What Resources Matter for Interlinear Glossing? Using LLMs and RAG for the Low-Resource Mapudungun Language

Anaís Almendra[1], Arianna Bisazza[2], Claudio Gutierrez[3], Felipe Hasler[1]
[1] Department of Linguistics, Universidad de Chile, Santiago, Chile
[2] CLCG, University of Groningen, Groningen, Netherlands
[3]Department of Computer Science, Universidad de Chile, Santiago, Chile
Correspondence: anaisalmendra21.aa@gmail.com

Abstract

Interlinear glossing is essential for the study and revitalization of endangered languages. However, it remains a time-consuming process that requires extensive linguistic expertise. Recent advances in Large Language Models (LLMs) offer a potential solution. In this research, we study the case of Mapudungun, an endangered language spoken in Chile and Argentina, to generate automatic interlinear glosses using the Gemini 2.5 Pro model. Our study investigates which information configuration through Retrieval-Augmented Generation (RAG) yields the best results. We compare the integration of a formal grammar, a dictionary, a small annotated corpus, and a combination of all these resources. Our evaluation shows that while dictionary integration causes a significant degradation in performance, grounding the model with a structured corpus maximizes accuracy relative to the resources employed. Notably, we find that a remarkably small dataset of 589 meaning units provides enough normative guidance to significantly improve the morphological tagging task. This work highlights the viability of utilizing minimally annotated corpora to assist in the documentation of morphologically complex languages.

1 Introduction

Currently, numerous languages worldwide are in a vulnerable situation (SIL International, 2026). Among them is Mapudungun (Gundermann et al., 2011; Zúñiga, 2019), an indigenous language of Chile and Argentina with approximately 380,000 speakers in the former country (Instituto Nacional de Estadísticas, 2024). This language has a morphological profile classified as agglutinative and polysynthetic (Golluscio and Hasler, 2017; Zúñiga, 2017) with a templatic structure for morpheme incorporation (Fortescue et al., 2017).

In the field of linguistic documentation, there are data annotation techniques through which text, audio, or video are structured and aligned with metadata (Woodbury, 2011). Among these procedures, interlinear glossing stands out as a technique commonly used in projects of this nature (Ginn and Palmer, 2023). This annotation format operates through the segmentation of morphemes and the assignment of grammatical tags (Elsner and Liu, 2025), which allows capturing the morphosyntactic features of the language under study and performing precise analyses of them (Ginn et al., 2024b). In this context, the Leipzig Glossing Rules provide a series of syntactic and semantic rules, in addition to a standardized set of tags, to systematize the interlinear glossing process (Comrie et al., 2015).

In the case of endangered languages, interlinear glossing has established itself as a documentary tool capable of facilitating their revitalization (Elsner and Liu, 2025). Despite the benefits that the generation of interlinear glossing brings to endangered languages, this task constitutes a highly complex endeavor, heavily dependent on experts, and requiring large amounts of time for its generation (Ginn et al., 2024b). To illustrate this task, the following example corresponds to a Mapudungun text organized according to the Leipzig Glossing system:

(1) kuydaufisayawürkey
kuyda -ufida -ya -w -ürke
take.care -sheep -FUT -REFL -EVID
-y
-IND.[3]

'It is say he was tending his sheep.'

In this example, Mapudungun phenomena can be observed, such as object incorporation into the verbal root and the suffixation of elements such as evidentiality, tense, and person.

Despite the utility of interlineal glossing, Mapudungun lacks a corpus with these characteristics.

In response to this limitation, and in relation to

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 64–73
July 3, 2026 ©2026 Association for Computational Linguistics

previous works (Ginn and Palmer, 2023; Ginn et al., 2024a,b; Elsner and Liu, 2025), we propose the use of Large Language Models (LLMs) for automatic interlinear glossing. Since Mapudungun, like most languages in the world, lacks a sufficient corpus to train large language models (Zhang et al., 2024), our study leverages the linguistic capabilities of these models to generate glosses for Mapudungun without the need for further training.

We investigate what is the most efficient approach for generating automatic interlinear glossing for Mapudungun. Specifically, we evaluate the Gemini 2.5 Pro model and explore the use of the Retrieval-Augmented Generation (RAG) technique to provide the LLM with three different sources of language information: (i) grammar, (ii) dictionary, and (iii) a Mapudungun corpus newly annotated according to the Leipzig Glossing system.

Our work identifies key issues regarding the generation of automatic glossing for Mapudungun. Firstly, the results indicate that none of the additional information sources provide statistically significant improvements in the text and segmentation tiers of the Leipzig glosses. Secondly, the tagging task is identified as the most complex. In this tier, the integration of additional information from the annotated corpus and the use of all materials combined do provide a statistically significant difference. Thirdly, it is more effective to use solely standard information, such as the annotated corpus, rather than the combination of multiple resources (dictionary, grammar, and corpus together), since the former produces better results and requires fewer resources. Consequently, these findings contribute to research seeking to perform automatic interlinear glossing for languages with scarce annotated corpora, such as Mapudungun, establishing the most efficient alternative with the lowest environmental impact.

2 Background and Related Work

2.1 Mapudungun Language

Mapudungun is a language of isolated genealogy (Golluscio and Hasler, 2017; Zúñiga, 2017) spoken in Chile and Argentina, which is currently in a vulnerable situation (Gundermann et al., 2011; Zúñiga, 2017). In Chilean territory, it has approximately 380,000 speakers (Instituto Nacional de Estadísticas, 2024). From the point of view of its morphological typology, it is classified as an agglutinative and polysynthetic language (Zúñiga, 2017). In turn, its affix integration structure responds to a templatic model (Fortescue et al., 2017; Zúñiga, 2017), given that its morphological incorporation is rigid regarding the positions in which morphemes can be added in relation to the verbal root. This morphology of the language has been classified as complex due to the diversity of linguistic phenomena it exhibits (Zúñiga, 2017). In this regard, processes such as noun incorporation, the adjunction of multiple morphemes onto a single root, and the large number of elements that possess the capacity to operate as roots within a morphological construction stand out (Zúñiga, 2017).

Regarding the available linguistic resources for Mapudungun, although grammars, dictionaries, and various digital tools exist, there is a lack of a specialized corpus in interlinear glossing. In this context, the automation of annotated corpus generation emerges as a strategy with the potential to accelerate the creation of resources for the study and revitalization of the language.

2.2 Automatic Interlinear Glossing

Given the complexities associated with the creation of interlinear glossing, recent research has explored the application of LLMs to accelerate the processes of generating corpora with interlinear glossing. Along these lines, Ginn and Palmer (2023) approached the task through the fine-tuning of a pretrained model, reporting an increase of two percentage points in performance compared to the unrefined architecture. This study highlights the inherent complexity that the automation of this labor represents. Subsequently, Ginn et al. (2024b) performed a model adjustment for gloss generation in multiple languages, employing a database of approximately 450,000 examples distributed across 1,800 languages. The results show an improvement of about 7 percentage points against state-of-the-art models. Despite this, the authors note that performance varies largely with the amount of available data.

Instead of model refinement, Ginn et al. (2024a) evaluated the use of few-shot prompting, exploiting the in-context learning abilities of LLMs. In this design, the system receives as input a transcription line and its respective translation, from which it must generate the corresponding gloss. The findings indicate that providing interlinear glossing examples in the prompt substantially impacts the model's performance. Likewise, it is concluded that, although interlinear gloss generation remains

highly complex, the strategic selection of pertinent examples can yield significant improvements.

Finally, Elsner and Liu (2025) also applied the prompting technique for automatic interlinear glossing, adopting the Leipzig Glossing Rules as a normative standard. They worked with different languages and used a prompting system with examples for each one that contained sentences, glosses, and translations. The study shows promising results and highlights the potential of LLMs as support tools for linguists.

2.3 In-context Learning for Low-Resource Languages

In the context of LLM use and low-resource languages, recent research has opted to utilize the linguistic capabilities of the models instead of further training them (Court and Elsner, 2024; Zhang et al., 2024; Spencer and Kongborrirak, 2025; Zhu et al., 2025). In the aforementioned studies, the models are operated through prompting with different types of linguistic information, depending on the expected task and the language being worked with.

Specifically, Court and Elsner (2024) work with the translation task from a Quechua variant to Spanish. They utilize the RAG and prompting technique to work with three types of materials, both separately and jointly: translations of morphemes and words, grammatical descriptions, and usage examples from parallel corpora. Spencer and Kongborrirak (2025) experiment with different RAG and prompting configurations with the goal of assisting the creation of a grammar for a low-resource language.

Zhu et al. (2025) use a puzzle-based methodology with features of varying complexity to identify whether LLMs can capture linguistic features of unseen languages. To deliver the language data, they work with step-by-step prompting. Finally, Zhang et al. (2024) test the translation task between English and languages unseen in the training corpora. For this, they utilize three materials through prompting: dictionary, grammar, and morphologically analyzed text.

3 Methodology

The present research works with the RAG technique on the Gemini 2.5 Pro model and uses as materials *A Grammar of Mapuche* (Smeets, 2008), *Diccionario mapudungun-español español-mapudungun* (Augusta, 2017), and adapts part of the AVENUE corpus (Levin et al., 2002), enriching it through the use of Leipzig Glosses.

Our interaction system, as shown in the Prompt Used Appendix A.1, instructs the model to process an input text and return its analysis in a single message response according to the Leipzig Glossing Rules. In this way, we encourage the model to treat the four Leipzig tiers as a single unit, thereby aiming to minimize potential hallucinations or the omission of units during the segmentation and tagging processes.

3.1 Evaluation

The results of the experiment were evaluated according to three Leipzig Glossing tiers: **text**, **segmentation**, and **tagging**. The translation tier is excluded from the analysis given that the task of evaluating how well a translation is performed requires theoretical and practical approaches beyond the scope of this study, which we leave to future work. We assess text, segmentation, and tagging in a binary manner based on the manually annotated evaluation corpus. That is, only an exact match of the expected meaning is considered correct, while any deviation is counted as an error.

McNemar's test was used to compare the models, and analyses of Accuracy and p-values were conducted. Each model was compared against the evaluation corpus; subsequently, the baseline was compared independently with corpus, dictionary, grammar and all the materials models. Because McNemar's test strictly requires paired data of equal length, we developed a tier alignment system to handle the LLMs' tendency to omit information or hallucinate content. In cases where the model generated an unnecessary morpheme or omitted a required one, our system aligned the output by inserting empty spaces, which were scored as incorrect predictions. This alignment method ensured that the evaluation sets maintained the same length.

In Table 1, we present representative examples for each evaluation tier using the dictionary and all materials settings. This exact cell-by-cell alignment and evaluation protocol was systematically applied to all experimental configurations.

3.2 Materials Used

We experiment with delivering three different linguistic resources to the model via RAG, as well as their combination:

Tier / Setting	Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6
1. Text Tier						
Evaluation	nierpuafuy	–	–	–	–	–
Baseline	nierpuafuy (✓)	–	–	–	–	–
Dictionary	nierpuafuy (✓)	–	–	–	–	
All	nierpuafuy (✓)	–	–	–	–	–
2. Segmentation Tier						
Evaluation	nie	r	pu	a	fu	y
Baseline	nie (✓)	Ø (×)	rpu (×)	a (✓)	fu (✓)	y (✓)
Dictionary	nie (✓)	Ø (×)	rpu (×)	a (✓)	fu (✓)	y (✓))
All	nie (✓)	Ø (×)	rpu (×)	a (✓)	fu (✓)	y (✓)
3. Tagging Tier						
Evaluation	have	ITR	TRANS	FUT	FRUS	IND.[3]
Baseline	have (✓)	Ø (×)	TRNSL (×)	FUT (✓)	IRR (×)	IND.[3] (✓)
Dictionary	have (✓)	Ø (×)	TRNSL (×)	FUT (✓)	IRR (×)	IND.[3] (✓)
All	have (✓)	Ø (×)	TRANS (✓)	FUT (✓)	IRR (✓)	IND.3 (×)

Table 1: A representative example *nierpuafuy* 'They would gradually have' from the evaluation corpus, illustrating the cell-by-cell alignment and binary evaluation across all tiers. The table contrasts the baseline, dictionary, and all materials configurations against the evaluation corpus. This alignment method provides equal sequence lengths across outputs, yielding the exact paired binary metrics required to compute McNemar's test.

- Grammar: *A Grammar of Mapuche* (Smeets, 2008) has been used by already existing morphological analyzers for Mapudungun (Almendra, 2025) and presents a chapter entirely dedicated to the morphology of the language and its structuring.

- Dictionary: *Diccionario mapudungun-español español-mapudungun* (Augusta, 2017) has been classified as one of the most comprehensive lexical works of the language (Augusta, 2017) and has been used for the development of a morphological analyzer for Mapudungun (Almendra, 2025).

- Corpus: the AVENUE project includes an open-access corpus of transcribed spoken Mapudungun (Levin et al., 2002). For this research, a part of it was selected and enriched with annotation according to Leipzig Glosses (cf. section 3.3)

Each of these materials provides different kinds of information regarding Mapudungun morphology. The grammar presents examples alongside explanations of the rules that Mapudungun follows regarding its morpheme structuring. The dictionary constitutes a repository of equivalencies between Mapudungun and Spanish. This resource presents the lemmas containing the language's morphemes together with their equivalencies in the majority language. Finally, the corpus is a set of demonstrations regarding how to gloss Mapudungun according to the Leipzig system. This information provides annotation examples in specific contexts of use; however, it does not present explicit explanations regarding the functioning of Mapudungun morphology.

3.3 Corpus

The corpus was obtained from a total of four audios that were aligned in ELAN (Max Planck Institute for Psycholinguistics, n.d.) with their transcriptions and translations for their subsequent analysis in FLEx (SIL International, 2026). It consists of a series of texts in Mapudungun glossed according to the Leipzig rules. Each sentence or word in Mapudungun appears in the corpus with four tiers of information: text, segmentation, tagging, and translation (see example gloss in 1). As Mapudungun has multiple orthographic systems (Llanquimán et al., 2025), we standardized the orthography using the tools and guidelines of *KMT - Kümewirin Mapudüngun Trapümwe* (Chandía, n.d.).

Then, we applied the following annotation principles:

1. The corpus was cleaned by eliminating interjections, incomplete or erroneous speech onsets, and proper nouns.

2. The inclusion of lexical borrowings was permitted, provided they did not imply code-

switching to Spanish within the same sentence, thus maintaining the focus on the structure of Mapudungun.

3. For the generation of Leipzig glosses, truncated or abbreviated forms in speech were represented by their full form in the gloss. For example, a form such as *feli* was segmented as *feley* for its subsequent tagging.

4. The glossing process allowed the correction and validation of translations and the assignment of tags to morphemes. Samples in which there was doubt along any of these two aspects were excluded.

The annotation process was carried out by two of the authors in two stages. First, a linguist specialized in ELAN and FLEx with intermediate knowledge of Mapudungun grammar performed the audio annotation and generated the preliminary analyses. Subsequently, a linguist specializing in the language's grammar reviewed and finalized the data in FLEx.

Given the polysynthetic and agglutinative nature of Mapudungun, the corpus split into the RAG retrieval subset and the evaluation subset was not measured by individual words. Instead, we used the concept of unit of meaning as the base metric. We define a unit of meaning as any element that possesses an independent linguistic tag. For example, the lexeme *ñuke* 'mother' counts as one unit. In contrast, a complex verbal form such as *tunualengün* is segmented into *tu-nu-a-*l=*engün*, receiving the tags [grab-NEG-FUT-NMLZ-3.PL], and is therefore counted as five meaning units. Once the entire corpus was processed, we counted the total number of meaning units and divided the RAG and evaluation data based on this metric.

Since the corpus was derived from audios, to avoid biases in the use of RAG and the evaluation, we aimed to allocate half of the content of each audio to each subcorpus, thus ensuring equitable representativeness. In its entirety, the total corpus corresponds to 40 speech turns and 1,175 units of meaning. The RAG subset contains 589 units and 22 speech turns, while the evaluation subset consists of 586 units and 18 speech turns.

In comparison to other studies, where hundreds of thousands of examples are utilized (Ginn et al., 2024b), the size of the corpus in this research is minute. Given the cost of manual annotation, we are interesting in assessing whether LLMs could accelerate the annotation process of further data.

3.4 Infrastructure Selection

The model used was Gemini 2.5 Pro on Google Cloud (Vertex AI), specifically within the Model Garden infrastructure employing the RAG Engine technique. This was selected for two essential reasons: (i) data management, and (ii) accessibility to the use of RAG. On the one hand, the Vertex AI platform provides guarantees regarding data use, indicating that these will not be used for training its models and that the generated results do not belong to the company either (Google LLC, 2025). On the other hand, Vertex AI offers the possibility of applying RAG without the need to develop code. This aspect is crucial to our approach, since it facilitates access for individuals who are not specialized in this area, such as linguists, and allows working with specialized models and data like grammars, dictionaries, or other materials. The model temperature is set to 0 in all experiments.

4 Results

We present the results using accuracy and statistical significance. Since each tier of the Leipzig gloss was evaluated independently, we report the performance separately for each one. In all experiments, 'Baseline' refers to the results obtained by querying Gemini 2.5 Pro with a fixed prompt containing 3 glossing examples (see full prompt in A.1). The 'All' setting refers to the combination of the three materials (grammar, dictionary, corpus) provided altogether to the LLM via RAG.

4.1 Text

In the text tier, only the corpus configuration outperforms the baseline, but by less than one percentage point. While the combined materials configurations match the baseline performance, the dictionary and grammar configurations underperform.

As shown in Table 2, the observed differences were only significant in the case of the dictionary, which worsened the performance ($p < 0.05$). None of the other settings showed a significant difference in the text tier ($p > 0.05$). These results indicate that the baseline already achieves high performance in the text processing task, which limits the observable margin of improvement from material additions via RAG.

Setting	Text	Segmentation	Tagging
Baseline	98.26%	78.62%	41.08%
Corpus	**98.96%**	74.92%	**56.41%***
Grammar	96.18%	76.88%	39.46%
Dictionary	90.62%*	70.95%*	34.91%*
All	98.26%	**80.60%**	**55.69%***

Table 2: Evaluation results (Accuracy) across the three tasks: text, morphological and tagging. The table compares zero-shot execution against various RAG augmented contexts. Bold values indicate results that outperform the baseline, and * denotes statistical significance.

4.2 Segmentation

We first observe that the segmentation results are overall lower than those achieved in the text task. This indicates a greater complexity for the morphological segmentation task.

In the segmentation tier, only the use of all materials outperforms the baseline. The corpus, dictionary, and grammar configurations individually underperform.

As shown in Table 2, once again the observed differences were only significant in the case of the dictionary, which worsened the performance ($p < 0.05$). None of the other settings showed a significant difference in the segmentation tier ($p > 0.05$). Notably, even the configuration that outperformed the baseline (All) did not yield a statistically significant improvement.

4.3 Tagging

The tagging task appears as the most complex among the three evaluated tiers with accuracies below 60% in all configurations.

In this tier, the corpus configuration and the combined materials configuration outperform the baseline, with the former achieving better performance. On the other hand, the individual use of the dictionary and grammar underperforms.

As shown in Table 2, the observed pattern was maintained, and the use of the dictionary significantly worsened the performance ($p < 0.05$). Unlike the previous tiers, the all materials and corpus settings significantly improved the performance ($p < 0.05$). Only the grammar configuration did not show a significant difference, neither improving nor worsening the performance ($p > 0.05$).

5 Discussion

Our results show that the inclusion of the grammar does not yield statistically significant gains in the system's performance compared to the baseline. This can be due to the fact that the model already possesses in its weights knowledge of Mapudungun that overlaps with the one provided by the grammar. Alternatively, this could mean that the model is not capable of exploiting the information provided by grammar and applying it to novel examples.

Dictionary integration produces a significant degradation of results across all analyzed cases. Even in the text tier, which should not present complexities for the model given that it constitutes a replication of the text provided by the user, the use of the dictionary worsened performance.

In contrast, the configurations based on the use of the corpus and all materials consistently record the highest levels of accuracy.

Regarding the statistical significance of these findings, we found that no configuration made a significant improvement in the text processing and segmentation tasks. In these areas, the performance of the model without additional information overlaps with the all materials and corpus configurations. Despite this, in the tagging task, we identify statistically significant improvements when employing the corpus and all materials configurations. Among them, the former not only exhibits superior performance but also achieves this through the use of a smaller volume of data compared to the use of all materials. These results show that the greater the complexity of the task, the greater the relevance of using RAG; in this sense, the use of additional information only makes a real contribution to improving the tagging tier.

In the tagging tier, the all and corpus configurations exhibit the best performance, with the latter outperforming the rest. Whereas the dictionary and grammar fail to reach the baseline, the aforementioned configurations exceed it with statistical significance. We attribute the performance gains of

the all configuration to the inclusion of the corpus, given that it yields the best overall results when used in isolation.

The performance gains achieved through the corpus integration can be observed in the mitigation of recurring baseline errors. Examples of this involve the morphemes *-fu* and *-nie*. In the baseline output, *-fu* was tagged as IRR instead of the correct target tag, FRUS. The all and corpus configurations resolved this by correctly tagging the *-fu* morpheme as FRUS. Likewise, the baseline confused the morpheme *-nie* with the verbal root *nie* 'have'. Once again, the all and corpus configurations correctly assigned the PROG.PS tag to this morpheme. Despite the gains achieved, we note common errors in the tagging tier, primarily driven by homophonous morphemes. In Mapudungun, the suffix *-tu* can take different meanings depending on the context, functioning as a verbalizer, a repetitive marker, or a transitivizer. In cases where *-tu* appeared, the model exhibited a frequent error by assigning the verbalizer label even when incorrect, probably due to an overrepresentation of this specific function in the data. To prevent this, we believe that refining the selection of examples to ensure a balanced representation of these syntactic phenomena could improve performance in future work.

Our findings align with the high complexity of automatic interlinear glossing highlighted in previous research (Ginn and Palmer, 2023), especially within the tagging tier. Consistent with recent literature (Ginn et al., 2024a; Elsner and Liu, 2025), we observed that utilizing prompting and the examples provided by our developed corpus offers an effective alternative to model fine-tuning. Furthermore, our RAG methodology improved the baseline results for Mapudungun, aligning with its successful use in tasks with other languages (Spencer and Kongborrirak, 2025). Nevertheless, these outcomes emphasize that the nature of the augmented material is critical, as external information does not inherently guarantee better results, as also observed in other studies (Court and Elsner, 2024). For instance, while some studies found the use of grammar beneficial (Zhang et al., 2024), the grammar configuration tested in this study failed to produce statistically significant gains. Additionally, this contrasts with previous work (Court and Elsner, 2024), which observed performance drops when using a corpus in translation tasks. These discrepancies suggest that the effectiveness of specific RAG resources is task-dependent.

In view of the above, we propose that to optimize the automatic tagging of Mapudungun using LLMs, the most effective strategy consists of employing an annotated corpus that acts as a normative guide for the model. This finding is relevant, as it contradicts a possible hypothesis that the use of diverse linguistic resources favors better generalization or morphological analysis. In this sense, we believe that combining a corpus structured under the Leipzig Glossing Rules with a RAG approach is a methodological novelty that allows us to improve the performance in automatic interlinear glossing generation. For low-resource languages with complex morphology like Mapudungun, this approach prioritizes information density over large data volumes, providing multiple specialized tiers from just a single word or sentence. Despite these insights, the findings of our study should be interpreted within the scope of a single LLM configuration. While this framework can inform future work across other architectures, it is important to acknowledge this limitation.

Finally, the size of the corpus employed in this research provides relevant evidence regarding the processing of Mapudungun morphology by Gemini 2.5 Pro. As mentioned, the quantity of examples provided by the corpus in this study is considerably smaller than that utilized in other works. In this context, our results indicate that the use of 589 meaning units, equivalent to 22 speech turns, is sufficient to improve automatic interlinear glossing for Mapudungun. Consequently, these findings establish a minimum threshold of data necessary to replicate this technique in other low resource languages like Mapudungun.

6 Conclusions

In the context of the automatic generation of interlinear glossing for Mapudungun using RAG, the results indicate that the best cost-result strategy consists of utilizing a previously glossed corpus that operates as a normative standard, rather than integrating multiple documentary sources. Our findings show that integrating LLMs and external data is a viable approach for glossing low-resource languages such as Mapudungun.

Likewise, we observe that the usefulness of providing extra linguistic materials via RAG depends on the complexity of the task addressed. While in the text and segmentation tiers, external information sources do not yield substantial improvements,

their impact is statistically significant in the tagging tier, which constitutes the highest level of difficulty within the glossing process.

Consequently, these findings provide an empirical framework to guide the work and methodological decisions of those dedicated to glossing data-scarce languages, like Mapudungun. The evidence obtained delivers concrete results to guide the course of action in those projects and research endeavors seeking to accelerate the processes of interlinear glossing generation for this type of languages.

7 Ethical considerations

For the development of this research, an annotated corpus was constructed from the online resource AVENUE Project (Levin et al., 2002). The use of this resource, enriched with morphological annotation, is proposed as a methodological strategy for the reuse of preexisting materials. This approach is substantiated as an alternative to over-intervention in speaking communities, relying exclusively on already consolidated data.

In turn, the integration of the resources used seeks to establish a pathway that facilitates the work of glossers of this language. The purpose of this experimental design does not aim to replace the work of human annotators, but rather to optimize and accelerate their processes, thus allowing the generation of a larger volume of useful corpus for the study of Mapudungun morphology. Under this logic, we propose using a single annotated corpus rather than combining multiple materials, as this approach requires significantly fewer computational resources. This reduction of the processing footprint is fundamental from an ethical perspective, considering that marginalized communities are precisely those who suffer the most from the environmental costs involved in training LLMs (Bender et al., 2021).

Finally, the application of the RAG architecture and the identification of the most efficient technique represent a contribution to the democratization of access to LLM-based technologies. Given its low technical barrier, this strategy enables non-engineering profiles, such as linguists, documentarians, and community members, to leverage these technologies for their own languages. Ultimately, we hope that the results and techniques implemented here will contribute to the efforts of documentation, study, and revitalization of languages globally.

Acknowledgments

Claudio Gutierrez thanks funding from IMFD, ANID - Millennium Science Initiative Program - Code ICN17_002.

Arianna Bisazza is funded by the Talent Programme of the Dutch Research Council (NWO) under project VI.Vidi.221C.009.

Felipe Hasler and Anaís Almendra were supported by ANID FONDECYT Project No. 1251110 *Estructura argumental y cambio de valencia en lenguas de los andes del sur*.

Anaís Almendra thanks to National Center for Artificial Intelligence CENIA FB210017, Basal ANID for their support during the development of this research.

References

Anaís Almendra. 2025. ¿cómo estudiamos y aportamos a las lenguas indígenas desde la computación? In Nicolás Albornoz, Valentina Espinoza, Camila Cortez, and Joaquín Vásquez, editors, *Cartografías lingüísticas: Un abordaje desde y hacia la interdisciplinariedad*, pages 169–182. Ediciones Colegas.

Félix José de Augusta. 2017. *Diccionario mapudungún-español, español-mapudungún*. Universidad Católica de Temuco and Centro de Investigaciones Diego Barros Arana. Compilado y editado por Belén Villena Araya.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623.

Andrés Chandía. n.d. KMT – Kümewirin Mapudüngun Trapümwe: Unificador ortográfico de mapudüngun. Sin fecha. Accedido: 08-Abril-2026.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2015. The Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses.

Sara Court and Micha Elsner. 2024. Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA. Association for Computational Linguistics.

Micha Elsner and David Liu. 2025. Prompt and circumstance: A word-by-word LLM prompting approach to interlinear glossing for low-resource languages. In *Proceedings of the 22nd SIGMORPHON workshop*

on Computational Morphology, Phonology, and Phonetics, pages 1–14, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Michael Fortescue, Marianne Mithun, and Nicholas Evans. 2017. Introduction. In Michael Fortescue, Marianne Mithun, and Nicholas Evans, editors, *The Oxford Handbook of Polysynthesis*, pages 1–23. Oxford University Press.

Michael Ginn, Mans Hulden, and Alexis Palmer. 2024a. Can we teach language models to gloss endangered languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5861–5876, Miami, Florida, USA. Association for Computational Linguistics.

Michael Ginn and Alexis Palmer. 2023. Robust generalization strategies for morpheme glossing in an endangered language documentation context. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, page 89–98. Association for Computational Linguistics.

Michael Ginn, Lindia Tjuatja, Taiqi He, Enora Rice, Graham Neubig, Alexis Palmer, and Lori Levin. 2024b. GlossLM: A massively multilingual corpus and pretrained model for interlinear glossed text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12267–12286, Miami, Florida, USA. Association for Computational Linguistics.

Lucía Golluscio and Felipe Hasler. 2017. Jerarquías referenciales y alineamiento inverso en Mapudungun. *RASAL Lingüística*, pages 69–93.

Google LLC. 2025. Términos de servicio específicos de Google Cloud (es-419). Sin fecha. Accedido: 08-Abril-2026.

Hans Gundermann, Jaqueline Canihuan, Alejandro Clavería, and César Faúndez. 2011. El mapuzugun, una lengua en retroceso. *Atenea (Concepción)*, (503):111–131.

Instituto Nacional de Estadísticas. 2024. Resultados – censo 2024. Accedido: 15 de abril de 2026.

Lori Levin, Rodolfo Vega, Jaime G. Carbonell, Ralf D. Brown, and Carolina Huenchullan. 2002. Data collection and language technologies for Mapudungun. In *Proceedings of the LREC-2002 Workshop*, Las Palmas, Spain.

Eduardo Llanquimán, Cristian Lagos, and Elizabeth Torrico-Ávila. 2025. Los grafemarios de la lengua mapuche como herramienta de revitalización lingüística: una revisión bibliográfica. *Revista de Lenguas y Literatura Indoamericanas*, 26(01–02):28–57.

Max Planck Institute for Psycholinguistics. n.d. ELAN. Accessed: 08-Abril-2026.

SIL International. 2026. Ethnologue: Languages of the world. Accessed: April 13, 2026.

SIL International. 2026. FieldWorks. Accessed: 08-Abril-2026.

Ineke Smeets. 2008. *A Grammar of Mapuche*, volume 41 of *Mouton Grammar Library*. Mouton de Gruyter, Berlin and New York.

Piyapath T. Spencer and Nanthipat Kongborrirak. 2025. Can LLMs help create grammar?: Automating grammar creation for endangered languages with in-context learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10214–10227, Abu Dhabi, UAE. Association for Computational Linguistics.

Anthony C. Woodbury. 2011. Language documentation. In Peter K. Austin and Julia Sallabank, editors, *The Cambridge Handbook of Endangered Languages*, pages 159–186. Cambridge University Press.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.

Hongpu Zhu, Yuqi Liang, Wenjing Xu, and Hongzhi Xu. 2025. Evaluating large language models for in-context learning of linguistic patterns in unseen low resource languages. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 414–426, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fernando Zúñiga. 2017. Mapudungun. In Michael Fortescue, Marianne Mithun, and Nicholas Evans, editors, *The Oxford Handbook of Polysynthesis*, pages 1–24. Oxford University Press.

Fernando Zúñiga. 2019. Grammatical relations in Mapudungun. In A. Witzlack-Makarevich and B. Bickel, editors, *Argument selectors: A new perspective on grammatical relations*, pages 39–67. John Benjamins Publishing Company.

A Appendix

A.1 Prompt used

"Eres un lingüista experto en el análisis morfológico del mapudungun, una lengua genealógicamente aislada hablada en Chile y Argentina. Debes aplicar el formato de las Glosas de Leipzig para entregar tus análisis morfológicos del mapudungun al usuario. Los inputs del usuario pueden ser de los siguientes formatos:

Ejemplo 1:
Input del usuario: ilotukelan
Tu respuesta al usuario:
ilotukelan
ilo-tu-ke-la-n

carne-VRBZ-HAB-NEG-IND.1SG
“Yo no como carne”
Ejemplo 2:
Input del usuario: chaw
Tu respuesta al usuario:
chaw
chaw
padre
“padre”
Ejemplo 3:
Input del usuario: tañi chaw müley tañi rukamew
Tu respuesta al usuario:
tañi chaw müley tüfi rukamew
ta= ñi chaw müle-y tüfey ruka-mew
DET= 1SG.POSS padre estar-IND.[3] ese casa-PPOS
“Mi padre está en esa casa”

Si el input es un elemento que no presenta morfemas (por ejemplo, un sustantivo sin afijos), no realices segmentación morfológica y repite la forma tal como aparece, tal como se presenta en el ejemplo 2.

Para la entrega de tus respuestas piensa paso a paso para realizar la segmentación y análisis del input entregado por el usuario, pero solo entrega tu respuesta, es decir, el análisis. Ciñe tus respuestas a los ejemplos entregados. Las equivalencias entre el mapudungun y las etiquetas y traducción debe ser en español, no en inglés u otra lengua.

Cuando el usuario te entregue un texto en mapudungun responde solo con la estructura analizada según los lineamientos de las Glosas de Leipzig. No añadas información adicional, solo entrega tu análisis.”