CogniTextes
Revue de l'Association française de linguistique cognitive

Volume 19 | 2019
Corpora and Representativeness

Non-representativeness in corpora: perils, pitfalls and challenges

Thomas Egan

Electronic version
URL: http://journals.openedition.org/cognitextes/1772
DOI: 10.4000/cognitextes.1772
ISSN: 1958-5322

Publisher
Association française de linguistique cognitive

Electronic reference
Thomas Egan, « Non-representativeness in corpora: perils, pitfalls and challenges », *CogniTextes* [Online], Volume 19 | 2019, Online since 17 June 2019, connection on 19 June 2019. URL : http://journals.openedition.org/cognitextes/1772 ; DOI : 10.4000/cognitextes.1772

Non-representativeness in corpora: perils, pitfalls and challenges

Thomas Egan

1. Introduction

1 This article[1] discusses some of the potential challenges the linguist faces when working with corpora which may not be as representative or balanced as might first appear. The discussion is illustrated with examples from the author's own experience of some twenty years' research using various types of corpora: contemporary and historical, mono- and multilingual.

2 The article is divided into five sections. Section 2 contains a brief introduction to the notion of representativeness, followed by a description of some text types which may cause problems for various reasons if included in a general corpus. This description is set against the background of a discussion on the Linguist List in 2000. Section 3 discusses some problematic examples that are to be found in widely used contemporary and historical corpora. Section 4 is concerned with multilingual corpora and the question of how to achieve maximum balance and representativeness in these. Finally, section 5 contains a summary and conclusion.

2. Representativeness and non-representativeness

The notion of representativeness is absolutely central to the corpus linguistic enterprise. According to McEnery & Wilson (2001):

> A corpus in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration. (McEnery & Wilson 2001 : 32)

3 As Leech (2007: 135) succinctly puts it: "Without representativeness, whatever is found to be true of a corpus, is simply true of that corpus – and cannot be extended to anything

else”. There are, however, various ways in which one can conceive of representativeness. Three main aspects are distinguished by Biber (1993).

> Different overall corpus designs represent different populations and meet different research purposes. Three of the possible designs are organized around text production, text reception and texts as products. [...]. A production design would include the texts (spoken and written) actually produced by the individuals in the sample ; a reception design would include the texts listened to or read. (Biber 1993: 245)

4 The compilers of the *British National Corpus* (BNC) took both production and reception into account (Aston & Burnard 1998: 28). Leech is a particularly keen advocate for the inclusion of reception criteria, maintaining that a “conceptually simple way of measuring the importance of a text, for purposes of corpus building, is how many receivers it has” (Leech 2007: 139 ; see also Sinclair 1991: 1). Conceptually simple this measurement may be, but if corpus compilers were to apply it uncritically, this could lead to numerous pitfalls for the corpus researcher, as I aim to show in section 3. Possible sources of some of these pitfalls were mentioned in a lively exchange of views on the Linguist List in 2000 between linguists who espoused formal and functional approaches to grammatical analysis.

5 The thread on the Linguist List started with a review by Andrew Carnie in 2000 of Newmeyer’s 1998 book *Language Form and Language Function* which drew responses from some linguists who thought that the functional theoretical perspective had been misrepresented in Carnie’s post (though not necessarily in Newmeyer’s book). One participant, Marc Hamman, wrote the following:

> A key concern for me is the empirical vagueness which surrounds the notion of grammaticality judgements as a measure of competence. The major problem is that there are sentences which native speakers will judge unacceptable despite being able to understand them perfectly well (a trivial example is the double negative “I don’t got nothing left.”), while other sentences which ought to be “grammatical” sound very strange if not unintelligible (The rat that the cat that the dog bit chased ran.) (Hamman 2000: 11.109)

6 Much of the subsequent discussion on the thread was concerned with grammaticality judgements, and the example *The rat that the cat that the dog bit chased ran* figured prominently in this discussion, with some participants, myself included, arguing that the sentence is not even English, let alone grammatical. Phil Gaines took issue with the contention that grammaticality was related to understandability, writing:

> For the generativist, grammaticality has nothing to do with understandability. [...] Right now, I’m re-reading Ulysses, wherein Joyce famously does delightful acrobatics with grammar. One of his games in a long narrative section is to separate the verb from the subject by as much as 10 lines of text. Two or three careful re-readings of such sentences are necessary to parse them. (Gaines 2000: 11.269)

7 In reply to this post I myself raised the question of whether works like Ulysses should be included in a general, as opposed to a specialised, corpus, arguing that:

> Joyce is a very good example of someone whose output ought to be approached with care. For instance, if one chose to make Ulysses the backbone of one’s corpus of early twentieth century English, one could be landed with serious problems in tracing the evolution of English clause structure. Just take two sentences, from the “Oxen in the Sun” episode, written in 1920 : “Before born babe bliss had. Within womb won he worship.” In this episode Joyce goes on to mimic the style and structure of Middle English writers, Elizabethan writers, etc. etc. I don’t think that

> anyone would argue that these passages should be allowed to influence our description of contemporary English usage. (Egan 2000: 11.322)

8 More particularly, I suggested that the compiler of a general corpus should be wary of three text types, namely those containing:

- utterances specifically produced to be cited, rather than used (typically by linguists, language teachers, etc.)
- utterances which are produced in a conscious effort to stretch the boundaries of the language (typically by poets)
- utterances which are produced in a conscious attempt to ape the expressive modes of a previous era (typically by writers of historical fiction) (Egan 2000: 11.322)

9 I stated explicitly that I did not mean that we should necessarily completely exclude the second and third text types from general corpora, but we should certainly be conscious of the danger of allowing them to be over-represented. Section 3 contains illustrations of some of the pitfalls their inclusion may give rise to.

3. Some possible problems in contemporary and historical corpora

10 This section deals with some examples of problems the researcher may encounter in working with both contemporary and historical corpora. There are five sub-sections, each of which addresses a different aspect of representativeness and illustrates this with an appropriate example. Sections 3.1 – 3.3 are devoted to some constructions in the BNC which are not representative of how the expressions in question were used in the latter half of the twentieth century. Sections 3.4 and 3.5 discuss two sorts of problem encountered while working with historical corpora. These illustrate two different types of problem caused by non-representativeness. Section 3.4 is concerned with intersubjective uses of *not fail to* in historical fiction, and how these may contribute to distort our perception of the construction's evolution. Section 3.5 is concerned with the *continue to* and *continue -ing* constructions and illustrates how a lack of balance between various components in a historical corpus can mislead the historical researcher.

3.1 The invented example problem: the 'see x to be' construction

11 Half of the tokens of the active voice 'see x to be' construction in the BNC are actually taken from Patrick Duffley's (1992) book on the English Infinitive. Moreover, Duffley has in part relied for his examples on Bolinger (1974) and Jespersen (1940). Bolinger, in turn, has taken examples from Jespersen, and one of his examples, *I saw them to be obnoxious*, is repeated as recently as in Horie (2000).

12 In actual fact, the 'see x to be' construction is not used in Present-Day English to encode a one-off judgment, such as that some people are obnoxious, but rather to encode a revised construal (see Egan 2011). (1)–(3), taken from the BNC and the Corpus of Contemporary American English (COCA) may serve to illustrate contemporary usage.

> (1) As they approached, Reni rose from his seat at a table near the large rectangular pool which was the centrepiece of what – as Huy <u>now</u> *saw it to be* - was an unconventionally asymmetrical garden. (BNC H84 2027)
> (2) It could have been the dazzling sunlight, or his eyes playing tricks on him ; but, for a moment, Lucian thought they were half-horse, half-human [...] <u>As they</u>

galloped closer, Lucian *saw them to be* men and women dressed alike in fringed tunics and trousers of soft leather. (COCA: Lloyd Alexander :*The Arkadians*)
(3) They were not all as neatly round as I had first *seen them to be*.
(COCA: Goldman, E S.: *Yellow Jackets*)

13 (1) and (2) both imply an element of revised construal, of re-perception after the mists have cleared, so to speak. Thus in (1) Huy's perspective on the garden changes and he realises that it is asymmetrical. Similarly in (2) Lucian first perceives the riders to be centaur-like. As they come into focus he perceives their outline more closely and the new input leads to a recategorisation of the objects. The point at which the revision occurred is signalled by the underlined adverbial *as they galloped closer*, as it is in (1) by *now*. The construction instantiated by the two tokens may be assigned the following schematic characterisation: "In the dark/from a distance, we imagined x to be y: we later saw it to be z" (see Egan 2008: 152). The revised perceptions encoded in the 'see x to be y' clauses in (1) and (2) encode what the speaker profiles as the correct construal of the situation perceived by the subject. It is the *actual* state of affairs. The exact opposite is the case in (3), which encodes a construal on the part of the subject which is encoded by the speaker as *false*. Thus in (3) the adverbial *first* in the *seen* clause refers to the locus of an original misperception, rather than that of the revision.

14 In this case of the 'see x to be' construction, the inclusion in the corpus by its compilers of a textbook, containing made-up examples designed to illustrate points of grammar, can serve to muddy the waters for the analyst. There are generally taken to be three main methods of procuring data for linguistic inquiry. These are introspection, corpus studies and experiments. The results of these methods can be mutually supportive. However, in order to be so, they must be kept strictly separate. This is obviously not the case when data from introspection are included in a corpus. One could argue that people do read grammar books, so their inclusion in a corpus can be justified by reception criteria. Nevertheless the inclusion of made-up examples without some sort of mark-up flagging them may cause problems for the corpus analyst.

3.2. The quotation problem: the 'remember to have' construction

15 In section 3.1 the problem for the corpus analyst was caused by the occurrence in the corpus of examples especially constructed to illustrate points of grammar. Another type of problem is the inclusion by compilers in the corpus of examples of genuine, but non-contemporary, usage. These typically occur in quotations. Take for example the use of a perfect form of the infinitive to indicate that the situation in the complement clause pertained prior to the time of the matrix verb. As shown by Bowie and Wallis (2016), the construction with the infinitival perfect has been in decline for several centuries. It is still hanging on in constructions containing *seem*, *appear* and verbs expressing judgements, as in *He believed her to have lived in London*. It is however archaic with many matrix verbs, such as *remember* in the pattern *I remembered to have seen a photograph of her*. As Fanego points out, writing of earlier forms of English, "*remember* takes a perfect infinitive that explicitly signals that the reference is to past time. This type of construction is no longer acceptable today, but was in common use in Modern English [...] and can even be found in texts dating back to the middle of the twentieth century" (Fanego 1996: 75). The data in the BNC may be taken as support for this judgement of Fanego's. There are two instantiations of the 'remember to have' construction in the corpus. The most recent of them occurs in an extract from a letter written by John Stuart Mill, cited in a scientific

treatise. The inclusion of archaic examples cited in biographies or scientific texts is impossible for the corpus compiler to avoid, unlike the inclusion of examples in grammar books. The latter genre can, after all, be excluded totally from the corpus. It would, however, be helpful if historical quotations were marked as such. As it stands, the analyst, using a concordancer, must rely on his or her own historical antennae to trigger the need for a closer examination of the tokens in question. It goes without saying that not all competent researchers into the language of one era are equally well-versed in the language of other eras.

3.3. The disguised text problem: the 'beseech x bare infinitive' construction

16 Like the examples in 3.2, those in the present section are also archaic expressions, but in this case they are contained neither in a grammar book, nor in a letter from an earlier age. Consider (4):

> (4) I *beseech you say* not one word but 'yes' or 'no' till I have said all I have to say. (FU4 1774)

17 (4) is one of two examples containing 'beseech X *bare infinitive*' in the BNC. It is taken from the text of the stage play *Pamela*, which is a 1987 adaptation of, and contains dialogue lifted verbatim from, Richardson's novel from 1740.[2] The other token of the construction is from a work entitled *Warriors of Christendom*, which is said by the BNC to have been written by John Matthews and Bob Stewart and published in 1988. The actual text in the corpus however is Robert Southey's 1808 translation of a Spanish text from 1637. As was the case with the 'remember to have *past participle*' construction, no indication of the original provenance of the 'beseech X *bare infinitive*' can be gleaned from the corpus itself. Analysts are thus again left to rely on their own intuitions about possible archaisms to prompt them to delve further into the source material.

18 These two examples differ in that the Southey text is obviously bogus on anything but reception criteria. It ought to have no place in a corpus designed to represent English produced in the latter half of the twentieth century. In the case of the stage play of *Pamela* , there is no attempt to conceal the fact that it is an adaptation, the full title being *Pamela, or, The reform of a rake: a play adapted from the novel by Samuel Richardson*. The adaptation, however, is a close one, resulting in a lot of dialogue more typical of eighteenth than twentieth century English. In fact, the difference between this example and the quotations in section 3.2 is merely one of quantity – most of the *Pamela* text consists of quotations !

19 At a pinch the inclusion in a general corpus of both the Southey text and the *Pamela* play could be defended on grounds of reception, since they were both read/heard by late twentieth century audiences. When it comes to production, however, they instantiate re production of the language of an earlier era, rather than the language of the late twentieth century.

3.4 The historical fiction problem: intersubjective uses of *not fail to*

20 This section deals with a sort of language that has been termed 'bygonese' by the historical novelist David Mitchell (see Stocker 2012: 313). It does so by examining in detail one construction, the 'not fail to' construction, which was borrowed from French in the

fourteenth century, along with its 'fail to' counterpart (see Egan 2010). While the latter non-negated, negative polarity, construction was little used before the Early Modern English (EModE) period, the negated, positive polarity, construction took root quickly, and, in the fifteenth century, came to be used in the second person to encode injunctions and, in the sixteenth century, combined with first person *will* or *shall* to encode promises. (5) and (6), both from the *Corpus of Early English Correspondence Sampler* (CEECS), illustrate the two senses.[3]

> (5) And that ye *faille not thus to doo* as ye tendre our pleasure.
> (CEECS, Henry VII to Sir Gilbert Talbot, ca. 1500)
> (6) According to my promise, *I will not faile to let you understand* of my proseedings last week. (CEECS, Anne Lady Meautys to Jane Lady Bacon, 1632)

21 In (5), the writer imposes an obligation on the addressee. The negated construction as used here carries deontic force, as it does in the earlier French example below (7).

> (7) et vous touz, juges, ne failliez Pas a ce faire.
> 'and all of you, judges, don't fail to do this.'
> *Miracle de Saint Lorens*, written before 1339, from *Corpus de la Littérature Médiévale des Origines au 15e Siècle.*

22 Pragmatically, the *promise* intersubjective sense, illustrated in (6), is the mirror image of the injunction sense. When used with first-person subjects and the modals *will* or *shall*, 'not fail to' codes a promise on the part of the speaker. The expression *I will not faile to* in (6) could be paraphrased 'I promise to', and indeed the writer actually uses the word *promise* to refer to a previous commitment to keep the addressee informed of her actions.

23 We can see in (8)–(9), taken from the 1710–1780 sub-period of *Corpus of Late Modern English Texts* (CLMET), that both promise and injunction senses continue to be used in the eighteenth century.

> (8) *I will not fail to make* your compliments to the Pomfrets and Carterets.
> (CLMET, letter from Robert Walpole to Horace Mann, 1744)
> (9) I desire, therefore, that one of you two *will not fail to write* to me once a week.
> (CLMET, letter from Chesterfield to his son, 1748)

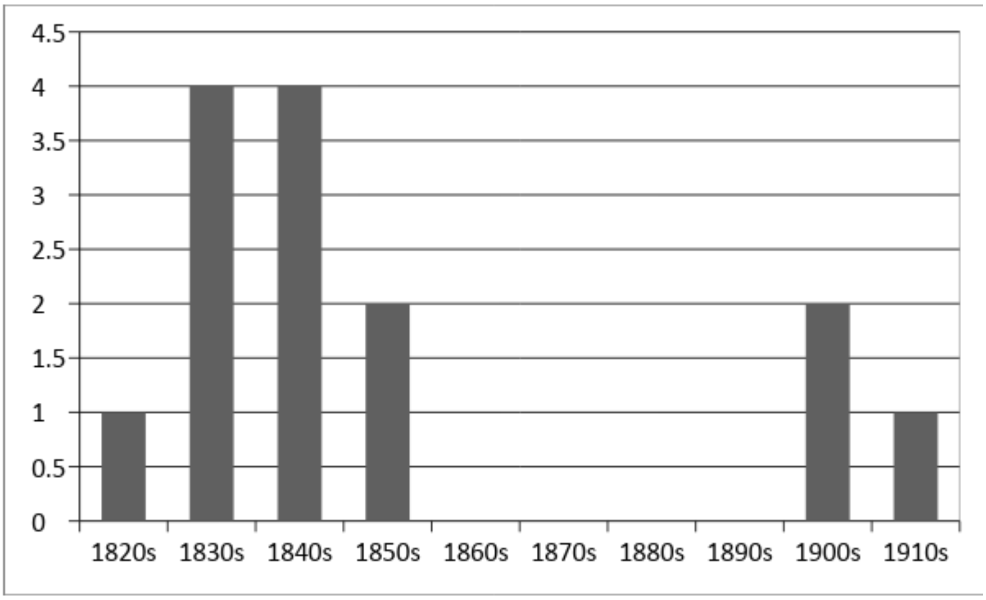
24 Example (8) resembles (6) in encoding a promise on the part of the speaker, while (9) resembles (5) in encoding an injunction on the addressee.

25 The first half of the nineteenth century witnessed a decrease in the use of the intersubjective construction. There are only four tokens of the intersubjective senses in the 1780–1850 period of CLMET, two of which are cited as (10) and (11).

> (10) "*I shall not fail to do so*, madam," replied Suffolk. "Your majesty will have strict justice."
> (CLMET, Ainsworth, *Windsor Castle*, 1843)
> (11) "Your grace acts as beseems a loyal gentleman," replied Surrey. "Hereafter *I will not fail to account* to you for my conduct in any way you please."
> (CLMET, Ainsworth, *Windsor Castle*, 1843)

26 *Windsor Castle*, the text in which (10) and (11) occur, is a work of historical fiction, set in the reign of Henry VIII. It is quite possible therefore that the author employed what he felt to be a somewhat archaic mode of expression in order to lend his narrative a period feel. Whether or not this is the case, the construction was certainly losing ground by the 1840s. In the 1850–1920 period of the CLMET, the construction is not attested. There are, however, a few later examples in the much larger *Corpus of Historical American English* (COHA). Details of the incidence of the intersubjective *promise* construction in the century 1820–1919 are given in Figure 1.

Figure 1: Raw frequencies for the intersubjective sense of ‘will not fail to’ in COHA, 1820–1919.



27 In what sort of texts do we find intersubjective ‘will not fail to’ in American English? Consider (12)–(13):

(12) “Nevertheless,” continued Amador, “*I will not fail to make* thy petition, backed with my own request, to the seor Narvaez”. (COHA, Robert M. Bird, *Calavar: Or The Knight of the Conquest, A Romance of Mexico, vol.* 1, 1834)

(13) “*I will not fail to wait* on thee, my liege.” (COHA, Horatio N. Moore, *Orlando*, 1835)

28 It would not be necessary to know the titles of these works to assign them to the genre of historical fiction. For example, both texts employ the archaic form of the second-person singular pronoun. Another vocabulary item not in current use in nineteenth century America is the noun *liege*. There can be little doubt that first-person ‘will not fail to’ is also considered archaic, or at least exotic, by these authors. More evidence of the expression contributing to lending an exotic tinge to a narrative may be seen in (14).

(14) “I am called Master Anseau, and am the goldsmith of our seigneur, the king of France, at the sign of St. Eloi. Promise me to be in this field the next Sabbath, and *I will not fail to come*, though it were raining halberts.” (COHA, Maturin M. Ballou, *The Sea-Witch Or, the African Quadroon: A Story of the Slave Coast*, 1855)

29 Like (10)–(13), (14) is clearly the product of an author attempting to recreate what he takes to be the dialogue of a previous age. The author is in fact H. W. Loring, not Maturin M. Ballou, as indicated by COHA, and the quotation is from a short story entitled *The Goldsmith of Paris.*[4] Among the archaic features in (14) is the title *Master*, the description of the French king as *seigneur*, the address *at the sign of*, the use of *the next Sabbath* as the date for an appointment, and the raining of *halberts* instead of the more usual English *cats and dogs*. As far as I know, it has never rained halberts in English, although it occasionally did so in Early Modern French. In addition to these archaic and/or exotic expressions, we find the equally archaic ‘will not fail to’.

30 The use of ‘will not fail to’ to lend an exotic air to the dialogue of historical fictional texts peters out in the course of the nineteenth century, with a few outliers from the early twentieth century. One of these is cited as (15).

(15) “My lady, go to thy tiring room and make thee ready. *I will not fail to wait* thee.”

(COHA, Beulah M. Dix, *Road to Yesterday*, 1906)

31 (15) contains the archaic pronouns *thee* and *thy* as well as *tiring* room for dressing room and *wait* used as a transitive verb. It is taken not from a novel, but from a popular play which was turned into an even more popular film, directed by Cecil B. DeMille in 1925. This particular utterance was not included among the dialogue intertitles in the film, but there are many others that contain archaic language, such as (16)–(18).

(16) What traffic dost thou hold with that black witch ?
(17) Thou knowest my face, even as I dreamed of knowing thine.
(18) In heaven's name, what do you in this array ?

32 The question was raised in section 2 as to whether such language should be included in a general corpus, as opposed to a specialised corpus of historical fiction. Does the sort of dialogue in (16)–(18) actually bear any relationship to the state of the English language in 1925, the year the film was made? After all, we have no reason to think that cinema-goers left the cinema thou-ing and thee-ing one another. In other words, they would have recognised exotic language for what it was, and this includes the 'will not fail to' construction. Should then this sort of language, to which a contemporary audience was exposed, but which it did not itself practice, be included in a general corpus? The answer will no doubt in part be influenced by the target audience of the corpus compilers and the sorts of research questions these scholars are engaged in investigating. If their interests are primarily diachronic, it obviously makes good sense to avoid including texts which contain a plethora of archaisms, without flagging these in some way or other. But even if their interests are primarily synchronic, will not the inclusion of second person singular *thou* complicate their description of the contemporary pronominal system? And what of the use of the simple present form of *do* in (18), where we would expect the progressive in early twentieth century English?

33 The question of how much, if any, historical fiction should be included in a general corpus is related to the question of how much emphasis we should place on reception as opposed to production. After all, not all many people <u>write</u> historical fiction. On the other hand, far more people <u>read</u> it. Moreover, people today continue to read genuine older fiction, written by authors such as Jane Austen and the Bröntes, for example. So there are lots of words and constructions that literate language users have an understanding of, but never employ themselves (unless of course they are attempting self-consciously to speak in the language of a bygone age). We would not expect to come across them in spoken corpora, for instance. Is the fact that people understand expressions sufficient to warrant their inclusion in a general corpus? What about expressions in other languages? From the point of view of reception, what is the difference between an English speaker reading Shakespeare and reading a French text? If one is researching the language of individual speakers, both are equally relevant. However, if one is researching the language of a speech community, both are arguably equally irrelevant. Sinclair (1991: 18) points out that "the phraseology of Shakespeare and the King James Bible still exert an influence on present-day English usage". Be that as it may, as far as I am aware, no one has suggested that we include Shakespeare or the King James Bible in a corpus of Present-Day English. I would suggest that the corpus compiler needs to be equally careful about including historical fiction, and that researchers need to be on their guard about its possible inclusion in a corpus.

3.5 The skewed genres problem: the *continue to/-ing* constructions

34 This section is devoted to a discussion of the historical evolution of the *continue to* and *continue –ing* constructions with particular reference to the representation of these two constructions in the first version of CLMET. Late Modern English (LModE) saw a general increase in the employment of *–ing* complements in what has come to be known, following Rohdenburg (2006), as *The Great Complement Shift* (see also Fanego 2004, Rudanko 2000, Vosberg 2003).

> Over the past few centuries, English has experienced a massive restructuring of its system of sentential complementation, which may be referred to as the Great Complement Shift. [...]Perhaps the most important set of changes is provided by the establishment of the gerund as a second type of non-finite complement [...] at the expense of infinitive (and *that*-clauses). (Rohdenburg 2006: 143)

35 One may wonder how constructions containing matrix verbs which occurred predominantly with *–ing* complements from their first appearance in the language in Middle English (ME) are affected by this shift. One such verb is *prefer*, the evolution of which is described in Egan (2012). Another such verb is *continue*, which is not listed by the OED with *to*-infinitive complements before the seventeenth century. Example (19) is from ME, while (20) is from EModE. (19) and (20) are both from the OED.

> (19) 1382 Wyclif Luke xxiii. 23 And thei contynueden axinge with greete voices, that he schulde be crucified.
>
> (20) 1651 Hobbes Leviath. ii. xxvi. 139 By whose authority they now continue to be Lawes.

36 Further evidence of the late arrival of the *to* complement form is provided by the Helsinki corpus, which contains five tokens of *continue –ing* as opposed to just two tokens of *continue to*, both of which are from EModE. Both non-finite forms of complement were thus firmly established in the language in EModE. To investigate their further development in LModE, I turned to CLMET (the original, not the extended versions, which were not yet available at the time I carried out this study). Table 1 contains the raw figures for both constructions in all three parts of CLMET, together with normalised frequencies per 1 million words.

Table 1: Raw numbers and frequencies per million words of both non-finite *continue* constructions in three parts of CLMET

	1710–1780	1780–1850	1850–1920
continue *to*	145 (69)	215 (57)	206 (51)
continue *–ing*	16 (8)	65 (17)	27 (7)

37 At first glance, Table 1 appears to contain evidence that the matrix verb *continue* did indeed partake in the development labelled *The Great Complement Shift*, since there is a statistically significant increase (Pearson's chi.sq. = 12.017, p =0.000527) in the use of the –*ing* form of complement in the second period. However, there is an equally significant decrease in its employment in the third period. The reason for the latter development is far from obvious, and required further investigation of the corpus data.

38 An exhaustive examination of the syntax of all the tokens in the corpus having yielded no clue that would serve to explain the rise and fall in distribution of the continue *–ing* construction, it proved necessary to look more closely at the text types in which the examples occur, starting with a crude division between fiction and non-fiction texts. The results of this division show that it is the non-fiction texts that are responsible for the increase in the *-ing* form from the first to the second period, and its subsequent decline. There is no difference to speak of in the ratio of *to* to *–ing* complements in fiction texts in the first two periods (it is roughly 4: 1 in both). On the other hand, there is a very clear, statistically significant, difference in the ratio in the non-fiction texts (20: 1 compared to 3: 1).

39 Let us first consider some tokens of the *continue to* construction in CLMET 1710–1780.

> (21) Ninety years is time sufficient to reduce any commodity, of which there is no monopoly, to its natural price, or to the lowest price at which, while it pays a particular tax, it can *continue to be sold* for any considerable time together. (Adam Smith, *The Wealth of* Nations, 1776)
>
> (22) The small number of Irish cattle imported since their importation was permitted, together with the good price at which lean cattle still *continue to sell*, seem to demonstrate, that even the breeding countries of Great Britain are never likely to be much affected by the free importation of Irish cattle. (Adam Smith, *The Wealth of* Nations, 1776)

40 Both examples, (21)–(22), are from one and the same work. In fact this single work is responsible for more than half of the 102 non-fiction tokens of the construction in the 1710–1780 sub-corpus. Smith's intimate friend, David Hume, is responsible for a further 10 %. Moreover, Smith was a member of *The Literary Club*, co-founded by Johnson in 1763, as were Gibbon, Reynolds and Burke, also represented in the corpus, as is Chesterfield who sponsored Johnson's dictionary. Together these seven authors account for two-thirds of the writers of non-fiction represented in CLMET 1710–1780. In other words, there is no doubt that the authors represented in the corpus constitute a rather narrow cross-section of the contemporary literate public. Whether the nature of this restricted sample actually influenced the results is, of course, another question, but the researcher should at least bear this possibility in mind when examining the data.[5]

41 Having noted that more than half of the examples of *continue to* in the non-fiction texts of CLMET 1710–1780 are to be found in an economic treatise, we can turn to the question of the types of text in which *continue -ing* occurs in CLMET 1780–1850. (23)–(26) are typical examples.

> (23) We *continued travelling* northward, in a zigzag line ; sometimes stopping a day to geologize. (Darwin, *The Voyage of the Beagle*, 1839)
>
> (24) We *continued riding* the greater part of the day, but had very bad sport, not seeing a kangaroo, or even a wild dog. (Darwin, *The Voyage of the Beagle*, 1839)
>
> (25) During three or four hours that we *continued ascending*, the scene increased in sterility and desolation. (William Beckford, *Dreams, Waking Thoughts, and Incidents,* 1783)
>
> (26) We *continued straying* from cloister to cloister, and wandering along the winding passages and intricate galleries of this immense edifice, whilst the Coadjutor was assisting at vespers. (William Beckford, *Dreams, Waking Thoughts, and Incidents,* 1783)

42 The complement clause predicates in (23)–(26) are all motion verbs, with all four examples describing some part of a journey. A further tranche of examples, illustrated here by (27)–(30), describe events which occurred during a journey.

(27) When we were on shore the party looked rather alarmed, but *continued talking and making gestures* with great rapidity. (Darwin, *The Voyage of the Beagle*, 1839)
(28) We *continued discoursing* until we arrived at Pegoens. (Borrow, *The Bible in Spain,* 1843)
(29) He *continued playing and singing* for a considerable time, the two younger females dancing in the meanwhile with unwearied diligence. (Borrow, *The Bible in Spain,* 1843)
(30) The rocks here formed a spacious terrace ; along which I *continued surveying* the distant groves, and marking the solemn approach of night. (William Beckford, *Dreams, Waking Thoughts, and Incidents,* 1783)

43 In addition to the eight examples (23)–(30), a further 20 examples of *continue –ing* occur in the same three texts, as shown in Table 2. *The Voyage of the Beagle* is actually a mixture of travel memoir and scientific treatise, but since the *continue –ing* examples occur in the narrative travel sections I have classified it for present purposes as a travel memoir and listed it as such in Table 2, which contains an overview of the distribution of the two *continue* constructions according to genre in the non-fiction texts of CLMET 1780–1850.

Table 2: Both *continue* constructions in non-fictional texts in CLMET 1780–1850 by genre

Genre	Texts	continue *to*	continue *–ing*	Total
Travel memoirs	4	31	28	59
History/ Biography	4	36	9	45
Letters	4	21	0	21
Social Studies	3	19	0	19
Verse/ Essays	2	3	0	3
Total	17	110	37	147

44 As noted above, 28 examples of *continue -ing* are from three travel memoirs. There is a fourth such memoir in the corpus, Mary Wollstonecraft's *Letters on Norway, Sweden, and Denmark*, which contains no examples of the construction. This work, however, does not so much contain narrative descriptions of journeys from point A to point B, as reflections on society in the three countries.

45 To sum up, we have seen that the steep rise in the incidence of the *continue -ing* construction in the 1780–1850 period in CLMET does not represent a genuine increase in the use of the construction in the language. As a matter of fact, the travel memoir genre, in which the construction is most prevalent in this sub-corpus, is not represented at all in the previous sub-corpus. It should be pointed out that the compiler of CLMET, Hendrik de Smet, did not attempt to achieve a balance of genres across the sub-periods.[6] In fact, such a balance may be virtually impossible to achieve, given the fact that genres come and go over the course of history. In any case, we find further evidence for the likelihood that the matrix verb *continue* was not affected by *The Great Complement Shift* in the data from COHA presented in Figure 2.

Figure 2: The total number of both *continue* constructions in the nineteenth century in COHA

46 Figure 2 shows that the comparative distribution of the two constructions remained relatively stable throughout the nineteenth century, which is in line with the evidence of the first and third periods of CLMET. This case study underlines the need for researchers to note the danger that diachronic corpora may not be consistent in balancing texts from the various periods included in the corpus. The researcher must be aware that appearances can be deceptive!

4. Balanced representativeness in multi-lingual corpora

47 In section 3.5 we saw how a researcher can be misled when confronted with a corpus which does not contain a strict balance between sets of texts taken from different historical periods. The same point may be made with respect to selections of texts from different contemporary varieties, and selections of texts taken from different languages. According to Leech (2007: 142) : "The requirement of comparability depends at least partly on that of representativity: comparable corpora permit precise comparisons between two varieties or states of a language, but only if the corpora are reasonably representative of their respective varieties". The present section is concerned with problems related to the comparison of corpus data from two or more languages, and with two solutions that have been proposed to address these problems. Section 4.1 describes briefly the problems involved and sections 4.2 and 4.4 two proposed solutions, which are illustrated in 4.3 and 4.5 respectively.

4.1 The challenge of 2-text corpora

48 There are two main types of multilingual corpora, comparable corpora, consisting of original texts in two or more languages, and translation corpora, consisting of original texts in one language with their translations into two or more languages. The term 'parallel corpus' has been used in the past for both types, but the terminological differences appear to have been resolved, with the term 'parallel' being restricted nowadays to translation corpora (Aijmer 2008: 275, Borin 2002: 2, Kenning 2010: 487,

McEnery & Xiao 2008: 19). Both types of corpus pose problems for the researcher. No matter how careful compilers of comparable corpora are to ensure similarity of text types in various languages, there will always remain the dangers that, on the one hand, there is a mismatch between the languages when it comes to genres and, on the other, that the texts chosen are not equally representative of these languages. When comparing translations with original texts, the question of representativeness is not equally urgent, since the one text is necessarily a mirror image of the other (Kenning 2010: 489, McEnery & Xiao 2008: 20). However, here the question of text types is even more pressing, since one is faced with the problem that these are fundamentally different, insofar as one set of texts may display translation effects, the other not. By translation effects (Johansson 1998: 5), or translationese (Gellerstam 1996: 54), are meant the retention in the target language texts of features of the source language that are not equally felicitous in the target. So prevalent are these features that, as has been demonstrated by Cappelle (2012), it is sometimes possible to use them to predict the original language of a translated text.

49 Any comparison between two or more features is dependent on the availability of a viable *tertium comparationis*. According to Johansson (2001: 584), "The advantage of a corpus of original texts and their translations is that the translation is intended to express the same meaning as the original text". Ebeling & Ebeling (2013: 21), however, express some reservations about relying on identity of meaning as a starting point for contrastive analysis, contending that: "One of the difficulties in starting with meaning is how to delimit it. Starting with form, the boundaries are already set, while meaning is much more elastic."

50 Underpinning the difficulty of establishing equivalence of meaning in two texts, one of which is a translation of the other, is the fact that that these two text types are produced subject to two different sorts of constraints. As researchers we only have access to one of these. That is, we can be reasonably sure that we know what the translator is trying to convey,[7] but we can only guess at the intentions of the author of the original text. The discrepancy is illustrated in Table 3, where the term '2-text corpora', borrowed from Krzeszowski (1990), is used for translation corpora containing original texts in one language and translations of these into another language.

Table 3: Sources and targets in 2-text translation corpora

	To be encoded	Encoded by
Translator	Expression in source text	Expression in target text
Original author	?	Expression in source text

51 We see in Table 3 that the expression in the source text occurs in two columns, in the second column as a prompt for the expression in the translated text and in the third as an utterance to be compared to the latter. The prompt in the third row, the content of which is represented by the question mark, is nebulous compared to its counterpart in the second row. Since the production of a meaningful utterance involves making a series of lexical and grammatical choices, it is important for the analyst to be aware of the parameters within which these choices are made. However, in the case of the original author, as opposed to the translator, the analyst is in the dark as to the exact nature of

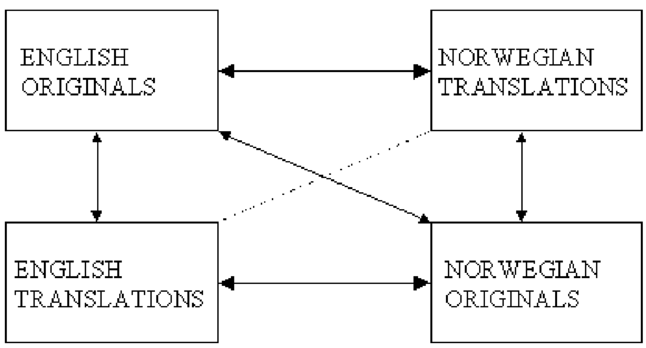
the prompts in question, having to reason backwards from their expression in the third column.

52 There are thus serious problems attached to drawing conclusions based on the contents of just two texts, whether they be from comparable (parallel) or translation corpora.[8] The next two sections contain descriptions of two strategies designed to tackle these problems.

4.2 Comparing original texts and translations: the 4-text solution introduced

53 One method of tackling the problems described in 4.1 involves the construction of corpora containing both original texts in two languages and their translations into the other language. This method was pioneered in the early 1990s by Stig Johansson, Karin Aijmer, Bengt Altenberg and Knut Hofland in their assembly of the English–Norwegian Parallel Corpus (ENPC) and the English–Swedish Parallel Corpus (ESPC).[9] Both these corpora contain extracts from fifty texts in English with their translations into Norwegian or Swedish, and fifty texts in Norwegian or Swedish with their translations into English. The original texts in the two languages thus constitute a comparable corpus, and each set of originals with its translations a parallel corpus. Care was taken by the compilers to ensure that the original texts in the pairs of languages were as comparable as possible. The arrows in Figure 3 illustrate various possible avenues of comparison.

Figure 3: The structure of the English-Norwegian Parallel Corpus (see Johansson 2007: 11)



54 The fact that there are various avenues of comparison allows for conclusions based on a comparison of original texts to be checked against translated texts and vice versa. The structure presented in Figure 3 can be expanded to include a third language, which would then include nine sub-corpora, three of original texts and six of translations, or even a fourth, with sixteen sub-corpora (Johansson 2002: 49). However, despite the value of such corpora to the researcher, the problems involved in locating suitable texts and translations into more than two or three languages, and obtaining permission to include these in a corpus, renders their compilation extremely challenging, if not completely impractical.

4.3 The 4-text solution exemplified: GIVE constructions in English and Norwegian

55 In the following I illustrate some of the possibilities provided by a 4-text corpus with some details from a case study of the cognate ditransitive verbs *give* in English and *gi* in Norwegian (Egan forthcoming).

56 The two verbs *give* and *gi* both occur in a ditransitive construction, as in (31) and a prepositional object construction (32).

(31) Macon *gave* her a credit card. (AT1)[10]
Macon *ga* henne et kredittkort. (AT1T)
(32) Jeg *ga* romnøklene til resepsjonisten. (LSC2)
I *gave* the room keys to the desk clerk. (LSC2T)

57 The corpus investigation was designed to address the following three research questions:

- How similar/different to one another are the distributions of the ditransitive and prepositional constructions containing the verbs *give* and *gi* in the original texts in the two languages ?
- Are there some kinds of tokens that are usually translated by syntactically congruent constructions ? What characterises these ?
- Are there some kinds of tokens that are seldom translated by syntactically congruent constructions ? If so, what characterises these tokens ? What characterises the syntactically divergent translations ?

58 The study was restricted to active tokens of both verbs with three explicit participants, a subject (except in the case of imperatives) and two objects, a direct object and either an indirect or a prepositional object. There were 381 such tokens in the original English texts and 435 in the original Norwegian texts. The entity coded by the direct object is referred to as the THEME, and the entity coded by the indirect or prepositional object as the RECIPIENT. The tokens categorised as instances of the ditransitive, labelled ‘OO’, and instances of the prepositional dative, labelled ‘OP’.

59 The tokens were classified according to the following four binary semantic distinctions.

- All subjects were classified as Agentive (*Ag*) or non-agentive, labelled ‘*St*’ for Stimulus.
- All verbs were classified as either encoding Transfer (*Tr*) or not (*NTr*).
- All $_{\text{recipients}}$ were classified as either animate (*An*) or inanimate (*In*).
- All $_{\text{themes}}$ were classified as either concrete (*C*) or abstract (*Ab*).

60 Eleven of the sixteen possible combination of these features are attested in English and twelve in Norwegian. Totals for all twelve attested combinations are given in Table 4. Note that the RECIPIENT is always listed before the THEME in the semantic classification: thus *Ag-Tr-An-CK* also subsumes *Ag-Tr-C- to An.*

Table 4: Numbers of tokens of various semantic combinations encoded as OO and OP in the original English and Norwegian texts in ENPC

Semantics	English		Norwegian	
	OO	OP	OO	OP
Ag-Tr-An-C	78	23	104	22

Ag-Tr-An-Ab	29	4	31	4
Ag-Tr-In-C	0	1	0	2
Ag-Ntr-An-C	67	3	30	3
Ag-Ntr-An-Ab	41	1	68	10
Ag-Ntr-In-C	13	1	1	0
Ag-Ntr-In-Ab	14	11	17	5
St-Ntr-An-C	23	1	23	8
St-Ntr-An-Ab	60	2	70	2
St-Ntr-In-C	1	2	1	2
St-Ntr-In-Ab	4	2	24	6
St-Tr-An-C	0	0	2	0
Total	330	51	371	64
Percentages	86.6 %	13.4 %	85.3 %	14.7 %

61 Even a cursory glance at Table 4 conveys the distinct impression that the overall distribution of the OO and OP constructions in the two languages is very similar. This is confirmed by statistical calculations (Pearson's chi.sq. with 1 df = 0.082, p =0.774637). Moreover, this similarity does not just hold for the more frequent senses in the GIVE network, such as *Ag-Tr-An-C* and *St-Ntr-An-Ab*, but also for the less frequent, more peripheral senses, such as *Ag-Tr-In-C* and *St-Ntr-In-C*.

62 The constructions that differ most in distribution in Table 4 are *Ag-Ntr-In-C*, with fourteen times as many tokens in English than in Norwegian, *Ag-Ntr-An-C*, with twice as many tokens in English than in Norwegian, and *St-Ntr-In-Ab*, with five times as many tokens in Norwegian than in English. The first two of these constructions both encode the agentive act of non-transfer of a concrete THEME to either an animate or an inanimate RECIPIENT. The concrete themes in the case of *Ag-Ntr-An-C* include many actions of smiling and looking, which are encoded in light verb constructions in English (give a smile, give a look), but not in Norwegian. Taken together three types of actions, looking, smiling and providing transport ('give a lift') account, between them, for some half of concrete THEME s in the original English texts. Similarly in the *Ag-Ntr-In-C* examples we find light verb constructions in English (give a push, give a stir) corresponding to constructions without the GIVE verb in Norwegian. As for *St-Ntr-In-Ab*, there do not appear to be any particular types of actions that can account for the greater incidence of these in the Norwegian texts.

63 I turn now to the evidence of the translations. Table 5 contains details of how often the various constructions are translated by syntactically congruent, as opposed to syntactically divergent constructions.

Table 5: Congruent translations of various sub-constructions encoded as OO and OP

Semantics	into Norwegian		into English	
	of OO	of OP	of OO	of OP
Ag-Tr-An-C	57 73 %	15 63 %	79 76 %	20 91 %
Ag-Tr-An-Ab	20 69 %	3 100 %	18 58 %	3 60 %
Ag-Tr-In-C	Ø	1 100 %	Ø	2 100 %
Ag-Ntr-An-C	30 45 %	1 33 %	24 80 %	2 67 %
Ag-Ntr-An-Ab	27 66 %	0 0 %	33 49 %	7 70 %
Ag-Ntr-In-C	3 23 %	0 0 %	0 0 %	Ø
Ag-Ntr-In-Ab	7 50 %	6 55 %	9 53 %	2 40 %
St-Ntr-An-C	12 52 %	1 100 %	14 61 %	7 88 %
St-Ntr-An-Ab	45 75 %	1 50 %	49 70 %	2 100 %
St-Ntr-In-C	1 100 %	0 0 %	1 100 %	2 100 %
St-Ntr-In-Ab	3 75 %	1 50 %	14 58 %	2 33 %
St-Tr-An-C	Ø	Ø	2 100 %	Ø
Total	204 62 %	29 57 %	243 66 %	49 77 %

64 The data in Table 5 show that the overall behaviour of the two sets of translators is very similar, at least with respect to the translation of instances of OO, with the Norwegian translations employing the congruent OO in 62 % of all cases, and the English translators in 66 %. That the difference is slight is borne out by a chi. sq. test: Pearson's chi. sq. with one df = 1.024, p = 0.311588. As in Table 4, the distribution of the constructions in the two languages is not only similar in the case of what we may surmise to be the more central senses in the GIVE networks, but also for the more peripheral senses, such as *Ag-Tr-In-C* and *St-Ntr-In-C*. We may also note that the constructions that contain many light verbs in English, *Ag-Ntr-An-C* and *Ag-Ntr-In-C*, are those with the fewest congruent translations into Norwegian. These are the very constructions that display the most marked differences in distribution in the original texts in the two languages. The evidence of the translated texts thus buttresses that of the original texts and vice versa.

65 In this study I started by comparing the original texts in the two languages before looking at the evidence of the translations. It is actually more common to start by comparing original texts in language A with their translations, before comparing the translations in language B with original texts in that language. Whichever methodology one adopts, the 4-text structure designed by researchers in Norway and Sweden and illustrated here by

the case study of *give* and *gi* has been employed widely in contrastive studies over the last twenty five years, to the extent that it must now be considered the default corpus structure for conducting such studies. Another method of addressing the deficiencies of 2-text corpora is presented in the next section.

4.3 Comparing translations of same text: the 3-text solution

66 Another way of better ensuring comparability of corpus data is to use translations of the same source text into different languages. This method, like the 4-text method described in section 4.2, also dates from the 1990s, which saw the publication of studies comparing translations of one and the same text, such as Paulussen (1999). It was also in the 1990s that a team at the University of Oslo, under the direction of Stig Johansson, developed the Norwegian–English–French–German sub-part of the Oslo Multilingual Corpus, which consists of original Norwegian texts with their translations into the other three languages.

67 In the 3-text method the original texts function as a *tertia comparationis* for the material in the target texts (see Egan 2013, 2016b). A popular method of ensuring that the linguistic items being compared are produced under similar constraints is to provide informants with a *tertium comparationis* from another modality. In research into predications of location and motion, this *tertium comparationis* may take the form of drawings, picture books or video snippets (see, for instance, Berman & Slobin 1994). Instead of using pictures, still or moving, as prompts, we can use verbal texts. Translators are viewed as informants who are provided with prompts in the form of verbal rather than visual *tertia comparationis*. Table 6 summarises the process and the results of the translation endeavours.

Table 6. Sources and targets in 3-text translation corpora

	To be encoded	Encoded by
Translator 1	Expression (a) in source text	Expression in first target text
Translator 2	Expression (a) in source text	Expression in second target text

68 The expressions in the two target texts in Table 6 consist of translations and are therefore likely to be coloured by translationese (Gellerstam 1996: 54) or translation effects (Johansson 1998: 5). The translations may be expected to be influenced to some extent by what are called ‘translation universals’ (Borin 2002: 5, Kenning 2010: 494). For instance, one translation may instantiate simplification, while another may instantiate conventionalisation (Mauranen 2008). One may wonder whether the choice of two different translation pathways by translators can compromise the results of the comparison of the expressions in the two target languages. The simple answer to this question is that there is no reason why this should be the case. As long as the two translations conform qualitatively to the linguistic norms of the target languages, the very fact that the translators have chosen different options may be indicative of pertinent differences in the languages in question.[11]

4.4 The 3-text solution exemplified: some prepositional constructions in English and French

69 The 3-text approach may be exemplified, to begin with, by (33) and (34), which consist of [exit] predications produced by English and French translators in response to the Norwegian originals in the Norwegian–English–French–German part of the Oslo Multilingual Corpus (see Egan & Graedler 2015).

(33) a. Jeg åler meg *ut av* vinduet igjen. (NF1)[12]
b. Wriggling *through* the window.... (NF1TE)
c. Je me suis glissé à nouveau *par* la fenêtre. (NF1TF)
(34) a. Hun holdt hesten an da hun var kommet *ut av* den siste kløfta. (HW2)
b. When she rode *out of* the last crevice, she reined in her horse. (HW2TE)
c. Elle retint le cheval après avoir *passé* le dernier ravin. (HW2TF)

70 In (33) both the English and the French translator code the manner of motion in the verb and the path in an adverbial, thus preserving the coding options of the original text. In (34) on the other hand, in which the original text contains a neutral verb of motion and a path adverbial, the English translator employs a manner motion verb and a path adverbial and the French translator the path verb *passer* (pass). Note that the inclusion of the Norwegian originals in the examples is not for the purpose of comparing them to the translations, but rather to illustrate the common prompts to which the translators are exposed.

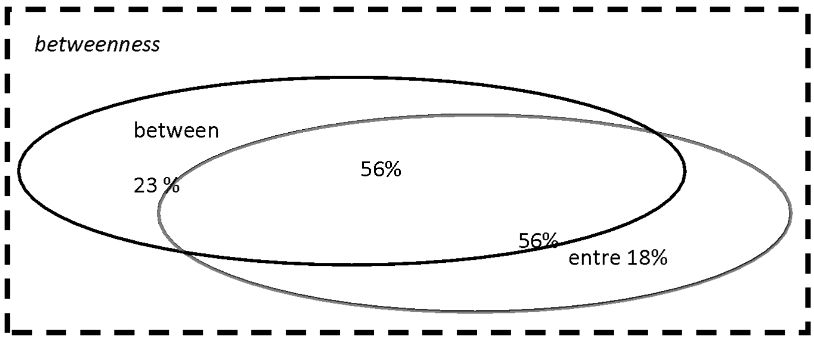
71 This sort of investigation of 3-text corpora has been used to contrast English and French correspondences of the Norwegian prepositions listed in Table 7.[13]

Table 7: Some contrastive studies of prepositions in the Oslo Multilingual Corpus

Norwegian preposition	Most common English correspondent	Most common French correspondent	Article(s)
mellom	between	entre	Egan (2013) Egan (2015b)
gjennom	through	à travers	Egan (2014) Egan (2015a) Egan (2015b)
over	over	sur	Egan & Rawoens (2013) Egan (2015b)
til	to	à	Egan (2015a)
ut av	out of	Path verb + par/de....	Egan & Graedler (2015)
inn i	Into	Path verb + dans	Egan & Graedler (2015)

72 There are big differences between the prepositions in Table 7 in the degree of similarity between the most common English and French correspondences. For instance, of a total of 393 tokens of Norwegian *mellom*, 74 % are translated into English by *between* and 69 % into French by *entre*, while of 313 tokens of Norwegian *gjennom,* 72 % are translated into English by *through*, but only 24 % into French by *à travers.*[14] The overlap between two English and French prepositions is illustrated in Figure 4, which shows the coverage by the two prepositions of the semantic space of [BETWEENNESS].

Figure 4: ***between*** **and** ***entre*** **used to translate Norwegian** ***mellom***



73 We see from Figure 4 that there is an overlap of 56 % between *entre* and *between* in tokens that translate Norwegian *mellom*. We can also use the 3-text corpus to work out the degree of total mutual correspondence between the English and French prepositions *between* and *entre* using a method based on Altenberg (1999). Altenberg's method involves the division of the total number of occurrences in target texts in a 4-text corpus of item *a* translating item *b* and vice versa by the total number of occurrences of both terms in the two sets of source texts. Multiplying the result of this calculation by 100 gives us the percentage overlap of the two items, which Altenberg labels their 'mutual correspondence'. We can adapt this method to 3-text corpora by replacing the total of *a* translating *b* and vice versa by the total number of mutual occurrences in the two translated texts of *a* and *b* multiplied by two. We have to multiply by two, since the correspondence is in both directions, i.e. we count the *a*s corresponding to *b*s plus the *b*s corresponding to *a*s. We then divide the result by the total number of tokens of both *a* and *b* in the two translated texts rather than the source texts, which by definition do not contain either item. There are a total of 365 tokens of *between* in the English translations in the OMC and 477 of *entre* in the French translations. 242 of these occur in parallel translations. Using the above formula, their degree of mutual correspondence in the translations, which we may label $MC_{t,}$ can be calculated as follows:

74 MC_t = (overlap *between/entre*)*2*100

75 total *between* + total *entre*

76 MC_t = (242)*2*100 = 59.61 %

77 365 + 477

78 This figure of 59.61 % may be compared to the total for mutual correspondence of the two items in translations of *mellom,* 56 %. These figures may, of course, also be compared to results of calculations using Altenberg's original formula on the correspondence of the two items in 4-text corpora. Needless to say, the greater the degree of correspondence

between results arrived at using different corpora and different calculations, the greater confidence we can have in our results.

79 Although the argument in this section has been illustrated with a contrastive study of expressions in just two target languages (hence the name '3-text corpus'), there is no reason why one cannot compare translations of the same source text into a greater number of languages. Indeed the corpus used here to compare expressions in French and English can also be used to compare one or both of these with German, since it also contains parallel translations into this language. Some scholars have made their own corpora containing translations into several languages. Among these are Åke Viberg who has compiled a corpus containing extracts from ten Swedish novels together with their translations into Finnish, French, German and English (Viberg 2012, 2013), Dan Slobin who compiled a corpus with one chapter of *The Hobbit* translated into ten languages (Slobin 2005), and Annemarie Verkerk whose corpus consists of predications of motion events in three literary works in two source languages, *Alice in Wonderland* and *Through the Looking Glass* and Paulo Coelho's *O Alquimista,* with their translations into twenty Indo-European languages (Verkerk 2014, 2015). The total number of texts in a multilingual corpus will depend on both the number of original texts that are translated, and the number of languages into which these originals are translated. A disadvantage with a very large number of target languages is the difficulty of identifying suitable texts and of obtaining permission from the various authors and publishers to include their products in the corpus. Moreover, if there is only one source text the likelihood is that each target language will only be represented by one translator, with the obvious danger of results being skewed by idiolectal factors.

5. Summary and conclusion

80 A lack of representativeness in a corpus may pose various sorts of challenges for the corpus linguist. Some of these have been described in this article. In section 2 I asserted that compilers of general, as opposed to specific, corpora should be wary of including certain text types. Dialogue in historical fiction is a good example of problematic content, since it can seriously misrepresent the contemporary state of the language, at least from the perspective of language production (see section 3.4). If the researcher, on the other hand, is primarily interested in reception or comprehension, it makes sense to include not only historical fiction, but also actual older fiction that is still widely read, such as the novels of Austen, or Dickens or George Eliot. It may also make sense to include experimental fiction or poetry, the authors of which may twist lexis or syntax in ways one would be unlikely to encounter in spontaneous speech. However, it is probably true to say that most linguists are interested in exploring the synchronic state of a language or, in the case of diachronic studies, a succession of such synchronic states. It is language production, rather than reception, that provides us with the more reliable guide to such states.

81 A second type of challenge is related to the notion of comparability. This was illustrated in section 3.5 which charted the development of non-finite complements of the verb *continue* in LModE as reflected in the texts in CLMET. It was shown that the results of the corpus investigation were seriously skewed because of a lack of balance between the text types in the three sub-corpora. Particularly challenging in this respect is the nature of a text such as *The Voyage of the Beagle*, which is part travelogue and part scientific treatise.

82 The question of comparability is acute when compiling and analysing multilingual corpora. In section 4.1 I argued that both comparable corpora of original texts and translation corpora pose challenges when it comes to representativeness. Translations, in particular, bearing as they often do traces of the source language, are unlikely to be completely representativeness of the target language in general. Sections 4.2 –4.5 describe two attempts to address these challenges, the four text model developed by Stig Johansson and his associates, and the three text model involving the comparison of translations into different languages of one and the same source text. Which method one chooses to adopt will, of course, depend on one's research question.

83 To round off, I would like to point out that the reservations expressed in this paper about corpus data should not be taken as criticism of corpus methodology. I believe that if corpus data is available, one should always use it. In section 2 I referred to a discussion on the Linguist List, in which various participants debated the grammaticality of *The rat that the cat that the dog bit chased ran*, and during the course of which I wrote:

> Until I see more evidence to the contrary I think I'll stick with my intuition that "the rat" [sentence] just isn't English. I am emboldened in this stance by the asterisking of such a sentence in Quirk *et al.* (1985: 1040). I am, nevertheless, open to convincing. I've had enough experience of being led astray by intuition (especially when coloured by stylistic preferences) to be more than willing to give way in the face of more evidence of actual usage. However, until "the rat..." has been proven to be English, surely any discussion of its grammaticality is a mite premature. (Egan 2000: 11.322)

84 In the nineteen years that have passed since this debate, I have had more experiences of being led astray by intuition. But I have also had experiences where my intuition saved me from being led astray by corpus data. If the corpus data smell funny, there may be something wrong with the corpus ; there may be something wrong with your sense of smell ; there may even be something wrong with both ! In which case, it would be an advantage if one also had experimental data to consult.

BIBLIOGRAPHY

Aijmer, Karin. 2008. Parallel and comparable corpora. In Anke Ladling and Merja Kytö (eds), *Corpus Linguistics. An International Handbook, Volume 1*. Berlin: De Gruyter, 275–292.

Altenberg, Bengt. 1999. Adverbial connectors in English and Swedish: Semantic and lexical correspondences. In Hilde Hasselgård and Signe Oksefjell (eds), *Out of corpora. Studies in honour of Stig Johansson*. Amsterdam: Rodopi, 249–268.

Aston, Guy and Lou Burnard. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Berman, Ruth and Dan Slobin. 1994. *Relating Events in Narrative: a Crosslinguistic Developmental study.* Hillsdale, NJ: Laurence Erlbaum.

Biber, Douglas. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8 :4 : 243–257, Oxford: Oxford University Press. Reprinted in Geoffrey Sampson and Diana McCarthy (eds), (2004) *Corpus Linguistics: Readings in a Widening Discipline*. London: Continuum.

Bolinger, Dwight. 1974. Concept and percept: two infinitive constructions and their vicissitudes. In *World Papers in Phonetics: Festschrift for Dr. Onishi's Kizyu.* Tokyo: Phonetic Society of Japan, 65–91.

Borin, Lars. 2002. ... and never the twain shall meet ? In Lars Borin (ed.), *Parallel corpora, parallel worlds: Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999*. Amsterdam: Rodopi, 1–43.

Bowie, Jill and Sean Wallis. 2016. The *to*-infinitival perfect: a study of decline. In Valentin Werner, Elena Seoane and Cristina Suárez Gómez (eds), *Re-assessing the present perfect: corpus studies and beyond.* Berlin: De Gruyter, 43–94.

Cappelle, Bert. 2012. English is less rich in manner-of-motion verbs when translated from French. *Across Languages and Cultures* 13:2 : 173–195.

Carnie, Andrew. 2000. *The Linguist List* 11.57.

Duffley, Patrick J.. 1992. *The English infinitive*. London: Longman.

Ebeling, Jarle. 2016. Linguistics in a new key. *The Nordic Journal of English Studies* 15 :3 : 7-14.

Ebeling, Jarle and Signe Oksefjell Ebeling. 2013. *Patterns in Contrast*. Amsterdam: John Benjamins.

Egan, Thomas. 2000. *The Linguist List* 11.322.

Egan, Thomas. 2008. *Non-finite complementation: a usage based study of infinitive and -ing clauses in English*. Amsterdam: Rodopi.

Egan, Thomas. 2010. The 'fail *to*' Construction in Late Modern and Present-day English. In Ursula Lenker, Judith Huber and Robert Mailhammer (eds), *English Historical Linguistics 2008. Selected papers fom the fifteenth International Conference on English Historical Linguistics. Volume 1: The history of English Verbal and Nominal Constructions.* Amsterdam: John Benjamins, 123–41.

Egan, Thomas. 2011. Perception and Conception: the 'see x to be y' construction from a cognitive perspective. In Doris Schönefeld (ed.), *Converging Evidence: Methodological and theoretical issues for linguistic research*. Amsterdam: John Benjamins, 57–80.

Egan, Thomas. 2012. *Prefer*: the odd verb out. In Irén Hegedűs and Alexandra Fodor (eds), *English Historical Linguistics 2010, Selected papers from the sixteenth International Conference on English Historical Linguistics.* Amsterdam: John Benjamins, 203–216.

Egan, Thomas. 2013. *Tertia comparationis* in multilingual corpora. In Karin Aijmer and Bengt Altenberg (eds), *Advances in Corpus-based Contrastive Linguistics: Studies in Honour of Stig Johansson.* Amsterdam: John Benjamins, 7–24.

Egan, Thomas. 2014. Encoding *throughness* in English and French. In Alejandro Alcaraz Sintes and Salvador Valera Hernandez (eds), *Diachrony and Synchrony in English Corpus Studies*. Frankfurt: Peter Lang, 233–257.

Egan, Thomas. 2015a. Motion *to* and motion *through*: Evidence from a multilingual corpus. In Philip Shaw, Britt Erman, Gunnel Melchers and Peter Sundkvist (eds), *From Clerks to Corpora: Essays on the English Language Yesterday and Today*. Stockholm: Stockholm University Press, 285–302

Egan, Thomas. 2015b. Manner and Path: Evidence from a multilingual corpus. *CogniTextes* 12. https://cognitextes.revues.org/788.

Egan, Thomas. 2016a. The subjective and intersubjective uses of 'fail to' and 'not fail to'. In Hubert Cuyckens, Lobke Ghesquière and Daniel van Olmen (eds), *Aspects of Grammaticalization: (Inter)subjectification and Pathways of Change.* Berlin: De Gruyter, 167–196.

Egan, Thomas. 2016b. Contrasting translations. *Nordic Journal of English Studies* 15:3 *:* 80–101.

Egan, Thomas. forthcoming. Giving in English and Norwegian: a contrastive perspective. In Melanie Röthlisberger, Eva Zehentner and Timothy Colleman (eds), *Ditransitive Constructions in Germanic Languages*. Amsterdam: John Benjamins.

Egan, Thomas and Anne-Line Graedler. 2015. Motion *into* and *out of* in English, French and Norwegian. *Nordic Journal of English Studies* 14 :1 : 9–33.

Egan, Thomas and Gudrun Rawoens. 2013. Moving *over* in(to) French and English : A translation-based study of 'overness'. *Languages in Contrast* 13:2 : 193–211.

Egan, Thomas and Gudrun Rawoens. 2014. *Amid(st)* and *among(st)*: A contrastive approach. In Caroline Gentens, Ditte Kimps, Lieven Vandelanotte and Kristin Davidse (eds), *Advances in Corpus Linguistics: Compilation and Applications.* Amsterdam: Rodopi, 207–228.

Egan, Thomas and Gudrun Rawoens. 2017. LOCATIVE *at* seen through its Swedish and Norwegian equivalents. In Thomas Egan and Hildegunn Dirdal (eds), *Cross-linguistic correspondences: from lexis to genre*. Amsterdam: John Benjamins, 121–145.

Fanego, Teresa. 1996. The development of gerunds as objects of subject-control verbs in English (1400-1760). *Diachronica* 13:1 : 29–62.

Fanego, Teresa. 2004. On reanalysis and actualization in syntactic change: the rise and development of English verbal gerunds. *Diachronica* 21 :1 : 5–55.

Gaines, Phil. 2000. *The Linguist List* 11.269.

Gellerstam, Martin. 1996. Translations as a source for cross-linguistic studies. In Karin Aijmer, Bengt Altenberg and Mats Johansson (eds), *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies, Lund, 4–5 March 1994.* Lund: Lund University Press. 53–62.

Hamman, Marc. 2000. *The Linguist List* 11.109.

Horie, Kaoru. 2000. Complementation in Japanese and Korean: A contrastive and cognitive linguistic approach. In Kaoru Horie (ed.), *Complementation: cognitive and functional perspectives*. Amsterdam: John Benjamins, 11–31.

Jespersen, Otto. 1940. *A modern English grammar: on historical principles*. Copenhagen: Munksgaard.

Johansson, Stig. 1998. On the role of corpora in cross-linguistic research. In Stig Johansson and Signe Oksefjell (eds), *Corpora and Cross-linguistic Research: Theory, Method and Case Studies*. Amsterdam: Rodopi, 3–24.

Johansson, Stig. 2001. The German and Norwegian correspondences to the English construction type *that's what*. *Linguistics* 39(3): 583–605.

Johansson, Stig. 2002. Towards a multilingual corpus for contrastive analysis and translation studies, In Lars Borin (ed.), *Parallel corpora, parallel worlds: Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999.* Amsterdam: Rodopi, 47–59.

Johansson, Stig. 2007. *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies.* Amsterdam: John Benjamins.

Kenning, Marie-Madeleine. 2010. What are parallel and comparable corpora and how can we use them ? In Anne O'Keefe and Michael McCarthy (eds), *The Routledge Handbook of Corpus Linguistics.* Abindgdon: Routledge, 487–500.

Krzeszowski, Tomasz P.. 1990. *Contrasting Languages: The Scope of Contrastive Linguistics*. Berlin : Mouton de Gruyter.

Leech, Geoffrey. 2007. New Resources, or Just Better Old Ones? The Holy Grail of Representativeness. In Marianne Hundt, Nadja Nesselhauf and Carolin Biewer (eds), *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 133–149.

Mauranen, Anna. 2008. Universal Tendencies in Translation. In Gunilla Anderman and Margaret Rogers (eds), *Incorporating Corpora: The Linguist and the Translator*. Clevedon: Mulltilingual Matters Ltd., 32–48.

McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, Tony and Richard Xiao. 2008. Parallel and Comparable Corpora: What is Happening ? In Gunilla Anderman and Margaret Rogers (eds), *Incorporating Corpora: The Linguist and the Translator*. Clevedon: Mulltilingual Matters Ltd., 18–31.

Newmeyer, Frederick. 1998. *Language form and language function*. Cambridge MA: MIT Press.

OED (1994). *The Oxford English dictionary*. On compact disc. Oxford: Oxford University Press.

Paulussen, Hans. 1999). *A corpus-based contrastive analysis of English on/up, Dutch op and French sur within a cognitive framework*. Unpublished doctoral dissertation, Ghent University.

Quirk, Randolph, Sidney Greenbaum, Geoffrey leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Rawoens, Gudrun and Thomas Egan. 2015. Coding *betweenness* in Swedish and Norwegian translations from English. *Meta* 60 :3 : 576–598.

Rohdenburg, Günter. 2006. The Role of Functional Constraints in the Evolution of the English Complementation System. In Christiane Dalton-Puffer, Dieter Kastovsky, Nikolaus Ritt and Herbert Schendl (eds), *Syntax, Style and Grammatical Norms: English from 1500-2000.* Bern: Peter Lang, 143–166.

Rudanko, Juhani. 2000. *Corpora and complementation: tracing sentential complementation patterns of nouns, adjectives and verbs over the last three centuries.* Lanham, Md.: University Press of America.

Schmid, Hans-Jörg and Annette Mantlik. 2015. Entrenchment in Historical Corpora: Reconstructing Dead Authors' Minds from their Usage Profiles. *Anglia* 133:4 : 583–623.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Slobin, Dan. 2005. Relating Narrative Events in Translation. In Dorit Diskin Ravid and Hava Bat-Zeev Shyldkrot (eds), *Perspectives on language and language development: Essays in honor of Ruth A. Berman.* Dordrecht: Kluwer, 115–129.

Stocker, Bryony D.. 2012. 'Bygonese' – Is This Really the Authentic Language of Historical Fiction? *New Writing* 9 (3) : 308-318.

Verkerk, Annemarie. 2014. Where Alice fell into: Motion events from a parallel corpus. In Benedikt Szmrecsanyi and Bernhard Wälchli (eds), *Aggregating dialectology, typology, and register analysis : Linguistic variation in text and speech.* Berlin: Walter de Gruyter, 324–354

Verkerk, Annemarie. 2015. Where do all the motion verbs come from? The speed of development of manner verbs and path verbs in Indo-European. *Diachronica* 32 (1) : 69–104.

Viberg, Åke. 2012. Language-specific meanings in contrast: A corpus-based contrastive study of Swedish *få* 'get'. *Linguistics* 50 :6 : 1413–1461.

Viberg, Åke. 2013. Seeing the lexical profile of Swedish through multilingual corpora: The case of Swedish *åka* and other vehicle verbs. In Karin Aijmer and Bengt Altenberg (eds), *Advances in corpus-based contrastive linguistics. Studies in honour of Stig Johansson.* Amsterdam: John Benjamins, 25–56.

Vinay, Jean-Paul and Jean Darbelnet. 1995. *Comparative Stylistics of French and English : A methodology for translation.* Translated and edited by Juan C. Sager and Marie-Josée Hamel. Amsterdam: John Benjamins.

Vosberg, Uwe. 2003. The role of extractions and horror aequi in the evolution of -ing-complements in Modern English. In Gunther Rohdenburg and Britta Mondorf (eds), *Determinants of Grammatical Variation in English.* Berlin: Mouton de Gruyter, 305–327

Corpora

British National Corpus. 2001. Oxford: Oxford University Computing Services.

CEECS : *Corpus of Early English Correspondence Sampler.* 1998. *The ICAME Corpus Collection on CD-ROM*, version 2 (1999): Bergen: Aksis.

Corpus de la Littérature Médiévale des Origines au 15e Siècle. http://www.classiques-garnier.com/

The Corpus of Contemporary American English. http://corpus.byu.edu/coca/

The Corpus of Historical American English. http://corpus.byu.edu/coha/

CLMET : Corpus of Late Modern English Texts. De Smet, Hendrik. 2005. A corpus of Late Modern English texts. ICAME Journal 29. 69–82. https://perswww.kuleuven.be/~u0044428/clmet.htm

The English–Norwegian Parallel Corpus. http://www.hf.uio.no/ilos/english/services/omc/

The English–Swedish Parallel Corpus. http://www.sol.lu.se/engelska/corpus/corpus/espc.html

Helsinki Corpus of English Texts. 1996. *The ICAME Corpus Collection on CD-ROM*, version 2 (1999). Bergen : Aksis.

The Oslo Multilingual Corpus. http://www.hf.uio.no/ilos/english/services/omc/

NOTES

1. I would like to thank the editors for inviting me to contribute to this special edition of *Cognitextes* and two anonymous reviewers for their helpful comments and suggestions.

2. Richardson's original version is as follows: "I beseech you say not one word but Yes or No, to my questions, till I have said all I have to say". It is quite possible to interpret the form *say* here as an imperative, rather than a *bare infinitive*. The classification of (4) as instantiating the *bare infinitive* construction was dictated by the orthography.

3. See Egan (2016a) for a discussion of the notion of intersubjectivity in relation to these constructions.

4. The story was published in a collection of short stories named *The Sea Witch* after the first story in the volume and edited by Ballou, who actually wrote none of the stories himself.

5. For an illustration of the extent of individual differences in the employment of another construction, the 'N+BE+*that*' construction, see Schmid & Mantlik (2015).

6. De Smet has produced two subsequent versions of the corpus, the latest one, CLMET3.0 containing some 30 million words, i.e. three times as many as the original CLMET. For details see https://perswww.kuleuven.be/~u0044428/clmet.htm

7. According to Vinay and Darbelnet (1995 : 30) : 'Translators are [...] faced with a fixed starting point, and as they read the message, they form in their minds an impression of the target they want to reach'.

8. For more on these problems see Aijmer (2008 : 277), Borin (2002 : 2), Kenning (2010 : 496), McEnery & Xiao (2008 : 21).

9. See Ebeling (2016) for background details.

10. The first part of the code 'AT1' refers to the text in the English – Norwegian parallel Corpus from which the example has been taken, with 'AT' being the initials of the author. 'AT1T' stands for translation of the same text. The full titles of the original works and the translations in the corpus are listed in Johansson (2007 : 329-338).

11. It is less likely that they will conform quantitatively ; differences in frequency of at least some items in original and translated texts in the same language are predicted by the unique items hypothesis (Mauranen 2008, Cappelle 2012).

12. The texts are from the Norwegian–English–French–German sub-part of the Oslo Multilingual Corpus. 'TE' and 'TF' stand for English and French translated text, respectively.

13. The fact that there are many identical English source texts in the ENPC and the ESPC has allowed for similar investigations of Norwegian and Swedish renderings of the English prepositions *amid(st)* and *among(st)* (Egan & Rawoens 2014), *between* (Rawoens & Egan 2015) and *at* (Egan & Rawoens 2017).

14. The degree of similarity between English and French encodings of [THROUGHNESS] differs greatly according to the semantic domain : there is thus considerable overlap between *through* and *à travers* in the case of perceptual predications, but none whatsoever in the case of temporal predications (Egan 2014).

ABSTRACTS

This article presents and discusses some problems of representativeness that the author has encountered in over twenty years of corpus-based research. It argues that the inclusion in a general corpus of certain text types, such as grammar treatises or works of historical fiction, can lessen the representativeness of the data, especially if the corpus is designed to reflect the linguistic production, as opposed to the linguistic reception, of a speech community. It is argued that less emphasis should be placed on reception in the compilation of general corpora. Also addressed are problems relating to the comparison of texts in different languages, as well as two solutions that have been proposed to counter these problems. The arguments are illustrated with examples from both contemporary and historical corpora.

Cet article présente et discute quelques-uns des problèmes de représentativité rencontrés par l'auteur au cours de plus de vingt ans de recherche basée sur corpus. Il démontre que l'inclusion dans un corpus général de certains types de texte, tels que les traités grammaticaux ou les oeuvres de fiction historique, peuvent nuire à la représentativité des données, surtout si le corpus vise à refléter la production linguistique, par opposition à la réception linguistique, d'une communauté linguistique donnée. L'article défend l'idée qu'il faudrait donner moins d'importance à la réception dans la construction de corpus généraux. Il aborde aussi des problèmes liés à la comparaison de textes dans différentes langues et présente deux solutions qui ont été proposées pour surmonter ces problèmes. Les différents aspects traités sont illustrés par des exemples tirés de corpus aussi bien contemporains que historiques.

INDEX

Mots-clés: représentativité, réception, langue archaïque, ditransitifs, tertium comparationis
Keywords: representativeness, reception, bygonese, ditransitives, tertium comparationis

AUTHOR

THOMAS EGAN

Inland Norway University of Applied Sciences