Multilingual Automatic Extraction of Linguistic Data from Grammars

Albert Kornilov
National Research University
Higher School of Economics
Moscow, Russia
albert.kornilov801@gmail.com

Abstract

One of the goals of field linguistics is compilation of descriptive grammars for relatively little-studied languages. Until recently, extracting linguistic characteristics from grammatical descriptions and creating databases based on them was done manually. The aim of this paper is to apply methods of multilingual automatic information extraction to grammatical descriptions written in different languages of the world: we present a search engine for grammars, which would accelerate the tedious and time-consuming process of searching for information about linguistic features and facilitate research in the field of linguistic typology.

1 Introduction

This work is dedicated to methods of information extraction, one of the subtasks of natural language processing. Methods of information extraction are widely used to create search engines. In addition to web services designed to search for internet websites that are relevant to the user's request, there is a need for highly specialized search engines for scientific publications, including linguistic ones.

One of the publication types in field linguistics is a descriptive grammar, which is a description of phonetic, morphological, syntactic, semantic and other characteristics of a particular language. Until recently, extracting language characteristics from descriptive grammars and creating databases based on them was done manually. For instance, The World Atlas of Language Structures[1], originally published as a book (Haspelmath et al., 2005), contains information on 144 characteristics for over 2600 languages.

Searching for information about a multitude of features is a long and labor-intensive process, albeit a portion of grammars is available not only in paper form, but also in digitized form: grammars from different time periods (from missionary grammars to modern papers) created by researchers from different countries do not have a single structure. Furthermore, a simple search for a word in a document can return dozens of occurrences, and not all of them will be relevant to the query.

The purpose of this work is to create a search engine for grammars, which would facilitate and speed up the process of finding information about language characteristics. The paper considers two methods of information extraction (BM25 and a reranking model based on BERT). The materials for the demonstration of the search engine include grammars presented on Google Drive[2].

Section 2 will analyze the already existing works pertaining to the task of automatic extraction of data from grammars; in Section 3, the methods used for data preprocessing will be described. Section 4 will discuss the two methods used for information extraction. In Section 5, we will compare the results obtained using the two methods and demonstrate the features of the search engine web application.

2 Review of Existing Approaches

At the moment, the subject of automatic information extraction of data from grammars is relatively little-studied. Several scientific papers regarding the methodology for extracting information from grammars using frame semantic parsers have been published by members of Språkbanken, a research and development unit at the University of Gothenburg, Sweden: (Virk et al., 2017; Virk et al., 2019; Virk et al., 2020; Virk et al., 2021). The methodology proposed by Språkbanken is illustrated in (Virk et al., 2019) using the following hypothetical sentence from a grammar as an example:

> The adjectives follow the noun they qualify.

[1] https://wals.info/

[2] https://drive.google.com/drive/folders/1FUunY_30HCKUsSixwczsxRaJ71fAb9Ii

Proceedings of the Second Workshop on NLP Applications to Field Linguistics, pages 86–94

An answer to the following question: "What is the order of adjectives and nouns in the language?" is to be chosen from the values "noun-adjective", "adjective-noun", and "both". Based on the labels assigned to the predicate "follow", the subject "adjectives", and the object "noun" by the semantic parser, the option "noun-adjective" is selected as an answer to the question and entered into the database.

Semantic parsers based on tagged text corpora are usually not sufficient to describe semantic frames found in grammars. (Virk et al., 2020) describes the functionality of a highly specialized semantic parser for linguistic publications, created on the basis of LingFN. LingFN is a corpus of grammars in English with annotated semantic frames, described in in (Malm et al., 2018). Extracting information from grammars written in languages other than English would require creating highly specialized parsers for each of the languages. Since a single specialized parser is not a multilingual solution, further this paper will discuss methods that are not based on frame semantics.

A simpler method is used in (Hammarström et al., 2020): to find out if a certain phenomenon is present in the language, the frequency of the corresponding term in the text of the grammar is counted. Occurrences of a term in the context of negative polarity items ("in language X [there is no phenomenon Y] | [missing category Y] | [category Y not found]") are excluded. Based on the distribution of occurrences of each term in grammars, a frequency threshold is calculated. Only terms with a frequency above the threshold are categories potentially present in the language. This method does not require significant time spent on annotating corpora and is universal for grammars written in any language, which greatly facilitates automatic creation of databases of linguistic characteristics.

However, the methods described in (Virk et al., 2017; Virk et al., 2019; Virk et al., 2020; Virk et al., 2021; Hammarström et al., 2020) are effective for building language databases in the form of tables, where at the intersection of a row with the name of the language and the column with the name of the category is an answer to a question (for example, "noun-adjective") or a truth value indicating presence or absence of a particular category in the language.

The table format does not fully meet the goals of our work, since it is not enough for a search engine to extract a single truth value; it is more crucial to extract a paragraph which describes the specific features of the desired language characteristic, together with the glosses and examples. Therefore, it has been decided to use methods that rank documents (paragraphs) according to their relevance to the search query entered by the user in order to return the original paragraph from the grammar in response to the query. The chosen approach does not perform any final feature extraction, but leaves the ultimate decision to the linguist.

3 Data

The grammars from which the search engine application extracts information are presented on Google Drive in the Grammars folder. The source code of the application is available on[3].

Each grammar is presented in a .pdf file. The table grammars_database.xslx (stored in the source code repository) contains meta-information for each grammar: the full path to the file, availability of an OCR layer ("Searchable"/"Not searchable"), the language described in the grammar, and the language the grammar is in. Initially, some files did not have an OCR layer. Such files were processed using the ocrmypdf[4] library.

For the subsequent information extraction, the contents of each file were preprocessed. The grammars were parsed using the pdftotext[5] library and divided into paragraphs. A combination of two spaces was taken as a separator. After separation, extra spaces were removed from the beginning and end of each paragraph. Since there are frequent cases of a paragraph being split between two pages, after separation, each pair of adjacent paragraphs was checked: if the second paragraph does not start with a capital letter and/or the first one does not end with a dot, ellipsis, question mark or exclamation mark, then they were connected again into a single paragraph.

Further, each paragraph was divided into tokens using the spaCy[6] library. spaCy was chosen because it currently implements text preprocessing methods for 22 languages. The paragraphs underwent tokenization; punctuation marks, numbers, and stop words were removed. The lists of tokens and their corresponding page numbers were saved

[3] https://github.com/grammars-data-extraction/linguistic_data_extraction

[4] https://github.com/ocrmypdf/OCRmyPDF

[5] https://pypi.org/project/pdftotext/

[6] https://spacy.io/

in .json files in the Grammars_Page_Numbers folder in the repository in order that the search algorithm would work with preprocessed data and not with the original .pdf file.

After the tokenization, the paragraphs were lemmatized, and the lists of lemmas were saved as .json files in the Grammars_Lemmas folder.

4 Methods for Ranking Paragraphs by Relevance to the Query

After the data has been divided into paragraphs and preprocessed, the search engine itself was implemented. It accepts a query from the user, determines which of the paragraphs are relevant to the query, and returns them. To calculate relevance, this paper uses the BM25 algorithm and a combination of BM25 with BERT embeddings.

4.1 BM25

BM25 is a family of functions that assign a relevance score to the search query to each of the documents (in our case, each of the paragraphs). The paper uses the function described in (Trotman et al., 2012) and implemented in the BM25Okapi class of the rank-bm25[7] library:

$$BM25(Q, d) =$$

$$= \sum_{t \in Q}^{n} IDF(t) \frac{(k_i + 1) \cdot tf_{td}}{tf_{td} + k_1 \cdot (1 - b + b \cdot \frac{L_d}{L_{avg}})}$$

$$IDF(t) = log \frac{N - df_t + 0.5}{df_t + 0.5}$$

Q: the query entered by the user;

d: the paragraph for which the relevance is determined;

tf_{td}: the number of occurrences of the token in the paragraph;

df_t: the number of paragraphs in the grammar that contain the token;

N: the total number of paragraphs in the grammar;

L_d: the number of tokens in the paragraph;

L_{avg}: the mean of the number of tokens for all paragraphs.

4.2 BERT

Among many other NLP tasks, BERT can be used to rank documents by relevance to a query: it assigns a vector to the query and to each paragraph from the document. The more relevant the query and the paragraph are to each other, the greater the cosine similarity is between them. For this paper, the bert-base-multilingual-cased model[8] was used, which supports 104 languages.

Since creating sentence embeddings using BERT and calculating the cosine similarity for each paragraph has a greater algorithmic complexity than BM25, in order to optimize the running time a decision was made to use the combined BM25 + BERT reranking method, decribed in (Nogueira and Cho, 2019).

4.3 BM25 + BERT Reranking

The combined method is structured as follows: using a simpler ranking method (in our case, BM25), n paragraphs relevant to the query are selected from the document, and afterwards k paragraphs ($k < n$) are selected from them using a more algorithmically complex method (in our case, the BERT embedder). When developing a search engine for grammars, it was decided to use only BM25 and the combined method, refraining from using BERT without BM25, since a search engine, unlike algorithms used for creating databases, works in real time, and significant time delays after the user enters a query are unacceptable.

5 A Solution to the Problem of Multilinguality

Since the goal of this paper is to create a search engine that is not exclusive to grammars written in English, it is necessary to implement an algorithm for automatically translating the user's query from English into other languages. Google Translate and libraries based on it are not suitable for this task: results for translating linguistic terms into other languages are in most cases incorrect. For instance, the term "reduplication" is translated from English into German as "Verdoppelung" ("doubling"), not "Reduplikation".

Consequently, it was decided to use another method of translating linguistic terms into different languages: using Wikipedia.

The HTML code of the Wikipedia page called "Reduplication" in English contains links to pages about the same term in other languages. The method for extracting page titles in the desired language was implemented using the beautiful-

[7] https://github.com/dorianbrown/rank_bm25

[8] https://huggingface.co/bert-base-multilingual-cased

soup4[9] library. In addition to titles of articles, it was decided to extract their summaries using the Wikipedia[10] library. Example: summary for the term "Ergative case"[11] in English (accessed 23 Feb. 2023):

> In grammar, the ergative case (abbreviated ERG) is the grammatical case that identifies a nominal phrase as the agent of a transitive verb in ergative–absolutive languages.

Using a summary as a query increases the likelihood of extracting a relevant paragraph from the grammar, as it may contain words, linguistic terms, and abbreviations that are often found in the context of the term requested by the user: for instance, the summary for the Wikipedia article "Ergative case" contains the abbreviation "ERG" and related terms "agent", "transitive verb", and "absolutive".

Each summary is extracted from Wikipedia, tokenized, and lemmatized only once. The summaries themselves and the lists of tokens corresponding to them are saved in .json files in the Grammars_Summaries folder in the repository.

6 Results

6.1 The Functionality of the Search Engine

In this section, the functionality of the search engine will be demonstrated on the example of the query "Plural" and a grammar of the Angami language (McCabe, 1887). The BM25 algorithm returns the five most relevant paragraphs from the grammar; in the combined algorithm, BM25 selects ten paragraphs and afterwards BERT selects the five most relevant ones out of them. The extracted paragraphs are shown in Table 1. In this particular case, the set of paragraphs selected by the two methods is the same; however, the paragraph containing the information about the most common method for expressing the singular and the plural in Angami (lack of marking) was placed higher by BM25 than by the combined method.

The interface of the search engine application is presented in Figure 1. The user is prompted to select an algorithm from the top menu and enter the name of the language and the desired linguistic feature. The application returns the five most relevant paragraphs from each grammar describing the

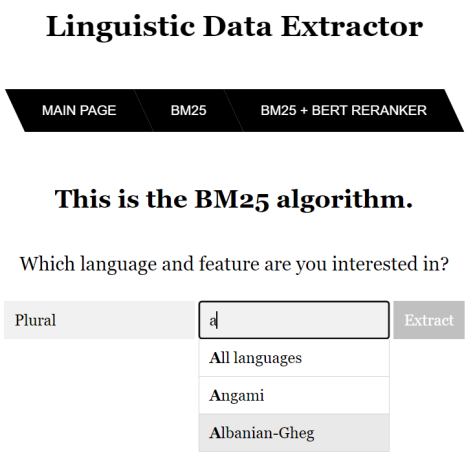


Figure 1: The web interface of the search engine.

language. After every paragraph, its source pages from the file with the grammar are displayed, in order for the user to be able to instantly see the relevant context and glosses with examples. The repository stores only a part of the grammars; the remaining grammars are copied from the Google Drive using the rclone[12] script upon being requested by the user.

The features currently available in the demo version of the search engine are the following: Reduplication, Plural, Declension, Nominative case, Ergative case, Absolutive case, Accusative case, Word order. Any feature with its own page on Wikipedia can potentially be integrated into the functionality of the application.

The demo version supports extraction of characteristics of the following languages: Samaritan Aramaic, Lule, Angami, Javanese, Sangir, Pampangan, Hawaiian, Albanian-Gheg, Karelian, Tibetan. Since for typological research entering the language name should be non-mandatory, an additional option "All languages" has been added to the interface.

6.2 A Qualitative Evaluation

The search engine has been tested on over 500 grammars written in some of the most spoken European languages (English, German, French, Spanish, Italian, Russian, Dutch). The testing procedure included extracting information on each of the linguistic features available in the demo version from each of the grammars.

While a quantitative evaluation of the search engine (e. g. calculation of metrics) is difficult to

[9] https://pypi.org/project/beautifulsoup4/
[10] https://pypi.org/project/wikipedia/
[11] https://en.wikipedia.org/wiki/Ergative_case
[12] https://rclone.org

Rank	BM25	BM25 + BERT Reranking
1	In these examples no inflections nor descriptive words are em ployed to denote the singular or plural.	The plural is the same as the third person plural of the personal pronoun Hāko these.
2	The plural is the same as the third person plural of the personal pronoun Hāko these.	As a general rule , however, when it is desired to clearly mark the singular and plural, the numeral adjective po = " one," is used to denote the singular, and the suffix ko the plural : I saw a dog in your house . Ā unki nu tefüh po ngulé.
3	As a general rule , however, when it is desired to clearly mark the singular and plural, the numeral adjective po = " one," is used to denote the singular, and the suffix ko the plural : I saw a dog in your house . Ā unki nu tefüh po ngulé.	In these examples no inflections nor descriptive words are em ployed to denote the singular or plural.
4	The reflexive pronoun " self," " myself," " himself, " &c. , is ren dered by the word the or tha . It is not declined, and has but one form for the singular and plural I came myself = A the vorwe.	The reflexive pronoun " self," " myself," " himself, " &c. , is ren dered by the word the or tha . It is not declined, and has but one form for the singular and plural I came myself = A the vorwe.
5	This section treats of nouns under the heads "Gender," " Number " and " Case." I.- GENDER .	This section treats of nouns under the heads "Gender," " Number " and " Case." I.- GENDER .

Table 1: Comparison of BM25 and BM25 + BERT Reranking on the example of the query “Plural” and the grammar (McCabe, 1887).

conduct due to the fact that final feature extraction is not performed, the empirical results show the following:

(i) Readability of outputs with glosses leaves room for improvement. This problem is mitigated by outputting the source pages from the file with the grammar. An example of an output with glosses and the corresponding fragment of the source page are given in Figure 2 and Figure 3 in Appendix A respectively.

(ii) It is not the case that division of grammars into paragraphs is optimal for all descriptive grammars, since they lack common structure: paragraphs that are overly long (containing large blocks of glosses and examples) or overly short (containing only one of the terms from the query) occasionally occur among the results. Outputting the source pages partially mitigates this problem as well, since the majority of the overly short paragraphs are titles of sections and subsections in the descriptive grammars. An example of an overly long output is given in Figure 4, and an example of an overly short output with its source page fragment is presented in Figure 5 in Appendix A.

6.3 Coverage of Linguistic Features in Wikipedia

In order to provide a quantitative evaluation of Wikipedia as the basis of the search engine, we took the list of linguistic features from The World Atlas of Language Structures[13] and manually annotated their corresponding Wikipedia entries. The titles of the entries were translated into German, French, Spanish, Italian, Russian, and Dutch using the Wikipedia[14] library. Statistics on the coverage of the linguistic features in Wikipedia articles have been calculated for all seven languages. The result of the evaluation is given in Table 2, and the table with the annotations (Coverage_of_linguistic_phenomena_in_Wikipedia.xlsx) is available in the source code repository.

Table 2 shows that 34 features out of 192 have their own Wikipedia entries in the English language. Several features are expressed by a com-

[13] https://wals.info/

[14] https://pypi.org/project/wikipedia/

Coverage	Yes	Partially	No
German	22	131	39
French	18	124	50
Spanish	17	137	38
Italian	19	91	82
Russian	18	131	53
Dutch	14	127	51
Average	**18**	**123.5**	**52.2**
English	**34**	**143**	**15**

Table 2: Coverage of linguistic features in Wikipedia articles (accessed 30 March 2023).

bination of Wikipedia entries instead of a single one. For instance, feature 52A, Comitatives and Instrumentals, is covered by three entries: Comitative case, Instrumental case, and Instrumental-comitative case. The average number of features marked with "Yes" for the other six languages is 18 (only 52.9% of the corresponding number for English), while the average number of missing features for the six languages is 348% of the number of missing features in the English Wikipedia.

Since the search engine outputs paragraphs and leaves the final decision to the linguist, the limitations on queries are less strict than for models intended for final feature extraction. Consequently, we introduce the third category, "Partially", in order to mitigate the imbalance: the linguistic features belonging to it are more specific than the corresponding articles. For example, feature 36A, The Associative Plural, has no matching article in the English Wikipedia and therefore corresponds to the article with the title "Plural".

The advantage of using Wikipedia is coverage of linguistic features that are not present in WALS: for instance, Assimilation, Aorist, Semelfactive, Mass noun, Cardinal numeral, and Vowel harmony.

7 Conclusion

This paper presents a search engine web application that allows automatic extraction of information from grammars written in different languages of the world. Two information extraction methods (classical BM25 and the combined method based on BM25 + reranking with BERT) have been compared to each other regarding the task of extracting linguistic information relevant to the user's query. The search algorithm has been integrated with Wikipedia.

The implemented system makes it possible to get an impression of the total complexity of the task of automatic information extraction from scientific publications and opens up the possibility for massive automated research in the field of linguistic typology, facilitating the routine task of extracting information from grammatical descriptions and allowing researchers to direct the time to solving problems that require advanced expertise.

Limitations

The work presented in the paper has potential limitations. To begin with, particular attention should be paid to normalization of terminology, which varies in grammatical descriptions belonging to different scientific schools and eras. Furthermore, the multilinguality of the system requires further development: testing of the search engine was only carried out for grammars written in some of the most spoken European languages, due to grammars in other languages being accessible in significantly smaller quantities. Moreover, the performance of the search engine can potentially be improved by using a faster system (for instance, S3) rather than accessing the Google Drive storage through rclone. In addition, while using Wikipedia is a potential solution to the problem of multilinguality, it is a user-generated source, and using it may potentially yield unexpected or unreliable results. Ultimately, the graphical interface can be supplemented with tools for collecting and analyzing user feedback. To further improve user experience, it is planned to carry out further testing of the system on experts conducting research in the field of linguistic typology.

Ethics Statement

The dataset originally used for testing the search engine partially consisted of grammars subject to copyright. In order to avoid any form of copyright infringement, we left only ten grammars in Google Drive and in the source code repository. The grammars are stored in the dataset solely for the purpose of demonstrating the functionality of the search engine. Each of the ten grammars is part of the open-access set maintained by Språkbanken[15], is at least 100 years old, and is not subject to copyright.

[15] https://spraakbanken.gu.se/blogg/index.php/2020/04/07/a-multilingual-annotated-corpus-of-worlds-natural-language-descriptions/

Acknowledgements

First and foremost, we would like to express our gratitude to Oleg Serikov (National Research University Higher School of Economics, AIRI, MIPT, RAS Linguistics) for his continuous support and guidance throughout the entire process of creating the search engine. In addition, we are grateful to the anonymous reviewers for their suggestions, which substantially improved this paper. Moreover, we would like to extend our thanks to Mason Gilliam (University of Texas at Austin) for valuable insights regarding information retrieval algorithms.

References

Antonio Machoni de Cerdeña. 1877. *Arte y vocabulario de la lengua lule y tonocoté...* Reimpreso por Pablo E. Coni.

Harald Hammarström, One-Soon Her, and Marc Allassonnière-Tang. 2020. Term spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions. In *Selected contributions from the Eighth Swedish Language Technology Conference (SLTC-2020)*, pages 27–34.

Martin Haspelmath, Matthew S Dryer, David Gil, and Bernard Comrie. 2005. *The world atlas of language structures*. OUP Oxford.

Itziar Laka. 1996. *A brief grammar of Euskara, the Basque language*. Univ. of the Basque Country.

Per Malm, Shafqat Mumtaz Virk, Lars Borin, and Anju Saxena. 2018. Lingfn: Towards a framenet for the linguistics domain. In *11th edition of the Language Resources and Evaluation Conference, 7-12 May 2018, Miyazaki (Japan)*, pages 37–43.

Robert Blair McCabe. 1887. *Outline Grammar of the Angāmi Nāgā Language: With a Vocabulary and Illustrative Sentences*. Superintendent of Government Printing.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Andrew Trotman, Xiangfei Jia, and Matt Crane. 2012. Towards an efficient and effective search engine. In *OSIR@ SIGIR*, pages 40–47.

Shafqat Mumtaz Virk, Lars Borin, Anju Saxena, and Harald Hammarström. 2017. Automatic extraction of typological linguistic features from descriptive grammars. In *Text, Speech, and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings 20*, pages 111–119. Springer.

Shafqat Mumtaz Virk, Daniel Foster, Azam Sheikh Muhammad, and Raheela Saleem. 2021. A deep learning system for automatic extraction of typological linguistic information from descriptive grammars. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1480–1489.

Shafqat Mumtaz Virk, Harald Hammarström, Lars Borin, Markus Forsberg, SK Wichmann, Maxim Ionov, John P McCrae, Christian Chiarcos, Thierry Declerck, Julia Bosque-Gil, et al. 2020. From linguistic descriptions to language profiles. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 23–27.

Shafqat Mumtaz Virk, Azam Sheikh Muhammad, Lars Borin, Muhammad Irfan Aslam, Saania Iqbal, and Nazia Khurram. 2019. Exploiting frame-semantics and frame-semantic parsing for automatic extraction of typological information from descriptive grammars of natural languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1247–1256.

Michael Weiers. 2013. *Die Sprache der Moghol der Provinz Herat in Afghanistan: Sprachmaterial, Grammatik, Wortliste*, volume 49. Springer-Verlag.

A Output Examples

Figure 2, Figure 3, Figure 4, and Figure 5 show examples of outputs of the search engine.

Grammars/Other/Basque Language, A Brief Grammar of Euskara (Laka).pdf Result №3.

Noun phrases are inflected for ergative case if they are subjects of transitive verbs:
(5)
a. zazpi gizonek ekarri dute pianoa
seven man-E brought have piano-det
'seven men have brought the piano'
b. etxeko txakurrak ikusi gaitu
house-of dog-det-E seen us-has
'the dog of the house has seen us'
c. Mirenen anaiek ez dakite kanta hau
Miren-gen brother-detpl-E not know song this
'Miren's brothers don't know this song'

Figure 2: An example of an output with glosses. Query: Ergative case. Method: BM-25. Language: Basque. Descriptive grammar: (Laka, 1996).

(5)

a. zazpi gizonek ekarri dute pianoa
seven man-E brought have piano-det
'seven men have brought the piano'

b. etxeko txakurrak ikusi gaitu
house-of dog-det-E seen us-has
'the dog of the house has seen us'

c. Mirenen anaiek ez dakite kanta hau
Miren-gen brother-det_{pl}-E not know song this
'Miren's brothers don't know this song'

(5a) illustrates our previous example Noun phrase as the subject of the transitive verb **ekarri** 'to bring'. (5b) illustrates a singular definite Noun phrase marked with ergative case, since it is the subject of the verb **ikusi** 'to see'. Finally, (5c) illustrates a plural definite Noun phrase inflected for ergative. Note that when the ergative marker **k** attaches to the plural determiner **ak**, the resulting form is **ek**. Again, this Noun phrase is the subject of a transitive verb, in this case, **jakin** 'to know'. Along these lines, it must also be noted that the combination of the proximity determiner **ok** and ergative **k** yields **ok**. Thus, regarding Noun phrases ending in the proximity dterminer **ok**, the absolutive and the nominative forms are identical; this is called 'syncretism'.

Figure 3: A fragment of a source page with glosses from a file with a grammar. Query: Ergative case. Method: BM-25. Language: Basque. Descriptive grammar: (Laka, 1996).

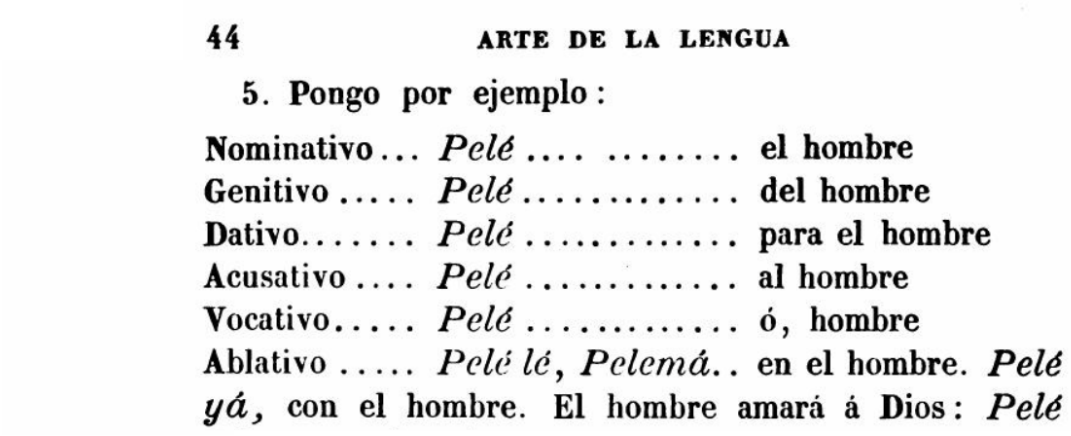
Grammars/Mongolic/Moghol, Die Sprache der (Weiers).pdf
Result №3.

Morphologie 113
B. Nomina
1. Pluralbildung 1
Die Moghol-Sprache besitzt die Pluralsuffixe -df-t, -nud und -s. Vielfach wird trotz eines verwendeten Pluralsuffixes nur die Einzahl zum Ausdruck gebracht. Um anzuzeigen, wann dies der Fall ist, wollen wir zwei Kategorien unterscheiden: 1. Die grammatisch-formale: Singular (S) und Plural (P). 2. Die semantische: Einzahl (E) und Mehrzahl (M). Hieraus ergeben sich hinsichtlich der Pluralbezeichnung durch Suffixe die Kombinationen EP und MP. Bei Mehrzahlwortern haben wir die Kombination MS. Die semantische Kategorie bezeichnen wir bei den Kombinationen immer als die erste. Die Kombination EP hat oftmals Kollektivbedeutung, worunter wir entweder die Bezeichnung einer Gesamtheit, z. B. "der Mensch" im Sinne der gesamten Menschheit, oder einer Gesamtgruppe verstehen, z.B. "Hirse" als Gesamtgruppe innerhalb verschiedener Getreidesorten. Als Belege fiihren wir nachstehend meist nur Einzelworter und deren Funktion an, da das Gesamtbeispiel ohne Schwierigkeiten in den Sprachmaterialien aufzufinden ist.
1. -df-t
Das weitaus haufigste Suffix -d steht iiberwiegend im Nominativ PI. von vokalisch auslautenden und n-Stammen, deren n beim Suffixantritt abfallt. -t steht nach den gleichen Stammen, jedoch meist in einem der obliquen Kasus oder vor enklitischen Personalpronomina, kurz als Silbenbeginn vor einem folgenden, oft akzentuierten Vokai. Das Suffix bezieht

Figure 4: An example of an overly long output. Query: Plural. Method: BM-25. Language: Moghol. Descriptive grammar: (Weiers, 2013).

Grammars/Arte y vocabulario de la lengua lule o tonocoté.pdf
Result №5.

44ARTE DE LA LENGUA
5. Pongo por ejemplo :
Nominativo ... Pelé ....



44 ARTE DE LA LENGUA

5. Pongo por ejemplo:

Nominativo... *Pelé* .... el hombre
Genitivo ..... *Pelé* ..... del hombre
Dativo..... *Pelé* ..... para el hombre
Acusativo .... *Pelé* ..... al hombre
Vocativo..... *Pelé* ..... ó, hombre
Ablativo ..... *Pelé lé, Pelemá..* en el hombre. *Pelé yá,* con el hombre. El hombre amará á Dios: *Pelé*

Figure 5: An example of an overly short output with its source page. Query: Nominative case. Method: BM-25. Language: Lule. Descriptive grammar: (de Cerdeña, 1877).