasher 2002: 47). A third reason to focus on differences between the two modalities is that the comparison of present-day spoken and written data may reveal diachronic change "in progress". It is often claimed that changes first occur in spoken language and only gradually make their way into written texts (cf., among many others, Hansen & Rossari 2005: 181). Fourthly, if a researcher finds differences between oral and written language, that may invite the researcher to zoom into this finding and perform a subsequent diachronic analysis including a distinction between the two modalities (see e.g., Degand & Fagard submitted), or restricting the research to one of the two modalities. For example, Lindström & Wide (2005) study the diachronic development of Swedish discourse particles of the type *you know*, and restrict their study to historical texts that have at least some interactive properties, and therefore may reflect colloquial language use in which such markers can be expected.

Analyzing present-day data – whether speech-like or written – in order to gain insight into diachronic processes is not completely unproblematic. First, these data are not instances of actual diachronic data, and whether it is possible to recognize early grammaticalization from synchronic data is a point considered controversial in the literature. As Mair (2004: 131) points out, some researchers are extremely skeptical about the possibility of observing grammaticalization processes unfolding in the field (cf. Compes, Kutscher & Rudorf 1993: 20), whereas others seem to take a middle road. For example, Lehmann (1991: 532) writes in his study of ongoing change in present-day German: "Given presently available methodological means, it is next to impossible to know which of the changes that speech habits currently exhibit are synchronic manifestations of ongoing language change, and which of them are but ephemeral fashions."

A second problem concerns the nature of the "spoken" data. For example, the *Corpus Gesproken Nederlands* (CGN) does not only contain spontaneous conversations and interviews, but also prepared speeches and stories that are read out loud. This means again that researchers studying spoken data need to be careful in their collection of data (cf. Section 4). A final problem concerning the use of spoken data is that they are most often analyzed from a transcribed version, which introduces problems of its own (cf. Halliday 2004:15-21).

Our final corpus then, Corpus 4, is compiled of present-day French and Dutch data; for both languages, it contains one subcorpus of written data and another of spoken data, thus enabling a comparison of the two modalities. Table 7 introduces some relevant information about Corpus 4.

Table 7. Corpus 4: data on the Dutch and French present-day corpora

	Dutch	French
Written data	Journalistic: NRC Handelsblad 1994	Journalistic: Le Soir 1997
Spoken data	CGN: spontaneous conversations and interviews (see Oostdijk 2000)	Valibel: spontaneous conversations and interviews (for more details, see Francard, Geron & Wilmet 2002)
Number of fragments	Written: 50 *want*, 50 *omdat* Spoken: 149 *want*, 124 *omdat*	Written: 50 *car*, 50 *parce que* Spoken: 50 *car*, 50 *parce que*
References	Degand & Pander Maat (2003); Spooren, Sanders, Huiskes & Degand (in press)	Degand & Pander Maat (2003); Simon & Degand (2007)

Analysis of the Dutch data in Corpus 4 reveals that *want* and *omdat* have different distributions in the two modalities (cf. Spooren, Sanders, Huiskes & Degand (SSHD), in press for a more detailed analysis): *want* is more frequent in spoken than in written Dutch (1640 vs. 686 instances per million words), whereas *omdat* is more frequent in written than in spoken Dutch (938 vs. 521 instances per million words). Table 8 shows the results of the domain analysis of the connective fragments.

Table 8. Domain type expressed by *want* and *omdat* in written and spoken Dutch[12]

	Written		Spoken	
	Content	Epist./speech act	Content	Epist./speech act
omdat	26	24	106	11
want	8	42	61	82

This analysis confirms the finding of previous subjectivity analyses that *want* is more subjective than *omdat* (χ^2 (1) = 84.1; p < .001). *Omdat* occurs more often with content relations (132 of 167 or 79.1%), whereas *want* is more frequently found in non-content relations (124 of 193 or 64.2%). Both causals show a consistent semantic profile in writing and in speech; there is no three-way interaction between connective, medium and domain (χ^2 (1) = 1.9; p = .17). Hence, no subjectification tendencies can be found.

Analysis of the French data shows that *car* is more frequent in written than in spoken French (0.32% vs. 0.02%), whereas *parce que* is more frequent in spoken than in written French (3.70% vs. 0.40%). In written French, the distribution of the two connectives is similar, but in spoken French, *parce que* is 185 times more frequent than *car*. The results of the domain analyses are given in Figure 4.

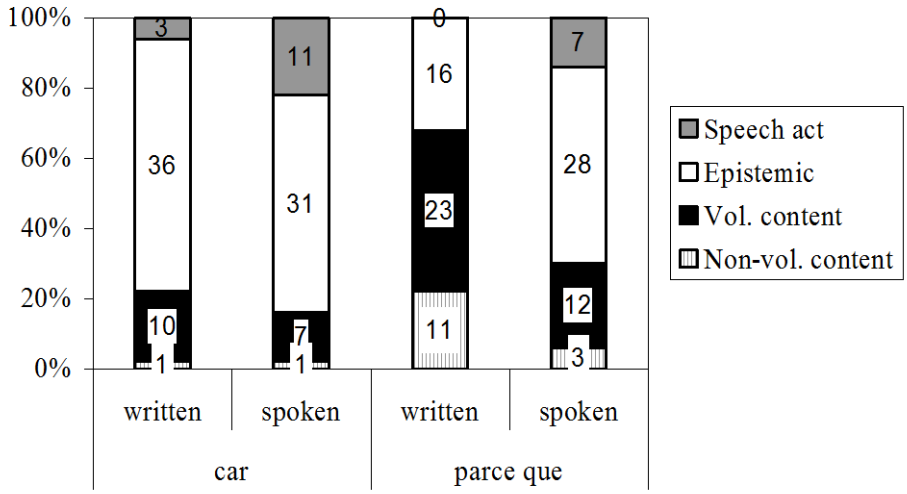

Figure 4. Domain type expressed by *car* and *parce que* in written and spoken French[13]

Previous subjectification analyses of these causals have shown that *car* has the same semantic profile in spoken and in written data (χ^2 (3) = 5.7; p =.14), but that *parce que* shows divergent semantic profiles (see Degand & Pander Maat 2003; Simon & Degand 2007). In written French, *parce que* is only used for objective functions; in spoken French, it can express both objective and

[12] These data are taken from Table 6 in Spooren, Sanders, Huiskes & Degand (in press). Because the written data contained hardly any speech act relations, the speech acts were grouped together with the epistemic relations.
[13] These data are taken from Figure 3 and Figure 4 in Simon & Degand (2007).

subjective functions (χ^2 (3) = 18.3; p < .001). Example (24) shows an instance of the speech act use of *parce que* in the Valibel corpus.

(24) je crois que ça s'appelle en français mais excusez-moi **parce que** je vais peut-être […] estropier le mot hein/un goupillon là [Valibel, chaBR1r]
'I think that in French this is called but excuse me parce que I might ruin the word / a "goupillon" [= sprinkler for holy water]'

It looks as if – in spoken French – *parce que* is taking over the role of *car*, including its more subjective functions. Hence, the difference between spoken and written *parce que* might reflect an ongoing change in French, one involving subjectification. Although the future will have to prove or disprove this claim, we can conclude that this subjectification tendency would not have become visible without making the distinction between the two modalities.

7. Conclusion

Our research started out with two questions: Does subjectification occur in the diachronic development of causal connectives? and 2) What does a corpus-based approach add to the study of this diachronic development? In answer to the first question, we can state that – for the four connectives we studied – subjectification is not an integral part of the diachronic development as a causal connective. It appears that subjectification does occur in the rise of these connectives. More specifically, we see this in the change a) from the preposition *om* 'for' plus the relativum *dat* 'that' to the subordinating connective *omdat*, b) from *par ce que* 'for this that' to the fixed phrase *parce que*, and c) from Latin *qua re* 'for which/what reason' to PDF *car*. In the later stages of their development, in which they remain to be used as a causal connective, only *parce que* seems to undergo subjectification. The other three causal connectives show hardly any subjectivity changes. On the contrary, *want* seems to loose some ground to intersubjectivity markers such as *immers* 'indeed', and *car* loses ground to *parce que* in the expression of intersubjective cause in Spoken PDF.

It appears that at least two general paths of development can be distinguished for causal connectives. The French case is an example of the first path, in which one causal marker (*parce que*) gradually takes over the function of another marker (*car*). This development goes hand in hand with an increase in subjectivity: *parce que* can nowadays express speech act relations, which it could not express in earlier periods. The Dutch case is an example of the second path, in which no large changes in the system occur: both *want* and *omdat* remain stable markers of causality, each with their own degree of subjectivity.

Overall, the semantic profile of three connectives under investigation remains relatively stable. Sometimes connectives take over functions that were previously expressed by other markers (e.g. *omdat* nowadays expresses relations that were previously marked with *dewijl* or *overmits*) and sometimes a connective looses ground to another marker (e.g. the finalistic use of *omdat* is taken over by *opdat* and *zodat* 'so that'). This stability within the connective use need not come as a surprise: subjectification is not an obligatory characteristic of diachronic developments. An explanation of the stable subjectivity profile might be that three causal connectives did not show real changes in their domains of use in the sense that they came to be used in a domain in which they could not occur earlier. The connectives *want*, *omdat*, and *car* could be used in all three domains from the earliest century on. It may be the case that subjectification only occurs or: in the

actual phases of grammaticalization, in which these items gain new (grammatical) meanings, and not when lexical items show a shift in the distribution over the meanings they can already express.

Our second research question concerned the merits of a corpus-based approach: what does a corpus-based approach add to the study of this diachronic development? Our analyses have shown that each of the four methods introduced in Table 1 has its own merits and limitations. Although each of these methods is valuable in its own respect, we would like to stress here the fact that they are most effective when combined. An analysis of translation corpora provides the researcher with a semantic network, including synchronic linguistic competitors that may be relevant for the diachronic analysis as well. Hence, this method can be regarded as a good starting point for diachronic research or for a functional-semantic analysis. The analysis of data from different texts in different periods may reveal diachronic changes. This analysis should be at the heart of the study: it provides authentic data in their context of use, and allows for both qualitative and quantitative research. The diachronic analysis of comparable material/data/texts (here, bible fragments) may reveal whether these changes also occur when the genre is kept constant. This enables the researcher to separate out genre effects from real diachronic changes. This latter analysis may also reveal whether the proportion of use of linguistic competitors changes, and hence, whether changes in the connective system as a whole occur. Finally, the comparison of written and spoken data may reveal change in progress. The analysis of spoken and written materials is also useful in tracing stable differences between spoken and written language.

We hope to have shown that "grammaticalization studies can gain from the systematic and principled use of large computerized corpora and the methods which have been developed within corpus linguistics" (Lindquist & Mair 2004: x). We think that subjectification processes can be studied in much more detail, and that the results thus obtained will lead to a refinement of the theoretical model.

Primary sources

BFM – *Base de Français Médiéval*. Lyon: UMR 5191 ICAR / ENS-LSH, 2005. Online: <http://bfm.ens-lsh.fr>.

CD-rom *Middelnederlands* 'Middle Dutch' (1998). Den Haag/Antwerpen: Sdu.

CD-rom *Klassieke literatuur: Nederlandse letterkunde van de Middeleeuwen tot en met de Tachtigers* 'Classical literature: Dutch literature from the Middle Ages to the Eightiers' (1999). Utrecht: Het Spectrum.

DB 1477 – *Delftse Bijbel* 1477. N. van der Sijs, 2008. Online: <http://www. bijbelsdigitaal.nl>.

DMF – *Base du Dictionnaire de Moyen Français*, UMR 7118 ATILF / Nancy2, 2007. Online: <http://atilf.atilf.fr/dmf>.

Frantext – *Base Textuelle Frantext*, UMR 7118 ATILF / Nancy2, 2008. Online: <http://www.frantext.fr>.

NBG 1951 – *NBG-vertaling 1951*. Nederlands Bijbelgenootschap, 1951. Online: <http://www.biblija.net>.

NBV 2004/2007 – *De Nieuwe Bijbelvertaling*. Nederlands Bijbelgenootschap, 2004/2007. Online: <http://www.biblija.net>.

Oostendorp, M. van (ed.). *Project Laurens Jansz. Coster: Klassieke Nederlandstalige literatuur in elektronische edities* 'Project Laurens Jansz. Coster: Classical Dutch literature in electronic editions'. Online: http://cf.hum.uva.nl/dsp/ljc/.

SV 1637 – *Statenvertaling 1637*. N. van der Sijs, 2008. Online: <http://www.bijbelsdigitaal.nl>.

References

Aijmer, K. (2004). The semantic path from modality to aspect: *Be able to* in a cross-linguistic perspective. In: H. Lindquist & C. Mair (eds.), *Corpus approaches to grammaticalization in English*. Amsterdam/Philadelphia: John Benjamins, 57-78.

Aijmer, K. & Altenberg, B. (1996). Introduction. In: K. Aijmer, B. Altenberg & M. Johansson (eds.), *Languages in contrast: Papers from a symposium on text-based cross-linguistic studies*. Lund: Lund University Press, 11-16.

Athanasiadou, A., Canakis, C. & Cornillie, B. (eds.) (2006). *Subjectification: Various paths to Subjectivity*. Berlin/New York: Mouton de Gruyter.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics* 19/2, 219-241.

Biber, D., Conrad S. & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Bybee, J., Perkins, R. & Pagliuca, W. (1994). *The evolution of grammar; Tense, aspect, and modality in the languages of the world.* Chicago/Londen: The University of Chicago Press.

Chafe, W. & Danielewicz, J. (1987). Chapter 3: Properties of spoken and written language. In: R. Horowitz & S. J. Samuels (eds.), *Comprehending oral and written language*. San Diego etc.: Academic Press, 83-113.

Claridge, C. & Wilson, A. (2002). Style evolution in the English sermon. In: T. Fanego, B. Mendez-Naya & E. Seoane (eds.), *Sounds, words, texts, and change: Selected papers from 11 ICEHL, Santiago de Compostela, 7-11 September 2000. Volume 2.* Amsterdam: John Benjamins, 25-44.

Condamines, A. (2008). Taking *genre* into account when analysing conceptual relation patterns. *Corpora* 3/2, 115-140.

Dasher, R. (1995). *Grammaticalization in the system of Japanese predicate honorifics*. Ph.D. Dissertation, Stanford University.

Degand, L. (2004). Contrastive analyses, translation and Speaker Involvement: the case of *puisque* and *aangezien*. M. Achard & S. Kemmer (eds.), *Language, Culture and Mind*. Stanford: CSLI Publications, 251-270.

Degand, L. & Fagard, B. (in press). Intersubjectification des connecteurs. Le cas de *car* et *parce que*. *Revista de Estudos Linguísticos da Universidade do Port*.

Degand, L. & Fagard, B. (submitted). *Alors* between Discourse and Grammar. The role of syntactic position. To appear in *Functions of Language.*

Degand, L. & Pander Maat, H. (2003). A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale, A. Verhagen & J. van de Weijer (eds.) *Usage-based approaches to Dutch.* Utrecht: LOT, 175-199.

Doherty, M. (1998). Clauses or phrases – a principled account of *when*-clauses in translations between English and German. In: S. Johansson & S. Oksefjell (eds.), *Corpora and cross-linguistic research: Theory, method, and case studies.* Amsterdam/Atlanta: Rodopi, 235-254.

Dyvik, H. (1998). A translational basis for semantics. In: S. Johansson & S. Oksefjell (eds.), *Corpora and cross-linguistic research: Theory, method, and case studies*. Amsterdam/Atlanta: Rodopi, 51-86.

Dyvik, H. (2004). Translations as semantic mirrors. From parallel corpus to WordNet. *Language and computers* 1, 311-326.

Erman, B. & Kotsinas, U.-B. (1993). Pragmaticalization: the case of *ba'* and *you know*. *Studier i Modern Språkvetenskap, Acta Universitatis Stockholmiensis, New Series* 10, 76-93.

Evers-Vermeul, J. (2005). *The development of Dutch connectives. Change and acquisition as windows on form-function relations.* Ph.D. dissertation, Utrecht University. Utrecht: LOT.

Evers-Vermeul, J. & Stukker, N. (2003). Subjectificatie in de ontwikkeling van causale connectieven? De diachronie van *daarom*, *dus*, *want* en *omdat*. *Gramma/TTT* 9-2/3, 111-139.

Evers-Vermeul. J. (2008). *"Want ghy zijt stof…" Over het belang van oude bijbelvertalingen voor diachroon onderzoek*. Paper presented at Cogling 3, Leiden, December 19-20, 2008.

Fagard, B. (2008). *Parce que, perché, porque* dans les langues romanes médiévales: l'utilité des études sur corpus. *Corpus* 7, 83-113. Also available online: <http://corpus.revues.org/>.

Fagard, B. & Degand, L. (2008). La fortune des mots: Grandeur et décadence de *car*. *Actes du Congrès Mondial de Linguistique Française* (CD-ROM), available online: <http://www.linguistiquefrancaise.org/index.php?option=article&access=standard&Itemid=129&url=/articles/cmlf/pdf/2008/01/cmlf08213.pdf>.

Francard, M., Geron G. & Wilmet, R. (2002). La banque de données VALIBEL: des ressources textuelles orales pour l'étude du français en Wallonie et à Bruxelles. In: C.D. Pusch & W. Raible (eds.), *Romanistische Korpuslinguistik – Korpora und gesprochene Sprache / Romance Corpus Linguistics – Corpora and Spoken Language (= ScriptOralia* 126*)*. Tübingen: Gunter Narr, 71-80.

Gellerstam, M. (1996). Translations as a source for cross-linguistic studies. In: K. Aijmer, B. Altenberg & M. Johansson (eds.), *Languages in contrast: Papers from a symposium on text-based cross-linguistic studies*. Lund: Lund University Press, 53-62.

Genetti, C. (1991). From postposition to subordinator in Newari. In: E.C. Traugott & B. Heine (eds.), *Approaches to grammaticalization*, vol. II. Amsterdam/Philadelphia: Benjamins, 227-255.

González-Cruz, A.I. (2007), On the subjectification of adverbial clause connectives: Semantic and pragmatic considerations in the development of while-clauses. In: U. Lenker & A. Meurman-Solin (eds.), *Connectives in the history of English*. Amsterdam/Philadelphia: John Benjamins, 145-166.

Gries, S. Th. (2006). Exploring variability within and between corpora: some methodological considerations. *Corpora* 1/2, 109-151.

Günthner, S. (1996). From subordination to coordination? Verb-second in German clausal and concessive constructions. *Pragmatics* 6/3, 323-356.

Günthner, S. & Mutz, K. (2004). Grammaticalization vs. pragmaticalization? The development of pragmatic markers in German and Italian. In: W. Bisang, N.P. Himmelmann & B. Wiemer (eds.), *What makes grammaticalization? A look from its fringes and its components*. Berlin/New York: Mouton de Gruyter, 77-107.

Halliday, M.A.K. (2004). The spoken language corpus: a foundation for grammatical theory. In: K. Aijmer & B. Altenberg (eds.), *Advances in corpus linguistics*. Amsterdam/New York, NY: Rodopi, 11-38.

Hansen, M.-B.M. & Rossari, C. (2005). The evolution of pragmatic markers: Introduction. *Journal of Historical Pragmatics* 6/2, 177-187.

Herring, S.C., Van Reenen, P.Th. & Schøsler, L. (2000). On textual parameters and older languages. In S.C. Herring, P.Th. Van Reenen & L. Schøsler (eds.), *Textual parameters and older languages*. Amsterdam: John Benjamins, 1-31.

Higashiizumi, Y. (2006). *From a subordinate clause to an independent clause: A history of English because-clause and Japanese kara-clause*. Tokyo: Hituzi Shobo.

Hoffmann, S. (2004). Are low-frequency complex prepositions grammaticalized? On the limits of corpus data – and the importance of intuition. In: H. Lindquist & C. Mair (eds.), *Corpus approaches to grammaticalization in English*. Amsterdam/Philadelphia: John Benjamins, 171-210.

Jacobs, A. & Jucker, A.H. (1995). The historical perspective in pragmatics. In: A.H. Jucker (ed.), *Historical pragmatics. Pragmatic developments in the history of English*. Amsterdam: John Benjamins, 3-33.

Johansson, S. (1998). On the role of corpora in cross-linguistic research. In: S. Johansson & S. Oksefjell (eds.), *Corpora and cross-linguistic research: Theory, method, and case studies*. Amsterdam/Atlanta: Rodopi, 3-24.

Keller, R. (1995). The epistemic weil. In: Stein, D. & S. Wright (eds.), *Subjectivity and subjectivisation: linguistic perspectives*. Cambridge: Cambridge University Press, 16-30.

Koch, P. & Oesterreicher, W. (1985). Sprache der Nähe - Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36/85, 15-43.

Labov, W. (1994). *Principles of linguistic change* (I: Internal Factors). Oxford: Blackwell.

Lehmann, C. (1991). Grammaticalization and related changes in contemporary German. In: E.C. Traugott & B. Heine (eds.), *Approaches to grammaticalization, Vol. II.* Amsterdam: Benjamins, 493-535.

Lewis, D. (2005). Corpus comparables et analyse contrastive: l'apport d'un corpus français/anglais de discours politiques à l'analyse des connecteurs adversatifs [Comparable corpora and contrastive analysis: the contribution of a French/English corpus of political discourse to the analysis of adversative connectives]. In: G. Wiliams (ed.), *La linguistique de corpus*. Rennes Cedex: Presses Universitaires de Rennes, 179-190.

Lindquist, H. & Mair, C. (eds.) (2004). *Corpus approaches to grammaticalization in English*. Amsterdam/Philadelphia: John Benjamins.

Lindström, J. & Wide, C. (2005). Tracing the origins of a set of discourse particles: Swedish particles of the type *you know*. *Journal of Historical Pragmatics* 6/2, 211-236.

Mair, C. (2004). Corpus linguistics and grammaticalisation theory: Statistics, frequencies, and beyond. In: H. Lindquist & C. Mair (eds.), *Corpus approaches to grammaticalization in English*. Amsterdam/Philadelphia: John Benjamins, 121-150.

Molencki, R. (2007). The evolution of *since* in medieval English. In: U. Lenker & A. Meurman-Solin (eds.), *Connectives in the history of English*. Amsterdam/Philadelphia: John Benjamins, 97-113.

Mortier, L. (2007). *Perspectives on grammaticalization and speakers' involvement. The case of progressive and continuative periphrases in French and Dutch.* Unpublished doctoral dissertation. Leuven: KU Leuven.

Noël, D. (2003). Translations as evidence for semantics: an illustration. *Linguistics* 41/4, 757-785.

Onodera, N. (2000). Development of *demo* type connectives and *na* elements: Two extremes of Japanese discourse markers. *Journal of Historical Pragmatics* 1/1, 27-55.

Onodera, N.O. (2004). Japanese discourse markers: Synchronic and diachronic discourse analysis. Amsterdam: John Benjamins.

Oostdijk, N. (2000). The spoken Dutch corpus project. *The ELRA Newsletter* 5/2, 4-8.

Pander Maat, H. & Degand, L. (2001). Scaling causal relations and connectives in terms of speaker involvement. *Cognitive Linguistics* 12/3, 211-245.

Pander Maat, H. & Sanders, T. (2001). Subjectivity in causal connectives: An empirical study in language use. *Cognitive Linguistics* 12/3, 247-273.

Pit, M. (2003). *How to express yourself with a causal connective*. *Subjectivity and causal connectives in Dutch, German and French.* Ph.D. dissertation, Utrecht University. Amsterdam/New York: Rodopi.

Partington, A. (2006). Metaphors, motifs and similes across discourse types: Corpus-assisted discourse studies (CADS) at work. In: A. Stefanowitsch & S. Th. Gries (eds.), *Corpus-based approaches to metaphor and metonymy*. Berlin/New York: Mouton de Gruyter, 267-304.

Prévost, S. (1999). *Aussi* en position initiale: évolution sémantico-syntaxique du 12^{e} au 16^{e} siècle. *Verbum* XXI/3, 351-380.

Prévost, S. (2003). *Quant a*: analyse pragmatique de l'évolution diachronique (14^{e}-16^{e} siècles). In: B. Combettes, A. Theissen & C. Schnedecker (eds.), *Actes du colloque 'Ordre et distinction dans la langue et le discours'*, Metz 99. Paris: Honoré Champion, p. 443-459.

Prévost, S. (2007). *à propos de, à ce propos, à propos*: évolution du 14^{e} au 16^{e} siècle. *Langue Française* 156, 108-126.

Rissanen, M., Kytö, M. & Heikkonen, K. (eds.) (1997). *Grammaticalization at work: Studies of long-term developments in English*. Berlin/New York: Mouton de Gruyter.

Simon, A.C. & Degand, L. (2007). Connecteurs de causalité, implication du locuteur et profils prosodiques. Le cas de *car* et de *parce que*. *Journal of French Language Studies 17*, 323-341.

Spooren, W., Sanders, T., Huiskes, M. & Degand, L. (in press). Subjectivity and causality: A corpus study of spoken language. In: S. Rice & J. Newman (eds.), *Empirical and Experimental Methods in Cognitive/Functional Research.* Stanford: CSLI Publications.

Stefanowitsch, A. (2006). Corpus-based approaches to metaphor and metonymy. In: Anatol Stefanowitsch & Stefan Th. Gries (eds.). Corpus-based approaches to metaphor and metonymy. Berlin: Mouton de Gruyter, 1-16.

Sweetser, E.E. (1990). *From etymology to pragmatics. Metaphorical and cultural aspects of semantic structure*. Cambridge: Cambridge University Press.

Traugott, E.C. (1995). Subjectification in grammaticalisation. In: D. Stein & S. Wright (eds.), *Subjectivity and subjectivisation: linguistic perspectives*. Cambridge: Cambridge University Press, 31-54.

Traugott, E.C. & Dasher, R.B. (2002). *Regularity in semantic change*. Cambridge: Cambridge University Press.

Van der Sijs, N. (2008a). *Verantwoording van de digitale uitgave van de Delftse bijbel uit 1477.* Available online: <http://www.bijbelsdigitaal.nl/info/DB1477_Verantwoording.pdf>.

Van der Sijs, N. (2008b). *Verantwoording van de digitale uitgave van de Statenvertaling 1637.* Available online: <http://www.bijbelsdigitaal.nl/info/SV1637_Verantwoording.pdf>.

Vincent, D. (2005). The journey of non-standard discourse markers in Quebec French: networks based on exemplification. *Journal of Historical Pragmatics* 6/2, 188-210.

Vogl, U. (2007). Het belang van conditionaliteit voor de ontwikkeling van temporeel naar causaal voegwoord. De geschiedenis van dewijl, terwijl, weil en while. *Nederlandse Taalkunde* 12/1, 2-24.