Interdisciplinary Journal of Linguistics
Volume [18] 2025, 69-82

Issues in Developing a Comparable Corpus for Indian Languages as a Key Linguistic Resource for Cross-Linguistic Research and Applications

Niladri Sekhar Dash[*]

Abstract

A comparable corpus (CC) is compiled by selecting two different languages with a set of predefined dimensions (e.g., topic, genre, text types, time period, etc.), and selecting large amount of texts from respective languages, with equal importance on texts considered for inclusion (Maia, 2003). The multipurpose applications of a CC, in a country like India, are unbound as it develops a new ecology for developing language-based assistive technology to serve the multilingual requirements of Indian societies. In this conceptual position paper, we propose methods and approaches for the generation, composition, content, and application of a CC in the Indian languages with reference to its functional utilities in linguistic and extralinguistic domains. We first proposed an applicable definition of a CC (Laviosa, 1997), then we discuss about its typological diversities (Johansson & Hasselgård, 1999), refer to the present Indian situation of CC generation, argue for generating it for the benefit of the Indian languages, describe the processes one should apply to design a CC, and finally, highlight applications of a CC in linguistics and language technology in the context of Indian languages.

Keywords: Comparable Corpora, Technolinguistics, Lexical Resource, Multilingual, Equivalents, Termbanks

1. Introduction

The multipurpose utilities of a comparable corpus (CC) in linguistics and techno-linguistics confirm the fact that future multilingual works will depend heavily on data and information obtained from it (Teubert, 1996). The modern multilingual technologies need a large amount of processed data from a CC to develop object-oriented and user-centric tools and technologies. A CC with an adequate amount of empirical and verifiable texts and a unique composition becomes conducive for cross-domain, interlingual, and crosslingual applications (Atkins *et al.*, 1992). In the last few years, a CC has been a constant source of linguistic information and data to develop linguistic resources (e.g., *comparable lexicons, termbanks, comparable grammars, translation systems, text miners, and bilingual dialogue systems*) and comparative studies on languages, societies, ecologies, and cultures (Maia, 1997;

[*] Indian Statistical Institute, Kolkata; nsdash@yahoo.com

Laviosa, 2002). Data and information collected from a CC have been useful for developing a better understanding of relational dependency between sister languages, looking into bilingual culture frames, and constructing ethnographic profiles of the sister language communities.

Keeping these advantages in view, in this conceptual position paper, we present some ideas about the form, composition, content, and application of a CC for the Indian languages with reference to its functional utilities within and beyond linguistics. We first define the concept of a CC, address its diversities, refer to the present Indian context, describe the process of designing a CC, and highlight applications of a CC in linguistics and language technology in the context of Indian languages.

2. Defining a Comparable Corpus (CC)

The concept of a digital comparable corpus (CC) is not very clear to many of us working in the area of linguistics and technolinguistics. As a result, we come across many views and definitions of CC, many of which show characteristic differences despite definitional and conceptual proximities (Maia, 2003). To get a clear understanding of the concept, we refer to some definitions to understand how a CC is to be formed, framed, and utilized.

- A CC contains corpora from two different languages that come from the same domains. A news text corpus from English and Bengali, based on this definition, is a CC, although from a critical perspective, they may vary in topic, composition, domain, time, place, size, balance, representation, and others. This definition, in the true sense, does not serve our purpose of designing a CC.

- A CC may select similar texts from more than one language. Although this logic is followed in building a CC, there are a few studies that reveal a different purpose and characteristics of a CC. It seems that it is still in a concept than a reality. In our argument, the basic goal of a CC is to compare different languages in similar circumstances, but avoid the inevitable distortion generated through translations (Borin & Prütz, 2000).

- A CC, in principle, should be made of pairs of texts produced independently of each other. The texts are put into a CC as they are considered comparable in certain aspects. For example, we compile a corpus of written Hindi sports texts similar in form, content, and context to a corpus of written English sports texts.

- A CC should refer to a multilingual corpus containing texts in different languages, where text samples may not be the same in content, genre, or register. It may cover a variety of corpus types, referring to texts designed and generated following similar design criteria. Texts may or may not be translations of each other. If texts (*on the same topic*) are translations of each other, it is a Parallel Translation Corpus (PTC); if texts are aligned, it is an Aligned Parallel Translation Corpus (APTC).

- A CC also consists of two sets of texts: one originally written in language A and one of similar texts translated into language A, from a different language. Its value is mainly theoretical; what is investigated is the linguistic nature of translated text, independently of the source language (Baker, 1995).
- A CC is permitted to contain similar texts from several languages. The parallel translation corpora of the Indian languages that were developed in the ILCI project are a nice example of this (Jha, 2010). Also, the text corpora compiled in *AI4Bharat*, *Vidyaapati (2022-2026)*, and *Ishaan* (2022-2026) of the *Bhashini* project of the Govt. of India are good examples of a multilingual CC. Both *Vidyaapati* and *Ishaan* follow the same policy, model, time frame, genres, and quantity of words from several Indian languages.
- A CC can be made from a variety of sources: texts found in digital form, scanned, or typewritten. The criteria for choosing texts depend on the homogeneity of texts in terms of features such as subject domains, text origin, text constitution (i.e., *single or joint texts*), factuality, technicality, and application (Biber, 1993).
- A CC may contain texts not collected from the same genres, types, and fields. Moreover, texts may not carry the same kind of content and linguistic information. With several disparities in composition and content, a text collection may be treated as a CC if it is made for a specific purpose, such as producing comparable lexicons and dictionaries. The Comparable Reference Corpus (CRC) of the *Commission of the European Community* (CEC) is a nice example of this type.

Our main argument is that the context of defining a CC is that any two natural language texts can be treated as 'comparable texts' provided the two texts adhere to some conditions and features defined beforehand for the purpose of a particular study or investigation. In this perspective, a written text and a spoken text of a language can be treated as a CC provided they match on selection criteria.

Advanced technology makes it easy now to create small or big, single or multidimensional CC in many languages. Technolinguists, corpus linguists, translators, and content developers create and use multilingual CC as a source of information, data, and guidance.

The size of a CC can vary based on how judiciously it is made to meet the criteria of specific requirements. In the case of texts with a high degree of technicality and written by subject specialists for peers, it is fine to collect relevant data and information from a few selected sample texts. In such a situation, the amount of text does not create much problem if samples of texts are of similar content and equal size. On the other hand, in the case of newspaper texts from two different languages, we argue for developing a larger CC taking various linguistic and extralinguistic issues into the design criteria.

We also argue for accepting a CC consisting of texts from similar sources representing different varieties of the same language. For instance, the samples of English texts that have been collected from the newspapers published in different cities of India are a fine example of a CC, the analysis of which provides a clear understanding of how Indian English is represented differently in various states in India. It also provides valuable data and insightful information for developing comparable lexical databases and designing primers in Indian English (Dash, 2020).

3. Types of Digital Comparable Corpus

In our opinion, based on the types of sources of texts, a CC, keeping further subdivision at bay, can be of three types:

(a) Monolingual Comparable Corpus (MCC)

(b) Bilingual Comparable Corpus (BCC)

(c) Multilingual Comparable Corpus (MLCC)

3.1 Monolingual Comparable Corpus (MCC)

An MCC includes texts from one language but incorporates text samples from different genres, domains, and registers. Here, the type of a particular text acts as a pointer to determine the separate identities of the texts included in the corpus. Thus, any comparable collection of texts may be put under an MCC to initiate comparative studies across registers within a single language (Table 1).

Hindi	
Medical texts	**Engineering texts**
Recent publication	Recent publication
Printed texts	Printed texts
Large collection	Large collection
Subdomains inclusive	Subdomains inclusive

Table 1: MCC with texts from a single language

We further classify an MCC using different parameters, such as genre (*written vs. spoken*), time (*synchronic vs. diachronic*), register (*dialectal vs. standard)*, text type (*informative vs. imaginative*), text content (*literary vs. scientific*), and text authorship (*male vs. female*). The basic argument is that an MCC is free to contain texts from two different types of texts, provided the selection criteria are uniform across the text types.

3.2 Bilingual Comparable Corpus (BCC)

It is also possible to create a BCC with texts taken from two different languages, maintaining a kind of uniformity in text selection criteria and composition patterns across the two languages (Church, 2025). Much better investigational advantages are available if it is made with texts taken from typologically (e.g., *Hindi & Punjabi*), genealogically (e.g., *Bengali & Odia*), or geographically (e.g., *Odia & Telugu*) related

languages. One can create a BCC with texts from Hindi and Bengali, as the languages are both typologically and genealogically related. It, however, should be kept in mind that the text samples, which are included from the two languages, should belong to a similar genre, text type, and content (Table 2).

Hindi	**Bengali**
Medical Texts	**Medical Texts**
Recent publication	Recent Publication
Printed textbooks	Printed textbooks
Large collection	Large collection
Subdomains inclusive	Subdomains inclusive

Table 2: BCC with texts from two languages

A BCC may appear similar to a parallel corpus. In the case of a parallel corpus, content, text type, and genre are identical, while texts belong to two separate languages. In the case of a BCC, all the parameters may vary based on the requirements of a study. Like that of a MCC, a BCC is also used in language teaching, translation, and dictionary compilation (Barlow, 1996).

3.3 Multilingual Comparable Corpus (MLCC)

It is not necessary that a CC should always contain texts from two languages or text varieties. The content of a CC may include three or more languages if the texts are both characteristically and compositionally similar in domains and content (Abdulmumin et al., 2024). For instance, the AI4Bharat is engaged in developing parallel comparable corpora in many Indian languages for machine translation purposes (Gawtham et al., 2022).

A MLCC includes texts from three or more languages, while the criteria for considering texts from different languages remain the same. Here, the goal is to collect texts of identical domains and disciplines from different languages in a uniform manner so that comparison among texts becomes organized, systematic, uniform and useful. In a multilingual country like India, a MLCC is indispensable for developing cross-lingual resources, starting cross-lingual studies, and developing multilingual technologies (Table 3).

Hindi	**Bengali**	**Tamil**
Written	Written	Written
Cultural event	Cultural event	Cultural event
New year	New Year	New Year
Celebration	Celebration	Celebration
Activities	Activities	Activities

Table 3: MCC with texts from three languages

The attempt to develop a MCC for the Indian languages began a few years ago, when the project *Indian Languages Corpora Initiative (ILCI)* started with a goal for collecting comparable sentences from

various disciplines and subject domains from several Indian languages (Dash 2009; Jha, 2010). This effort is further strengthened with the initiation of the *Vidyaapati* and *Ishaan* projects under *Bhashini*, as well as with the establishment of the national hub for Indian languages as a part of the AI4Bharat.

4. Designing a Comparable Corpus

There is a demand for CC as it is rich with information to carry out multidimensional intralingual and interlingual studies. Besides, it is an essential requirement for many tools and resources like machine translation, information retrieval, and parallel multilingual lexicons. A CC perhaps does not need further specification of criteria beyond those defined for monolingual corpora (Sabel et al., 2022). Any two monolingual corpora, if put into comparative analysis, can be considered a CC. In such a case, factors controlling the constitution of each monolingual corpus may not be identical or similar. It can be assumed that while one corpus may comprise texts from one language (L1) of a particular discipline, the other one may consist of texts from another language (L2), containing texts from a different discipline. Although both texts theoretically constitute a bilingual corpus, when put to comparative analysis, they show little similarity for comparison. There are a few linguistic criteria that mark their differences (Table 4):

Criteria	**L1**	**L2**
Genre	written	spoken
Register	literature	mass media
Agent	young female	aged male
Domain	story	political speech
Setting	family	market

Table 4: The Bilingual corpus is not a comparable corpus

The differences of the factors stated above prevent texts from being comparable despite their bilingual entity. Any study between the two may produce results and observations hardly comparable between the languages. To overcome this limitation, we argue for developing bilingual parallel corpora (BPC) for interlinguistic comparison and mutual information extraction. It will be developed in such a way that both texts are subjected to alignment at the word and sentence level. It is necessary to define the design criteria for the creation of aligned and bilingual parallel corpora, as predetermination of designing factors relieves investigators from unwanted problems (Váradi and Kiss, 2001).

In the case of a CC, we argue that both texts should follow similar design criteria. This will enable access to the text easily and more simplified manner. Besides, it will help to process texts in an efficient way and utilise results more fruitfully in translation, parallel lexicon generation, and language teaching. Imagine an instance where both texts are parallelly developed following identical design criteria. Suppose two texts are created in Bengali and Tamil on the news of the

election published in two newspapers in 2025 in Tamil and Bengali. If both texts are put in comparison, they can be an ideal example of CC due to their identical design criteria (Table 5).

Criteria	**Bengali**	**Tamil**
Genre	written	written
Register	mass media	mass media
Domain	news report	news report
Sub-domain	politics	politics
Topic	election	election
source	newspaper	newspaper
Time	2025	2025

Table 5: Identical designing criteria for BCC

In some cases, bilingual texts may include original and translated texts of the same domains and genres (Maia, 2021). This is useful in translation and parallel lexicon generation. When we search for the appropriate terms in the source and target corpora, the following things happen.

- A term in the source corpus may not have an appropriate corresponding term in the target corpus.
- Translation 'per se' is probably impossible, as a term in one language may not correspond to a word in another language.
- A translated term may be a hyponym of a term in the source language.
- Language-specific expressions in the source corpus may not have a correspondence in the target corpus.

It implies that a parallel corpus (with translationally equivalent texts) and a CC (made with texts from pairs of languages) can vary in roles and relevance (McEnery & Oakes, 1995). They, however, provide data for those people who are looking for information based on which words and phrases can be translated acceptably (Zanettin, 1994).

When we propose to develop a CC with a collection of scientific texts from two different languages for collecting translational terminology (Kontonatsios, 2015), we suggest considering the following criteria:

- author-reader relationship (expert-expert or expert-naive communication),
- published texts (because they ensure some degree of acceptance of the terminology of a particular field),
- text origin (product of individual or joint authorship),
- constitution of texts (single or composite),
- factuality (text is factual or imaginative),

- technicality (text is technical or semi-technical),
- intended outcome (informative or didactic).

It is known that word-class-based equivalents, which are found in bilingual dictionaries, are not enough for judging the quality of translations. Analysis of CC shows how different innovations largely spread across text-types until they permeate the entire language. We also observe that the terminologies of CC provide a good empirical basis for describing these terms from the perspective of their use in different texts, based on which conceptual equivalent terms are possible to derive from a CC.

What we argue here is that with proper modifications of the existing techniques and rules, we can design a uniform method for developing CC for Indian languages to fulfill the needs of Indian linguistics and language technology (Siripragada, 2020). We can adopt innovative strategies for specific language pairs (e.g., *Hindi & Manipuri*) or develop a new model to include several Indian languages so that existing models of CC are refined to enable a wide range of applications for the Indian languages.

5. Utilities of a Comparable Corpus

Compared to a monolingual corpus, a CC provides much better insight into paired languages. A CC is made to supply linguistic data and information for comparative studies on language-specific, typological, and cultural differences. It presents notable differences between the source texts and their translations or between the two different languages. It shows how domain-centric monolingual texts vary based on various linguistic properties and features. It also reveals striking differences between native and non-native texts, as well as provides structurally equivalent texts and paired lexical databases for multilingual dictionary compilation, language teaching, and translation. Some more applications are defined below.

- Texts of a CC are often used for translation analysis to reconstruct the correspondences between segments of a source text and the segments of its translation (Candel-Mora & Vargas-Sierra, 2013).
- A CC is analysed to search for words, terms, and phrases with special meanings and frequent usages in specific domains, genres, and areas (Dragsted, 2006).
- A CC is used to look for information on ways in which a given word or a phrase can be translated acceptably in a target language (Esselink, 2003).
- A CC may be used to develop the procedural skills of the learners. Instead of dealing with isolated words and rules for combining them

into a text, the learners are trained to use wider syntactic chunks like multiword units, phrases, idioms, and collocations to learn about the formation of larger constructions (Fantinuoli, 2018).

The value of a CC in translation is very high. Technically, a translation is a text production activity in which many factors influence the source and target language situation. The analysis of a CC produces information to understand the respective expectations, experience, and knowledge of the languages involved. After establishing structural and semantic equivalence between comparable texts, we gather information about the way in which discourses are framed in two languages. Evidence collected from CC may guide to creation of new texts in the target language based on the requirements of new purposes.

A CC is a good source for learning about terminology. Made with scientific and technical texts, it is a useful source of terminology and its usage information for compiling bilingual or multilingual terminology databases that have strong applicational roles in Indian linguistic contexts. On the other hand, in a classroom context, a CC can be referred to confirm translation hypotheses and to suggest possible solutions to actual translation problems relating to specific text types. Specialised CCs offer information on concepts and terminology along with attested usages and distribution patterns controlled by contexts.

Another unique application of a CC may be linked with the investigation of particular genres and topic areas, as pre- and post-translational activities. A CC, as a source of potentially endless search processes, creates opportunities for language investigators to find genre-based and topic-based linguistic representations in the comparable texts, depending on the intuition, proficiency, interests, and needs of investigators.

The application of a CC can be envisaged in testing the relationship between those expressions of language that we perceive as holding some kind of semantic similarity or conceptual equivalence. The issue of semantic proximity or referential closeness of words has been an area of serious study in sense disambiguation and machine learning. A CC has a great role in this kind of study as it provides ample data and texts to measure the degree of semantic proximity of synonymous words used in two different texts. We can measure similarities between texts and compare words and phrases by identifying categories that have a strong formal resemblance, such as proper names and cognates, or which are proposed as translation equivalents between the two languages (Kaji and Morimoto, 2002).

A CC may be used as a yardstick to train the translators. In search of embedded constructions in a CC, for instance, a frequent use of embedded sentences can indicate the level of complexity involved in

the composition of the texts. Moreover, contrastive analyses of a CC reveal how similar ideas and concepts are expressed in similar kinds of texts in different languages. The analyses may consider different stylistic preferences such as use of rhetoric, propositions, figurative expressions, collocations, lexicogrammatical features, and others. Also, a CC supplies information about how discourses are realised in texts of different languages, and how linguistic and extralinguistic factors are taken into strategies for constructing different discourse patterns.

A CC does not offer single translationally established correspondence; it actually offers a repertoire of recurring collocational or structural features, which provide a basis for establishing equivalence between stretches of texts in two languages. The frequency of use of various linguistic items in a CC provides an indication of the expectations the prospective researchers may have about the linguistic items and features of the texts of a particular genre.

The translation assessment methods based on a CC can be integrated into a curriculum of trainee translators to guide them about the aspects and issues relevant to the creation of acceptable translations. For instance, a list of concordance retrieved from a CC is critically examined in a translation evaluation task as an aid to the analysis of source and target text features. A paired concordance list can be highly instructive to the translation trainees while engaged in a translation evaluation task.

In sum, the functional value of a CC primarily lies in its strength in opening up unprecedented scopes for comparing similar or different sets of language data as well as referring to different language types and varieties used under similar circumstances of text composition and application. Thus, a CC helps to avoid inevitable distortions noted in a parallel and translation corpora.

6. Why a CC is a Good Option for India?

From a practical perspective, we argue that CCs in Indian languages are far better linguistic resources than others in the present multilingual situation:

(a) **General Linguistics:** A CC made with two or more Indian language texts is a good resource for gathering many new insights, information, and understanding about the Indian languages used, which cannot be found by studying monolingual corpora.

(b) **Comparative Linguistics:** A CC is suitable for a wide range of comparative studies across the Indian languages to expand knowledge and understanding about language-specific, typological, genealogical, and cultural similarities and differences.

(c) **Linguistic Universal:** A CC of Indian languages can be a representative corpus to study, examine, attest, or reject some 'universal features' claimed to be present in the Indian languages.

(d) **Lexical Database:** A CC of Indian languages can be a useful resource for compiling multilingual lists of semantically equivalent words, conceptually similar technical terms, and translational equivalents.

(e) **Translation:** A CC in Indian languages is useful for finding similarities and differences between source texts and translations (Dewangan et al. 2021).

(f) **Practical Application:** A CC of Indian languages is indispensable for producing multilingual lexicons, multilingual idioms, dictionaries, writing comparative grammars, and preparing language teaching materials.

In essence, we find several practical applications of CCs built with Indian languages. Because of many visualized advantages, we propose to develop CCs of various types, forms, and texts between two or more Indian languages. Also, there are many theoretical justifications and cognitive advantages due to which we are in favor of compiling CCs for the Indian languages and argue for generating different kinds of CCs for developing tools and technology for the benefit of the languages.

7. Conclusion

A comparable corpus can be fruitfully used for many works of linguistics and techno linguistics in Indian languages. A CC can be useful if it is built within and across 'language pairs' as suggested above. Due to genealogical, typological, orthographic, and other similarities, a CC made within the paired Indian languages will be easy to align, process, and analyse in the respective languages.

It has been more than three decades now that the Indian languages have been used to develop monolingual corpora. Very recently, there have been efforts to develop CCs in the Indian languages (e.g., *AI4Bharat*), although we have not developed pan-Indian guidelines or protocol for this. In this position paper, we therefore present some ideas about how to develop a CC in Indian languages. We also argue that a CC is indispensable for developing multilingual resources (e.g., *lexicon, dictionaries, frozen expressions, equivalent syntactic constructions, context-based databases for translational equivalents, and ambiguous cross-linguistic lexical sets)* for the Indian languages. Data and examples collected from CCs can be utilized to write comparative grammars and develop multilingual scientific and technical terminologies.

In the present global scenario of Artificial Intelligence, Large Language Models, and Machine Learning, the CCs in the Indian languages can be indispensable resources for developing multilingual assistive tools and techniques, as well as starting comparative synchronic studies among the typologically, genealogically, and geographically related Indian languages. To achieve all the goals stated above, we suggest a collective and pan-Indian initiative to develop CCs in Indian languages and utilize the same in all kinds of linguistic and techno-linguistic works to break the divides existing among the Indian languages.

References

Abdulmumin, I., Mkhwanazi, S., Mbooi, M., Muhammad, S.H., Ahmad, I.S., Putini, N., & Marivate. V. 2024. Correcting the FLORES evaluation dataset for four African languages. In *Proceedings of the 9th Conference on Machine Translation*, Miami, Florida, Association for Computational Linguistics, pp. 570-578.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing, 7*(1), 1–16.

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

Barlow, M. (1996). Parallel texts in language teaching. In S. P. Botley, J. Glass, T. McEnery, & A. Wilson (Eds.), *Proceedings of Teaching and Language Corpora 1996* (UCREL Technical Papers No. 9, pp. 45–56). UCREL.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243–257.

Borin, L., & Prütz, K. (2000). Translated text. In W. Daelemans, K. Siman, J. Veenstra, & J. Zavrel (Eds.), *Computational linguistics in the Netherlands 2000* (pp. 30–34). Rodopi.

Candel-Mora, M.A. & Vargas-Sierra, C. 2013. An analysis of research production in corpus linguistics applied to translation. *Procedia,* 95: 317-324.

Church, K. 2025. Comparable corpora: Opportunities for new research directions. In *Proceedings of the 18th Building & Using Comparable Corpora Workshop* (pp. 73–82)., ACL, 20 Jan. 2025.

Dash, N. S. (2009). *Generation of translation corpora in Indian languages* [Paper presentation]. 2nd Workshop of the Indian Languages Corpora Initiative (ILCI-I), Dravidian University, Kuppam, India.

Dash, N. S. (2020). Developing a news text corpus of Indian English for language technology, machine translation, and language teaching. *International Journal of Translation, 32*(1–2), 31–57.

Dewangan, S., Joshi, N. & Bhattacharyya P. 2021. Experience of neural machine translation between Indian languages. *Machine Translation*. 35: 71-99.

Dragsted, B. 2006. Computer-aided translation as a distributed cognitive task. In Dror, I. and Harnad, S. (eds.) *How Cognitive Technology Extends Our Minds.* (pp. 237-256) John Benjamins, pp. 237-256.

Esselink, B. (2003). Localisation and translation. In H. Somers (Ed.), *Computers and translation* (pp. 67–86). John Benjamins

Fantinuoli, C. 2018. The use of comparable corpora in interpreting practice and training. *The Interpreters' Newsletter,* 23: 133-149.

Gawtham, R. et al. 2022. Samanantar: the largest publicly available parallel corpora collection for 11 Indic languages. *Transaction Association of Computational Linguistics*. 10: 145-162.

Jha, G. N. (2010). The TDIL program and the Indian Language Corpora Initiative (ILCI). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)* (pp. 982–985). European Language Resources Association.

Johansson, S. & Hasselgård, H. 1999. Corpora and cross-linguistic countries. In Granger et al. (Eds.) *Contrastive Linguistics and Translation*, Benjamins, pp. 145-162.

Kaji, H., & Morimoto, Y. (2002). Unsupervised word sense disambiguation using bilingual comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)* (Vol. 1, pp. 1–7).

Kontonatsios, G. (2015). *Automatic compilation of bilingual terminologies from comparable corpora* [Doctoral dissertation, University of Manchester].

Laviosa, S. (1997). How comparable can comparable corpora be? *Target, 9*(2), 289–319.

Laviosa, S. (2002). *Corpus-based translation studies: Theory, findings, applications*. Rodopi.

Maia, B. (1997). Sentence structure and thematization in comparable and parallel texts. In K. Klaudy & J. Kohn (Eds.), *Proceedings of*

the 2nd International Conference on Current Trends in Studies of Translation and Interpreting (pp. 541–547). Scholastica.

Maia, B. (2003). Training translators in terminology and information retrieval using comparable and parallel corpora. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in translator education* (pp. 43–54). St. Jerome.

Maia, B. 2021. Corpora, translation, terminology, and beyond: objectives and perspectives. *TradTerm, 37*(1): 10-29.

Mcenery, A., & Oakes, M. P. (1995). Sentence and word alignment in the CRATER project: Methods and assessment. In S. Armstrong-Warwick & E. Tzoukerman (Eds.), *Proceedings of the EACL-SIGDAT Workshop* (pp. 77–86).

Sabel, H. B., Dell'Oro, F., Montrichard, C., & Rossari, C. (2022). Setting up bilingual comparable corpora with non-contemporary languages. In *Proceedings of the 15th Workshop on Building and Using Comparable Corpora (BUCC 2022)* (pp. 56–60).

Siripragada, S., Philip, J., Vinay, P., & Jawahar, C. V. (2020). A multilingual parallel corpora collection effort for Indian languages. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (pp. 3743–3751).

Zanettin, F. 1994. Parallel words: designing a bilingual database for translation activities. In A. Wilson & T. McEnery (Eds.), *corpora in language education and research: A selection of papers from TALC 94.* UCREL Technical papers No.4, Lancaster, pp. 99-111. UCREL,