Available online at www.jilile.in ISSN: 3048-4952

JOURNAL OF INDIAN LANGUAGES AND INDIAN LITERATURE IN ENGLISH

Journal of Indian languages and Indian literature in English, 03(01), 39-49; 2025

The Use of Parallel Corpora to Develop Dictionaries and Grammars for Minority Languages

Mr.V.Selvakumar,
Research Scholar
&
Dr.K.Umaraj,
Associate Professor and Head Department of Linguistics,
Madurai Kamaraj University,
Madurai-21

APA Citation:
V.Selvakumar., (2025). The Use of Parallel Corpora to Develop Dictionaries and Grammars for Minority Languages, Journal of Indian Languages and Indian literature in English, 03(01),39-49; 2015

Submission Date: 08.03.2025

Acceptance Date: 24.03.2025

Introduction

The world's linguistic landscape is characterized by immense diversity, with thousands of languages spoken across the globe. However, a significant portion of this diversity is represented by minority languages, often spoken by smaller communities and facing the threat of language endangerment. These languages frequently lack the robust linguistic resources that are readily available for dominant languages, such as comprehensive dictionaries, detailed grammars, and advanced language technology tools. This scarcity presents a significant barrier to language preservation, education, and the integration of these languages into the digital age.

The challenges faced by minority languages are multifaceted. Limited access to educational materials in the native language can hinder intergenerational language transmission, leading to a decline in the number of speakers. The absence of standardized dictionaries and grammars can impede the development of consistent language norms and contribute to language attrition. Furthermore, the lack of language technology tools, such as machine translation and speech recognition systems, can marginalize these languages in the increasingly digital world. In this

context, parallel corpora, consisting of texts in a minority language aligned with their translations in a majority language, emerge as a powerful and cost-effective solution. These corpora provide a rich source of linguistic data that can be leveraged to develop essential language resources. By comparing aligned texts, researchers can extract lexical equivalents, identify grammatical patterns, and analyze semantic relationships, all of which are crucial for creating dictionaries and grammars.

This paper delves into the methodology and benefits of utilizing parallel corpora to address the resource scarcity faced by minority languages. It will explore the various techniques involved in extracting linguistic information from parallel corpora, including word alignment, lexical extraction, and grammatical analysis. Furthermore, it will highlight practical examples of how these techniques have been applied to develop dictionaries and grammars for specific minority languages. Finally, the paper will discuss the potential applications of these resources in language preservation, education, and technology development, emphasizing the importance of community involvement and ethical considerations. By showcasing the transformative potential of parallel corpora, this paper aims to contribute to the ongoing efforts to safeguard and promote the world's linguistic diversity.

The Power of Parallel Corpora

Parallel corpora, especially for languages as structurally different as English and Tamil, reveal a wealth of information. The comparative nature of these corpora helps us understand the nuances of each language.

Lexical Equivalents: Corresponding words and phrases between the two languages. Example:

Consider an English-Tamil parallel corpus:

- English: "The book is on the table."
- Tamil: "புத்தகம்மமமஜை மமல்உள்ளF." (Putthagam mejai mēl ulladhu.)

By comparing these, we can identify that "book" corresponds to "புத்தகம்" (putthagam), "table" corresponds to "மமஜை" (mejai), and "on" corresponds to "மமல்" (mēl). If the English sentence was "the red book is on the big table" and the Tamil was "சிவப்பு புத்தகம் பபரிய மமஜை மமல்உள்ளF" (Sivappu putthagam periya mejai mēl ulladhu) then we can see that "red" and "சிவப்பு" (sivappu) and "big" and "பபரிய" (periya) are lexical equivalents. This process, when repeated across a large corpus, allows for the creation of bilingual dictionaries and the identification of subtle lexical variations.

Grammatical Structures: How grammatical relations are expressed in the minority language through comparison to the majority language.

Example: In the same English-Tamil parallel corpus:

- English: "She ate the apple."
- Tamil: "அவள்ஆப்பிள்சாப்பிட்டாள்." (Avaḷ āppiḷ sāppiṭṭāḷ.)

By analyzing these, we can see that Tamil often places the object ("ஆப்பிள்," āppiḷ, "apple") before the verb ("சாப்பிட்டாள்," sāppiṭṭāḷ, "ate"). This demonstrates a Subject- Object-Verb (SOV) word order, which is different from English's Subject-Verb-Object (SVO) order. Also it is seen that the subject "she" and "அவள்" (avaḷ) are at the begining of each sentence. Through such comparisons, researchers can deduce grammatical rules, identify verb conjugations, and understand

the syntax of Tamil.

Semantic Relationships: The meaning of words and phrases in context.

- Example:
- English: "It is raining."
- Tamil: "மழை பபய்கிறF." (Maḻai peykiṟathu.)

Here, "it" is an impersonal pronoun in English, while Tamil uses "மழை" (maḻai, "rain") as the subject. Analyzing the context reveals how each language expresses the same concept differently. If the English sentence was "He is running" and the Tamil sentence was "அவன்ஓடுகிறான்" (avan oodugiraan) then it shows that "he" and "அவன்" (avan) are equal. By examining how words are used in different contexts, researchers can clarify their meanings and identify subtle semantic distinctions.

Collocations and Idioms: Common word combinations and figurative expressions.

- Example:
- English: "to kick the bucket"
- Tamil: "குடத்ஜத உஜதக்க" (kuḍaththai uthaikka) (literal translation), but the actual translation is "இறந்F மபா" (iṟanthu pō) or "காலம்பசன்றார்"(kālam ceṉṟār).

Parallel corpora are excellent for identifying idioms, which can be challenging to translate accurately. In this case, a literal translation would be misleading.Also, the English phrase "take care" translates to "கவனித்Fக்பகாள்" (kavaṉittuk koḷ) in Tamil. By observing these frequent word combinations in the corpus, we can identify common collocations.By systematically analyzing these linguistic features across a large collection of aligned English- Tamil texts, researchers can build a comprehensive understanding of Tamil, laying the foundation for the development of dictionaries, grammars, and other essential language resources.

Developing Dictionaries with Parallel Corpora

Word Alignment: This is the foundational step. It's about determining which words or phrases in the English text correspond to which words or phrases in the Tamil text.

- Example:
- English: "The cat sat on the mat."
- Tamil: "பூஜன பாயின்மீF உடக்காரந்தF." (Pūṉai pāyiṉ mītu uṭkārntatu.)

Tools like GIZA++ or fast_align would analyze these sentences (and many more) and attempt to determine the most probable alignments:

"The" > (Implied, but not a direct word-to-word translation in this case, but the idea is present)

- "cat" > "பூஜன" (Pūṉai)
- "sat" > "உட்காரந்தF" (uṭkārntatu)
- "on" > "மீF" (mītu)
- "the" > "பாயின்" (pāyiṉ)
- "mat" > "பாய்" (pāy)

The algorithms use statistical methods to determine the most likely pairings, based on frequency and context.

Lexical Extraction:

Once alignment is done, we extract potential dictionary entries.

- Example:
- From the aligned data, we would extract:
- "பூஜன" (Pūṉai) > "cat"
- "உடக்காரந்தF" (uṭkārntatu) -> "sat"
- "மீF" (mītu) > "on"
- "பாய்" (pāy) > "mat" We would then compile a list of unique Tamil words and their English equivalents.

Contextual Analysis:

This is where we refine the extracted entries by examining the surrounding context.

Example:

If "ஓடு" (ōṭu) appears in the corpus, alignments and context might show:

"ஓடு" (ōṭu) > "run" (in "அவன்ஓடுகிறான்" (avan ōṭukiṟāṉ) - "He is running") "ஓடு"

(ōṭu) > "tiles" (in "வீடின்ஓடு" (vīṭṭiṉ ōṭu) - "The house's tiles")

Concordances would display all instances of "ஓடு" (ōṭu) with surrounding words, allowing us to distinguish its different meanings.

Example Sentences:

Parallel corpora provide ready-made examples. Example:

- For the entry "பூஜன" (Pūṉai) -> "cat," we could include:
- "பூஜன பாயின்மீF உட்காரந்தF." (Pūṉai pāyiṉ mītu uṭkārntatu.) - "The cat sat on the mat."
- "பூஜன பால்குடித்தF." (Pūṉai pāl kuṭittatu.) - "The cat drank milk." These

sentences illustrate the word's usage.

Addressing Polysemy:

Many words have multiple meanings.

Example:

As shown above, "ஓடு" (ōṭu) has multiple meanings. By analyzing the contexts, we create separate dictionary entries:

- "ஓடு" (ōṭu) (verb) > "run"
- "ஓடு" (ōṭu) (noun) > "tiles"

This provides a starting point for creating a dictionary entry, or even two entries. By following these

steps, researchers can effectively utilize parallel corpora to create valuable dictionaries for minority languages, even when starting with limited resources.

Developing Grammars with Parallel Corpora:

Part-of-Speech Tagging:

This involves assigning grammatical tags (noun, verb, adjective, etc.) to words. Example:

English: "The cat quickly runs."

Tagged English: "The/DET cat/NN quickly/RB runs/VBZ"

Tamil: "பூஜன மவகமாக ஓடுகிறF." (Pūṉai vēkamāka ōṭukiṟatu.)

By aligning the sentences and transferring tags, we can create a tagged Tamil corpus:
"பூஜன/NN மவகமாக/RB ஓடுகிறF/VBZ" (Pūṉai/NN vēkamāka/RB ōṭukiṟatu/VBZ)

This tagged data can then be used to train a Tamil POS tagger. This process is very helpful, because training a POS tagger requires large amounts of pre-tagged data, and this method allows researchers to use the larger amount of tagged English data to bootstrap the creation of the Tamil tagged data.

Syntactic Analysis:

This focuses on the structure of sentences and how words relate to each other.

- Example:
- English: "The boy saw the bird." (SVO structure)
- Tamil: "ஜபயன்பறஜவஜய பாரத்தான்." (Paiyaṉ paṟavaiyai pārttāṉ.) (SOV structure)

By comparing the aligned sentences, we observe the different word orders. We see that the object "bird"/"பறஜவஜய" (paṟavaiyai) comes before the verb "saw"/"பாரத்தான்" (pārttāṉ) in Tamil. We can also begin to understand the case marking that occurs in Tamil. The -ai suffix in பறஜவஜய marks it as the object.

This helps in understanding the syntactic rules of Tamil.

Morphological Analysis:

This involves studying word forms and how they change (inflection, derivation).

- Example:
- English: "run," "runs," "running" (inflection)
- Tamil: "ஓடு," "ஓடினான்," "ஓடுகிறான்" (ōṭu, ōṭināṉ, ōṭukiṟāṉ) (inflection)

By analyzing variations of "ஓடு" (ōṭu) in the corpus, we can identify suffixes that indicate tense and person. We can see that -ināṉ indicates past tense, and -kiṟāṉ indicates present tense. This helps in understanding Tamil verb conjugations and other morphological processes.

Grammar Rules Extraction:

This involves identifying recurring patterns and formulating grammatical rules.

Example:

If we observe that Tamil verbs consistently end with certain suffixes depending on the tense and person, we can extract rules for verb conjugation. If we also notice that certain nouns take specific case endings, we can extract rules for noun declension. As an example, we may see that many nouns that indicate a location end with the suffix -il. This would allow a grammarian to create a rule concerning locative case.

Addressing Language Variation:

Minority languages often have regional or social variations.

Example:

Tamil has regional dialects. A parallel corpus containing texts from different regions can help document these variations. For example, certain words or grammatical constructions might be more common in one region than another. Social variation is also important. The way that Tamil is spoken in a formal setting, will be very different from the way that it is spoken in an informal setting. Parallel corpora from both kinds of settings will show these variations. By utilizing parallel corpora, researchers can effectively develop grammars for minority languages, even when starting with limited resources.

Tools and Resources

When working with parallel corpora to develop dictionaries and grammars for minority languages, a range of tools and resources are essential.

Parallel Corpus Alignment Tools:

GIZA++: A statistical machine translation toolkit that is widely used for word alignment.

It helps determine the probability of word correspondences between two languages in a parallel corpus. It is very powerful, but can be computationally intensive.

fast_align:

- A faster and more efficient alternative to GIZA++.
- It uses simpler algorithms but often achieves comparable or better alignment accuracy.
- It is often preferred for large corpora due to its speed.

Hunalign:

- Primarily used for sentence alignment, ensuring that corresponding sentences in the parallel corpus are correctly paired.
- This is a crucial first step before word alignment.
- Example: Before you can use GIZA++ to align words, you must use Hunalign to align the correct sentences.

Corpus Analysis Tools:

AntConc: A free and versatile concordancer that allows users to search for words and phrases in a corpus and view their contexts. It's invaluable for contextual analysis, identifying collocations, and examining word usage patterns. Example: You can use AntConc to search for all instances of a

specific Tamil word and see how it is used in different sentences.

Sketch Engine:

A powerful online corpus analysis tool that provides a range of features, including word sketches, thesaurus generation, and collocation analysis. It's particularly useful for identifying semantic relationships and creating lexical resources. Example: Sketch Engine can generate a "word sketch" for a Tamil word, showing its typical grammatical relations and collocations.

WordSmith Tools:

A comprehensive corpus analysis suite that includes tools for concordancing, word frequency analysis, and keyword extraction. It's useful for identifying key terms and analyzing language variation. Example: WordSmith can be used to compare the frequency of words in different dialects of a minority language.

Machine Translation Systems: Moses:

A statistical machine translation toolkit that can be used to train translation models from parallel corpora. While not directly used for dictionary/grammar development, it can be used to evaluate the quality of the developed resources. Example: After creating a basic dictionary, you could use Moses to translate sentences and assess the dictionary's effectiveness.

OpenNMT:

An open-source neural machine translation toolkit that uses deep learning techniques. It can achieve high translation accuracy and is increasingly used in language technology development. Example: OpenNMT could be used to create a Neural Machine Translation system for the minority language that would greatly aid in the creation of more parallel corpus data.

Linguistic Annotation Tools:

TreeTagger:

A part-of-speech tagger that can be used to automatically assign grammatical tags to words in a corpus. It's essential for creating annotated corpora that are used to train grammar models. Example: TreeTagger can be trained on tagged Tamil data (created via transfer tagging) to automatically tag new Tamil texts.

Stanford CoreNLP:

A suite of NLP tools that includes part-of-speech tagging, named entity recognition, and syntactic parsing. It's a versatile tool that can be used for a wide range of linguistic annotation tasks.

Example: Stanford CoreNLP can be used to parse Tamil sentences, revealing their syntactic structure. These tools, when used in conjunction, provide a powerful toolkit for researchers working to develop dictionaries and grammars for minority languages using parallel corpora.

Benefits and Applications

The development of dictionaries and grammars for minority languages, facilitated by parallel corpora, yields a multitude of benefits, impacting language preservation, education, technology, and cultural understanding.

Language Revitalization:

Endangered minority languages face the risk of extinction due to declining speaker numbers and limited resources. Dictionaries and grammars serve as crucial tools for documenting and preserving these languages. By providing standardized resources, they empower communities to actively use and transmit their language to future generations. Example: Imagine a small community where the traditional language is spoken by only a few elders. A well-constructed dictionary and grammar, accessible online and in print, can inspire younger generations to learn and use the language, fostering a sense of cultural identity and continuity.

Language Education:

Access to comprehensive dictionaries and grammars significantly enhances language learning. Learners can use these resources to expand their vocabulary, understand grammatical rules, and improve their fluency. These tools can be integrated into educational materials, such as textbooks and online courses. Example: Children in a bilingual education program can use a newly created Tamil-English dictionary to understand unfamiliar words in their Tamil lessons. Similarly, language learning apps can incorporate grammar rules extracted from parallel corpora to provide interactive exercises.

Language Technology Development:

Annotated corpora and linguistic resources are essential for developing NLP tools, such as machine translation, speech recognition, and text-to-speech systems. These tools can help bridge the digital divide and make minority languages more accessible in the digital world. Example: A speech recognition system developed for a minority language can enable speakers to use voice commands on their smartphones. A machine translation system can facilitate communication between speakers of different languages.

Cultural Preservation:

- Language is an integral part of cultural heritage, reflecting the history, values, and traditions of a community.
- Documenting the lexicon and grammar of a minority language helps preserve its cultural knowledge and wisdom.
- Dictionaries and grammars can include information on cultural practices, traditional knowledge, and oral traditions.

Example: A dictionary of a native language might include entries for traditional medicinal plants, along with their uses and cultural significance. A grammar might document the unique narrative structures used in oral storytelling.

Cross-Cultural Communication:

Dictionaries and grammars facilitate communication between speakers of minority and majority languages. They enable accurate translation and interpretation, promoting mutual understanding and respect. These resources can be used in various contexts, such as education, healthcare, and legal proceedings. Example: A bilingual dictionary can help a healthcare provider communicate with a patient who speaks a minority language. A bilingual glossary of legal terms can ensure accurate interpretation in court.

Challenges and Future Directions

While the use of parallel corpora offers significant advantages, it also presents challenges that need to be addressed to ensure the successful development and application of linguistic resources for minority languages.

Data Scarcity:

Obtaining sufficient parallel corpora for minority languages is often the most significant hurdle. Many minority languages lack readily available digital texts, and creating parallel corpora requires significant effort. Challenges include finding existing translations, obtaining permission to use them, and funding translation projects. Example: For very small language groups, there may be very little written material at all. In these cases, researchers must gather oral histories, and then translate them.

Future direction: Crowd sourcing translation efforts, leveraging community-generated content, and utilizing low-resource machine translation techniques to create pseudo-parallel corpora are potential solutions.

Language Variation:

Minority languages often exhibit significant regional and social variation, which can complicate corpus design and analysis. Standardizing linguistic resources while respecting dialectal diversity requires careful consideration. Researchers must ensure that corpora represent the full range of language variation. Example: Tamil itself has variations between Sri Lankan Tamil, and Indian Tamil, and within India, there are many regional dialects. A corpus that only contained one of these, would not represent the whole language.

Future direction: Developing dialect-specific dictionaries and grammars, utilizing sociolinguistic metadata to annotate corpora, and creating flexible linguistic resources that accommodate variation are essential.

Ethical Considerations:

Respecting the cultural sensitivities of minority language communities is crucial. Researchers must obtain informed consent before using language data and ensure that the development of linguistic resources benefits the community. Issues such as data ownership, intellectual property rights, and the potential for cultural appropriation must be addressed. Example: Some languages contain culturally sensitive information that must be handled with great care, and not released to the general public.

Future direction: Establishing ethical guidelines for language documentation, engaging in collaborative research with community members, and ensuring that linguistic resources are used in a culturally appropriate manner.

Developing User-Friendly Resources:

Creating dictionaries and grammars that are accessible and useful to language learners and speakers is essential. Resources should be designed with the needs of the target audience in mind, considering factors such as literacy levels, technological access, and cultural preferences. Example: A digital dictionary might include audio pronunciations, images, and interactive exercises to enhance learning.

Future direction: Utilizing user-centred design principles, developing mobile applications, and creating multimedia resources are crucial for enhancing accessibility and usability.

Community Involvement:

Engaging minority language communities in the development of linguistic resources is crucial for ensuring their relevance and sustainability. Community members possess invaluable knowledge of their language and culture, and their input is essential for creating accurate and culturally appropriate resources. Example: Holding community workshops to gather feedback on dictionary entries, or having community members help to record audio pronunciations.

Future direction: Establishing community-based language documentation projects, providing training and support to community members, and ensuring that linguistic resources are owned and managed by the community.

Conclusion

In essence, parallel corpora represent a paradigm shift in minority language resource development. They move beyond the limitations of isolated linguistic data, offering a dynamic and contextualized approach to capturing language structure and usage. This methodology directly addresses the resource scarcity that hinders the vitality of many minority languages, providing a pathway to create robust dictionaries and grammars. These tools are not merely academic artifacts; they serve as active agents in language preservation, empowering communities to maintain their linguistic heritage. The impact of parallel corpora extends into the realm of education, where they facilitate more effective language learning, and into technology, where they enable the development of essential tools that bridge the digital divide. Furthermore, they play a critical role in cultural preservation, documenting the deep connection between language and culture. While advancements in AI and machine learning are poised to accelerate the creation of these resources, the human element remains paramount. Community collaboration is not simply a best practice; it is a fundamental requirement. Respecting cultural sensitivities, ensuring equitable access, and prioritizing community ownership are essential for the long-term viability and impact of these projects. Ultimately, parallel corpora, when used ethically and collaboratively, provide a powerful mechanism for safeguarding and celebrating the rich diversity of human languages. They offer a tangible means of ensuring that minority languages not only survive but thrive in the 21st century.

References

- McEnery, T., & Xiao, R. (2007). Parallel corpora and contrastive linguistics. Computer Assisted Language Learning, 20(1), 9-33.
- Kenny, D. (2001). Lexis and creativity in translation: A corpus-based study. Manchester: St. Jerome Publishing.
- Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. Target. International Journal of Translation Studies, 7(2), 223-243.1
- Bowker, L. (2002). Computer-aided translation technology: A practical introduction. University of Ottawa Press.

- Hale, S. (2010). Community interpreting. Palgrave Macmillan.
- Hutchins, J., & Somers, H. (1992). An introduction to machine translation. Academic Press.
- Oakes, M. P. (1998). Statistics for corpus linguistics. Edinburgh University Press.
- Sinclair, J. (1991). Corpus concordance collocation. Oxford University Press.
- References Specifically Related to Minority Languages:
- Bird, S. (2010). Where can I find language data? In J. Litman, & A. Renear (Eds.), Perspectives on Data Informatics (pp. 1-28). Morgan & Claypool Publishers.
- Hinton, L. (2001). Language revitalization: An annotated bibliography. Cambridge University Press.
- Grenoble, L. A., & Whaley, L. J. (Eds.). (2006). Saving languages: An introduction to language revitalization. Cambridge University Press.
- Ostler, N. (2010). The dying of language. Canon gate Books.
- Koehn, P. (2010). Statistical machine translation. Cambridge University Press.
- Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing.2 MIT press.
- Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python.3 O'Reilly Media, Inc.
- http://www2.statmt.org/moses/giza/GIZA++.html
- https://www.sketchengine.eu/