TRADI IMT (XX-XXI): Recent proposals for the alignment of a diachronic parallel corpus

Judit Martínez Magaz
University of León (Spain)

1 Introduction

In the course of the last decades, linguistic corpora have usually been compiled in an electronic format, which allows the processing of large amounts of data to be combined with a detailed analysis of the individual texts (*cf.* McEnery and Wilson 1996: 117–128; Sánchez 2001: 11–45). Electronic corpora have been improved so much that our knowledge of the different linguistic aspects is increasing everyday. However, corpus typology is not homogeneous at all.

Today, electronic advances are developed at high speeds, allowing for the creation of new textual combinations and new types of corpora (*cf.* Sinclair 1995; Ramón García 2001). Multilingual corpora comprise two groups: comparable[1] and parallel corpora. In a broad sense, a parallel corpus is formed by original texts in language A and their translations (and pseudo-translations) in language B (*cf.* Teubert 1996; Borin 2002). These texts cover the same meanings and have identical functions in both languages. Parallel corpora offer many different possibilities when comparing two or more languages, and together with comparable ones, parallel corpora have become an important tool for the comparison of different languages. Among other various uses,

> [...] the most common and also most well-known uses of parallel and comparable corpora are: (1) for contrastive and typological grammatical and lexicographical studies in linguistics, (2) for knowledge acquisition for machine translation in computational linguistics, and, (3) as a source of authentic contrastive language data in language learning and teaching. (Borin 2002: 14)

This paper focuses on the design and construction of a new English–Spanish parallel corpus: *TRADI IMT (XX-XXI)*.

2 TRADI IMT (XX-XXI): *A dialectal diachronic parallel corpus*

TRADI IMT (XX-XXI) is a new dialectal diachronic parallel corpus formed by a subcorpus A of English literary texts dating from the Early Modern English period (henceforth EModE) – 16th and 17th centuries – and a subcorpus B with the Spanish translations of those aforementioned EModE texts coming from the 20th and 21st centuries. What makes this type of corpus special is the fact that the English texts were selected to contain fragments written in the different regional varieties ('dialects') spoken in England in the EModE period.

Therefore, the dialectal diachronic parallel corpus is called *TRADI IMT (XX-XXI)*: *TRAducción de DIalectos* ('Translation of Dialects'), *IMT* (*Inglés Moderno Temprano*: Early Modern English, source text period), *XX-XXI* (20th–21st centuries, target text period). This corpus will become the tool that will allow for a descriptive study of the English regional varieties present in EModE texts and the different solutions taken by the Spanish translators when facing those dialectal features.

A careful selection of the texts is one of the main issues in the compilation of any corpus, bilingual or monolingual. In this particular case, the availability of Spanish translations has determined the path to be followed for the final selection of both English and Spanish texts. From a wide selection of EModE texts presenting regional varieties, only those which have been translated into the Spanish language are of interest for our purposes. Those are mainly theatrical plays by William Shakespeare, Ben Jonson and Thomas Middleton, in which certain characters present dialectal features from different English geographical areas in their speeches.

According to prior research carried out in order to build the *Catalogue of Translations*,[2] thorough statistical results have been achieved proving that, as expected, Shakespeare has been the English author most translated into the Spanish literature. Multiple translations have been found from this author: 367 entries out of 389 total entries of the *Catalogue* correspond to Shakespeare's translations into the Spanish language, while only 18 and 4 translations have been found from Jonson and Middleton, respectively. Such a difference in the number of target texts (henceforth TTs) available is mainly due to the fact that some of Shakespeare's most famous plays, such as *King Lear* or *Macbeth*, have been translated on multiple occasions.

The previous elaboration of the *TRADI IMT (XX-XXI) Catalogue of Translations* has been followed by a series of decisions taken in the later compilation of the corpus. Thus, the *TRADI IMT (XX-XXI)* corpus is formed by all the Source Text (henceforth ST) fragments written in non-standard English regional varieties, and only those fragments of TTs from the last published editions of each of the STs. Acknowledging the numerous cases of re-editions and re-printings of the STs into Spanish, only the most recent editions of each text have been selected for the corpus, in case of possible later revisions by the translator.

In view of the above considerations, the *TRADI IMT (XX-XXI)* corpus is defined as:

1. an open corpus, because it will be possible to add new fragments of TTs as new translations are discovered;
2. a corpus constituted by fragments of texts, as the focus is put on the fragments written in the different English regional varieties.

The compilation of the corpus in electronic format has required previous preparation of the available texts. All the English STs derive from *Literature Online Database*[3] (henceforth LION), an on-line library of English and North American poetic, prose and dramatic works. LION allows for the downloading of HTML versions of the first published editions of the EModE texts, avoiding modernized versions and reflecting the spelling and punctuation of that time. The Spanish TTs are in a printed format. For this reason, all of them have been converted into an electronic format by means of an optical scanner. This process has required a subsequent thorough review and correction of the texts, since in the majority of cases the scanner has been unable to recognise accents and other spelling features of the Spanish language.

Once STs and TTs are stored in the computer, and then subjected to a tagging process,[4] they are prepared for their respective alignment.

3 *Proposals for the alignment of* TRADI IMT (XX-XXI)

The process of alignment is a vital component in an electronic parallel corpus like *TRADI IMT (XX-XXI)*. A good corpus-based analysis will depend on the previous finding of an operative alignment unit. Although a parallel corpus need not be aligned as such, the most recent and advanced corpora of that type do appear aligned.[5] Text alignment is considered to be the crucial step for a forthcoming descriptive study of texts from different languages.

Corpus Presenter (Hickey 2000, 2003) is the software programme used in the electronic formatting of *TRADI IMT (XX-XXI)*; it allows the combination or alignment of two texts paragraph by paragraph. As long as ST and TT have the same number of paragraphs, the software programme aligns them automatically by displaying the whole text aligned in just one screen, numbering each paragraph as shown in Figure 1:

A1: Source text.
B1: Target text.

A2: Source text.
B2: Target text.

Etc.

Figure 1: System of alignment of TRADI IMT (XX-XXI)

However, the paragraph by paragraph automatic alignment provided by the programme is not enough for the purposes of the type of research presented in this paper. A thorough study of the presence of dialectal features in EModE texts and in the Spanish translators' strategies needs to account for every linguistic detail (e.g. spelling alterations), which makes clear the essentiality of the narrowest unit of alignment possible (and therefore, unit of comparison). That unit cannot be as wide as a paragraph for a detailed analysis.

At present, *TRADI IMT (XX-XXI)* is composed of theatrical plays, as no Spanish translation of EModE poetry or prose has been found to date. As a result, the paragraph by paragraph alignment coincides with the fragment of discourse uttered by each character. In *TRADI IMT (XX-XXI)* there are individual speeches of such a length that it would be very difficult to compare both ST and TT. Alternatively, there are discourses formed by one sentence or even by one word only. Hence, it is necessary to find an operative unit for our purposes.

The first decision has been to divide each character's dialogue into smaller units, taking into account proposals such as that of the *COMPARA Corpus* (*Portuguese – English Parallel Corpus*), whose unit of alignment is the source text orthographical sentence, understood as the fragment of discourse between two strong pauses, generally marked by a full stop (Frankenberg-García 2001). In principle, this type of proposal seems appropriate for *TRADI IMT (XX-XXI)* too.

However, it is necessary to bear in mind a very important issue: the focus of the research is on English texts dating from the 16th and 17th centuries. In fact, though punctuation was increasingly standardised in printed texts, EModE texts were still largely inconsistent. New punctuation marks appeared at that time, but their use was not at all clear for the English printers. For these reasons, the notion of 'orthographical sentence' is not self-evident in EModE texts; the use of the full stop, the colon or the semicolon is, in the majority of cases, completely arbitrary. Modern punctuation practices are different in this respect:

> Modern punctuation is uniform; the old punctuation was quite the reverse. It was natural that in the earlier stages of printing usage should be less settled, and it was certainly convenient for the printer. (Simpson 1969: 10)

Nevertheless, with a careful study of the fragments of the English texts comprising the corpus, it has been realised that strong pauses indicating the end of a sentence always appear in a coherent way by means of one of the following resources: stop, colon, semicolon and question mark. Their use seemed to be clear enough for the majority of the printers at that time.

The semicolon seems to have been introduced by chance in the English texts of that time by printer Richard Grafton, reappearing later in *The Scepter of Judah*, printed by John Wright in 1584 (Partridge 1964: 124). This punctuation mark was understood as a strong pause “to mark emphasis and to make the structure of the sentence clear” (Simpson 1969: 56).

In EModE, the colon was understood as a strong pause as well, even more emphatic than the semicolon: “… the colon is a stronger stop than the semicolon; indeed it is the function of the colon to mark an emphatic pause” (Simpson 1969: 67).

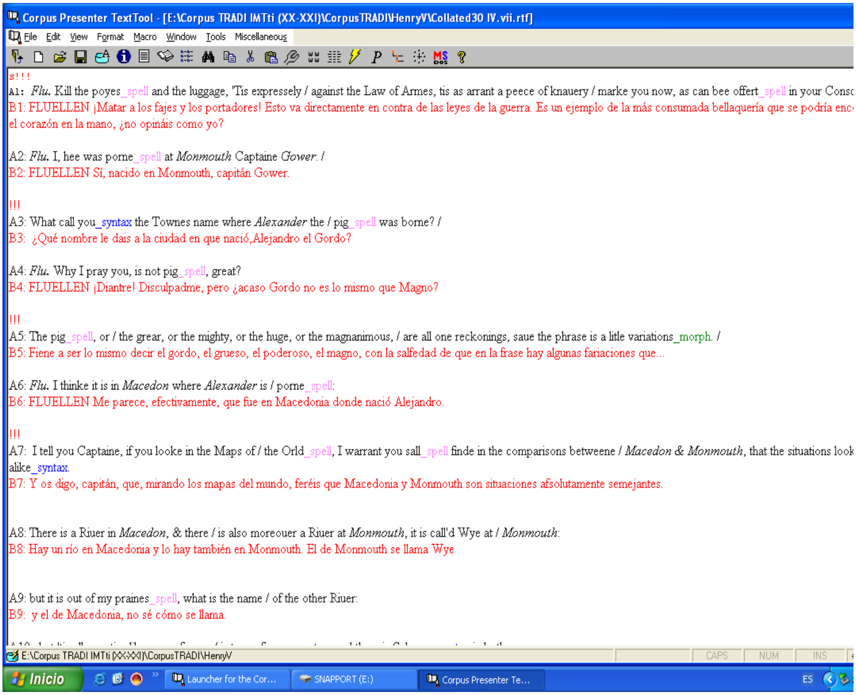
Moreover, we cannot overlook the fact that English texts (and specially those from Shakespeare’s First Folio) constitute a true summary of punctuation models: author, scribe, editor and compositor were only some of the figures who handled the texts before their final edition. Therefore, bearing in mind all the aforementioned facts, the criteria below have been followed to reach a final proposal:

1. the dialectal passages chosen for the research never present a systematic and consistent use of punctuation marks;
2. strong pauses seem to be marked by a full stop, semicolon, colon and question mark;
3. despite the length of the sentence between two strong pauses, one of the aforesaid punctuation marks always appears.

Thus, I decided to take the source text sentence between strong pauses (full stop, semicolon, colon, question mark) as the unit of alignment and comparison for *TRADI IMT (XX-XXI)*. In those cases in which there are no one-to-one correspondences between original and translation, the decision has been to follow the *COMPARA Corpus* criteria: the TT will be divided or joined in two or more sentences in order to adjust to the ST (Frankenberg-García 2001). Therefore, the corresponding target text can be one, two or more sentences, or even less than a sentence. Those ST units which are not translated have been aligned with blank units ('zero units'). Sentences or fragments which are added in the target text, and with no correspondence with the source text, are always included in the previous unit of alignment.

Following the aforementioned criteria, source texts will always be divided in the same way, and it will be possible to align the source sentence with multiple translations. One will be able to compare not only source and target texts, but also different target texts from the same source text.

Figure 2 shows how an aligned fragment of *TRADI IMT (XX-XXI)* appears on the computer screen.


Corpus Presenter TextTool - [E:\Corpus TRADI IMTti (XX-XXI)\CorpusTRADI\HenryV\Collated30 IV.vii.rtf]
File
Edit
View
Format
Macro
Window
Tools
Miscellaneous
MS DOS
E:\Corpus TRADI IMTti (XX-XXI)\CorpusTRADI\HenryV
CAPS
NUM
INS
Inicio
Launcher for the Cor...
SNAPPORT (E:)
Corpus Presenter Te...
ES

s!!!
A1: *Flu.* Kill the poyes_spell and the luggage, 'Tis expressely / against the Law of Armes, tis as arrant a peece of knauery / marke you now, as can bee offert_spell in your Consc
B1: FLUELLEN ¡Matar a los fajes y los portadores! Esto va directamente en contra de las leyes de la guerra. Es un ejemplo de la más consumada bellaquería que se podría enc
el corazón en la mano, ¿no opináis como yo?

A2: *Flu.* I, hee was porne_spell at *Monmouth* Captaine *Gower*: /
B2: FLUELLEN Sí, nacido en Monmouth, capitán Gower.

!!!
A3: What call you_syntax the Townes name where *Alexander* the / pig_spell was borne? /
B3: ¿Qué nombre le dais a la ciudad en que nació,Alejandro el Gordo?

A4: *Flu.* Why I pray you, is not pig_spell, great?
B4: FLUELLEN ¡Diantre! Disculpadme, pero ¿acaso Gordo no es lo mismo que Magno?

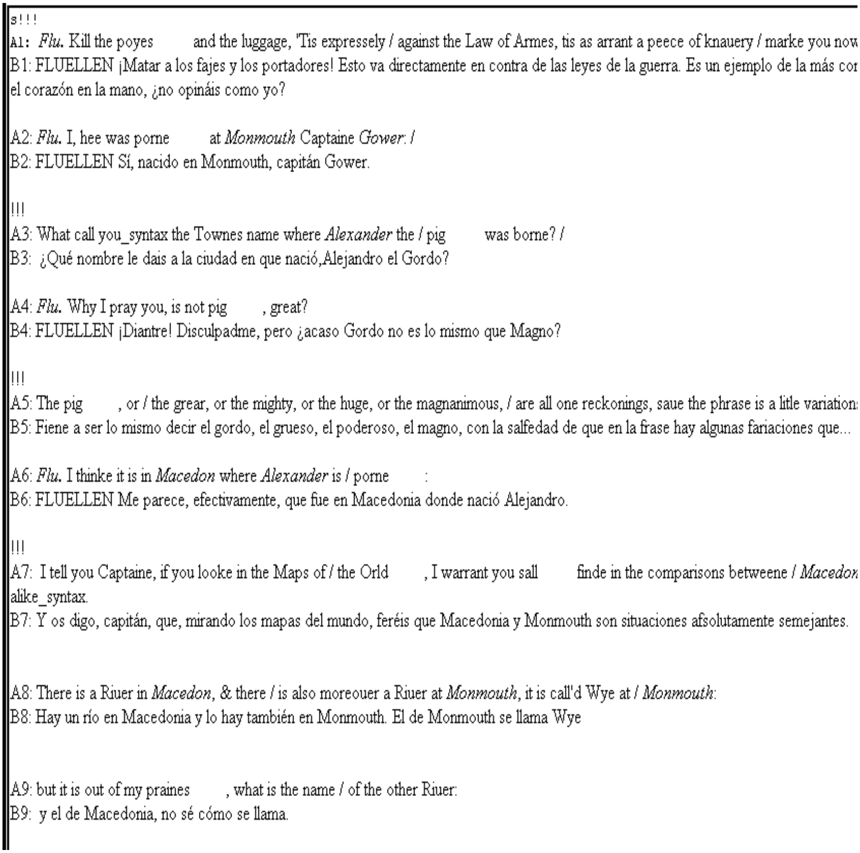
!!!
A5: The pig_spell, or / the grear, or the mighty, or the huge, or the magnanimous, / are all one reckonings, saue the phrase is a litle variations_morph. /
B5: Fiene a ser lo mismo decir el gordo, el grueso, el poderoso, el magno, con la salfedad de que en la frase hay algunas fariaciones que...

A6: *Flu.* I thinke it is in *Macedon* where *Alexander* is / porne_spell:
B6: FLUELLEN Me parece, efectivamente, que fue en Macedonia donde nació Alejandro.

!!!
A7: I tell you Captaine, if you looke in the Maps of / the Orld_spell, I warrant you sall_spell finde in the comparisons betweene / *Macedon* & *Monmouth*, that the situations look
alike_syntax.
B7: Y os digo, capitán, que, mirando los mapas del mundo, feréis que Macedonia y Monmouth son situaciones afsolutamente semejantes.

A8: There is a Riuer in *Macedon*, & there / is also moreouer a Riuer at *Monmouth*, it is call'd Wye at / *Monmouth*:
B8: Hay un río en Macedonia y lo hay también en Monmouth. El de Monmouth se llama Wye

A9: but it is out of my praines_spell, what is the name / of the other Riuer:
B9: y el de Macedonia, no sé cómo se llama.



s!!!
A1: *Flu.* Kill the poyes and the luggage, 'Tis expressely / against the Law of Armes, tis as arrant a peece of knauery / marke you now
B1: FLUELLEN ¡Matar a los fajes y los portadores! Esto va directamente en contra de las leyes de la guerra. Es un ejemplo de la más cor
el corazón en la mano, ¿no opináis como yo?

A2: *Flu.* I, hee was porne at *Monmouth* Captaine *Gower*: /
B2: FLUELLEN Sí, nacido en Monmouth, capitán Gower.

!!!
A3: What call you_syntax the Townes name where *Alexander* the / pig was borne? /
B3: ¿Qué nombre le dais a la ciudad en que nació,Alejandro el Gordo?

A4: *Flu.* Why I pray you, is not pig , great?
B4: FLUELLEN ¡Diantre! Disculpadme, pero ¿acaso Gordo no es lo mismo que Magno?

!!!
A5: The pig , or / the grear, or the mighty, or the huge, or the magnanimous, / are all one reckonings, saue the phrase is a litle variation
B5: Fiene a ser lo mismo decir el gordo, el grueso, el poderoso, el magno, con la salfedad de que en la frase hay algunas fariaciones que...

A6: *Flu.* I thinke it is in *Macedon* where *Alexander* is / porne :
B6: FLUELLEN Me parece, efectivamente, que fue en Macedonia donde nació Alejandro.

!!!
A7: I tell you Captaine, if you looke in the Maps of / the Orld , I warrant you sall finde in the comparisons betweene / *Macedon*
alike_syntax.
B7: Y os digo, capitán, que, mirando los mapas del mundo, feréis que Macedonia y Monmouth son situaciones afsolutamente semejantes.

A8: There is a Riuer in *Macedon*, & there / is also moreouer a Riuer at *Monmouth*, it is call'd Wye at / *Monmouth*:
B8: Hay un río en Macedonia y lo hay también en Monmouth. El de Monmouth se llama Wye

A9: but it is out of my praines , what is the name / of the other Riuer:
B9: y el de Macedonia, no sé cómo se llama.

Figure 2: A black-and-white snapshot of TRADI IMT (XX-XXI)

However, one must be aware of the fact that perfect and exact comparable units will never exist:

> Aligning source texts and translations is not a simple task, for translators do not always translate texts in a predictable and linear manner. Source-text sentences are sometimes divided into two or more sentences in the translation. Sometimes, translators join source-text sentences together, rendering them as a single translation sentence. In addition to this, translators may leave things out and insert elements which were not present in

> the source text, and sometimes they may reorder elements so that the order in which they appear in the translation differs from that in which they appear in the source text. (Frankenberg-García 2001: 3)

Still, the final decision to take the ST orthographical sentence as the unit of alignment and comparison has revealed itself to be very convenient for the type of research intended. Throughout a series of descriptive studies whose presence here would go beyond the scope of this paper, the suitability of the unit has already been confirmed.

By means of a unit of the type presented in this article, one is able to recognize three main strategies in the translation of written regional varieties:

1. trying to find a real diatopic variety in the target language (TL), although it is impossible to find equivalent varieties between two different languages;
2. translating into an imaginary dialect. This strategy involves the alteration of pronunciation features, or any other feature, just to mark a distance from the standard variety;
3. translating without taking into account the presence of a dialect in the ST; that is, translating as the standard variety, losing any type of dialectal feature in the TT: "Rendering ST dialect by TL standard has the disadvantage of losing the special effect intended in the ST" (Hatim & Mason 1990: 41).

By analyzing and comparing both STs and TTs in the corpus, it will be possible to find out which of the three aforesaid strategies is more likely to occur in the TTs.[6]

4 Final remarks

To carry out a descriptive study of the English dialectal features and the solutions taken by the Spanish translators, a new electronic corpus has been designed: *TRADI IMT (XX-XXI)*. This type of dialectal diachronic parallel corpus is made up of English literary texts from the Early Modern English period (16th and 17th centuries) and their translations into Spanish. The English texts selected for the corpus are those including passages written in the different English regional varieties of the time.

The finding of an appropriate and useful unit of alignment in an electronic parallel corpus like the one presented here is a crucial and difficult task, since punctuation in the EModE written texts was inconsistent. The different criteria followed in order to find an operative unit of alignment when dealing with texts dating from the 16th and 17th centuries and their corresponding translations from the 20th and 21st centuries will enable a later analysis and comparison of both STs and TTs.

The type of corpus presented throughout this paper will definitely allow identification of the solutions adopted by the translators when dealing with ST dialectal features. *TRADI IMT (XX-XXI)* will become a tool for a descriptive-comparative analysis by which we will be able to identify not only the type of dialectal features in the Early Modern English texts, but also the Spanish translators' norms and strategies when dealing with them.

It is important to bear in mind that the final emphasis of the intended research will be on the analysis and the results, never on the corpus itself. A corpus ought to be understood as a very useful tool, but one should not forget that it may have its own limitations and weaknesses. Still, an electronic parallel corpus has revealed itself as the most appropriate way to accomplish the task just undertaken here. In this sense, the first step is a prior search for an operative alignment unit, just as it has been described above.

Notes

1. For definitions of comparable corpora, cf. Baker (1995: 234); Ramón García (2001: 73); Sinclair (1995: 32).
2. *TRADI IMT (XX-XXI) Catalogue of Translations* is an electronic catalogue compiled in the *Windows Access* programme, comprising text details of the published available Spanish translations of the type of English texts mentioned (cf. Martínez Magaz 2004: 75-86).
3. LION – *Literature Online Database*: http://lion.chadwyck.com (only for subscribers).
4. The tagging process for *TRADI IMT (XX-XXI)* is described in Martínez Magaz (2004: 89–91).
5. "To *align* a text with a translation of it in another language is [...] to show which of its parts are translated by what parts of the second text" (Kay & Röscheisen 1993: 121).
6. No statistical results are presented in this study. The final conclusions of the research anticipated will appear soon in a forthcoming monograph by the same author.

References

Baker, Mona. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target* 7(2): 223–243.

Borin, Lars. 2002. ... and never the twain shall meet?. In L. Borin (ed.). *Parallel corpora, parallel worlds*, 1–43. Amsterdam / New York: Rodopi.

Frankenberg-García, Ana. 2001. *COMPARA*, the Portuguese–English Parallel Corpus. Article online: www.linguateca.pt/Repositorio/ Frankenberg-Garcia2001.doc

Hatim, Basil and Ian Mason. 1990. *Discourse and the translator.* London / New York: Longman.

Hickey, Raymond. 2000. Processing corpora with *Corpus Presenter*. *ICAME Journal* 24: 65–84. Also available at http://www.hit.uib.no/ icame/ij24/ hickey.pdf

Hickey, Raymond. 2003. *Corpus Presenter. Processing software for language analysis.* Amsterdam: Benjamins.

Kay, Martin and Martin Röscheisen. 1993. Text-translation alignment. *Computational Linguistics* 19(1): 121–142.

McEnery, Tony and Andrew Wilson. 1996. *Corpus linguistics.* Edinburgh: Edinburgh University Press.

Martínez Magaz, Judit. 2004. El catálogo y el corpus paralelo diacrónico dialectal TRADI IMTti (XX-XXI): propuestas, muestras y modelo de análisis. Facultad de Filosofía y Letras. León: Universidad de León. (Unpublished Licentiate thesis.)

Partridge, Astley C. 1964. *Orthography in Shakespeare and Elizabethan drama.* London: Edward Arnold.

Ramón García, Noelia. 2001. Estudio contrastivo inglés-español de la caracterización de sustantivos. León: Universidad de León. (Unpublished PhD thesis).

Sánchez, Aquilino. 2001. Investigación y análisis mediante corpus lingüísticos: el poder de atracción de las palabras. In P. Fernández Nistal and J. M. Bravo Gozalo (eds.). *Pathways of translation studies*, 11–45. Valladolid: Centro Buendía, Universidad de Valladolid.

Simpson, Percy. 1969 *Shakespearian punctuation.* Oxford: Oxford University Press.

Sinclair, John. 1995. Corpus typology: A framework for classification. In G. Melchers and B. Warren (eds.). *Studies in Anglistics* (Stockholm studies in English 85), 17–33. Stockholm: Almqvist and Wiksell International.

Teubert, Wolfgang. 1996. Comparable or parallel corpora? Special Issue of *International Journal of Lexicography* 9(3): 238–264.