Keyword Spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions

Harald Hammarström
Uppsala University
harald.hammarstrom@lingfil.uu.se

One-Soon Her
National Chengchi University
onesoon@gmail.com

Marc Tang
University Lumière Lyon 2
marc.tang@univ-lyon2.fr

1 Introduction

The present paper addresses extraction of information about languages of the world from digitized full-text grammatical descriptions. The typical instances of such information-extraction tasks are so-called typological features, e.g., whether the language has tone, prepositions, SOV basic constituent order and so on, similar in spirit to those found in the database WALS `wals.info` (Dryer and Haspelmath, 2013).

Given its novelty, only a few embryonic approaches (Virk et al., 2019; Wichmann and Rama, 2019; Macklin-Cordes et al., 2017; Hammarström, 2013; Virk et al., 2017) have addressed the task so far. Of these, some a keyword-based and some combine keywords with more elaborate analyses of the source texts such as frame-semantics (Virk et al., 2019). All approaches so far described require manual tuning of thresholds and/or supervised training data.

For the present paper, we focus on the prospects of keyword extraction, but in a way that obviates the need for either manual tuning of thresholds or supervised training data. However, this approach is limited the features for which a (small set of) specific keywords frequently signal the presence thereof, e.g., `classifier`, `suffix(es)`, `preposition(s)`, `rounded vowel(s)` or `inverse`. Keyword extraction is not applicable for features which are expressed in a myriad of different ways across grammars, e.g., as whether the verb agrees with the agent in person. It may be noted that the important class of word-order features, which are among the easiest for a human to discern from a grammar, typically belong to the class of non-keyword-signalled features unless there is a specific formula such as SOV or N-Adj gaining sufficient popularity in grammatical descriptions. Keyword-signalled features are, of course, far simpler to extract, but not completely trivial, and hence the focus the present study.

2 Data

The data for the experiments in this essay consists of a collection of over 10 000 raw text grammatical descriptions digitally available for computational processing (Virk et al., 2020). The collection consists of (1) out-of-copyright texts digitized by national libraries, archives, scientific societies and other similar entities, and (2) texts posted online with a license to use for research usually by university libraries and non-profit organizations (notably the Summer Institute of Linguistics). For each document, we know the language it is written in (the meta-language, usually English, French, German, Spanish or Mandarin Chinese), the language(s) described in it (the target language, typically one of the thousands of minority languages throughout the world) and the type of description (comparative study, description of a specific features, phonological description, grammar sketch, full grammar etc). The collection can be enumerated using the bibliographical- and metadata is contained in the open-access bibliography of descriptive language data at `glottolog.org`. The grammar/grammar sketch collection spans no less than 4 527 languages, very close to the total number of languages for which a description exists at all (Hammarström et al., 2018).

Figure 1 has an example of a typical source document — in this case a German grammar of the Ewondo [ewo] language of Cameroon — and the corresponding OCR text which illustrates the typical quality. In essence, the OCR correctly recognizes most tokens of the meta-language but is hopelessly inaccurate on most tokens of the vernacular being described. This is completely expected from the typical, dictionary/training-heavy, con-

temporary techniques for OCR, and cannot easily be improved on the scale relevant for the present collection. However, some post-correction of OCR output very relevant for the genre of linguistics is possible and advisable (see Hammarström et al. 2017). The bottom line, however, is that extraction based on keywords has good prospects in spite of the noise, while extraction of accurately spelled vernacular data is not possible at present.

3 Model

At first blush, the problem might seem trivial: simply look for the existence of the keyword and/or its relative frequency in a document, and infer the feature associated with the keyword. Unfortunately, to simply look for the existence of a keyword is too naive. In many grammars, keywords for grammatical features do occur although the language being described, in fact, does not exhibit the feature. For example, the grammar may make the explicit statement that there are "no X" incurring at least one occurrence[1] Also, what frequently happens is that comments and comparisons are made with other languages — often related languages or other temporal stages — than the main one being described[2]. Furthermore, there's always the possibility that a term occurs in an example sentence, text of reference title. However, such "spurious" occurrences will not likely be frequent, at least not as frequent as a keyword for a grammatical feature which actually belongs to the language and thus needs to be described properly. But how frequent is frequent enough? We will try to answer this question.

Let us assume that a full-text grammatical description consists of four classes of terms:

Genuine keywords: Terms that describe the language in question

Noise keywords: Descriptive terms that do not accurately describe the language in question (i.e., through remarks on other language or of things not present, as explained above)

Meta-language words: Words in the meta-language, e.g., *the*, *a*, *run* if the meta-language of description is English, that are not linguistic descriptive terms

Language-specific words: Words that are specific to the language being described but which do not describe its grammar. These can be morphemes of the language, place names in the language area, ethnographic terms etc.

We are interested in the first class, and in particular, to distinguish them from the second class. Except for rare coincidences, the words from these two classes do not overlap with the latter two, so they can be safely ignored when counting linguistic descriptive keywords. Now, a simple model for the frequency distribution of the keywords of a grammar $G(t)$ is that it is simply composed of sample of the "true" underlying descriptive terms according to their functional load $L(t)$ and a "noise" term $N(t)$, with a weight α balancing the two:

$$G(t) = \alpha \cdot L(t) + (1 - \alpha) \cdot N(t)$$

For example, if a language actually has duals, $L(dual) > 0$, perhaps close to 0.0 if the duals have low functional load, but higher if there is rampant dual agreement. For most languages, we expect the functional load of verbs to be rather high, perhaps $L(verb) > 0.2$. The purity level α, captures the fraction of tokens which actually pertain to the language, as opposed to those that do not. (Those tokens are typically of great interest for the reader of the grammar — they are "noise" only from the perspective of extraction as in the present paper.)

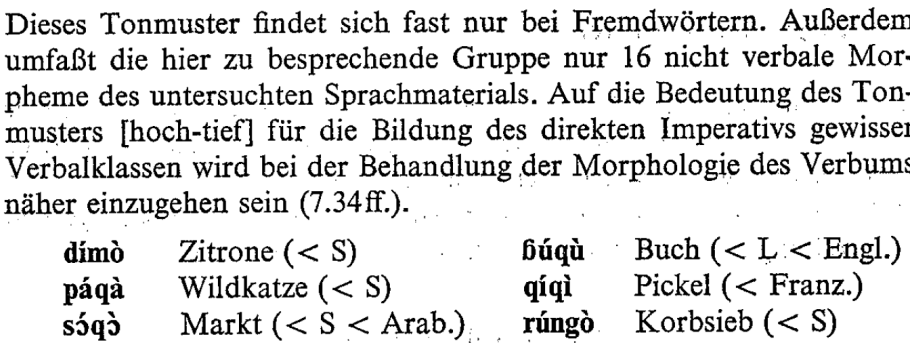
Suppose now that we have several different grammars for the *same* language. As they are the describing the same language, their token distributions are all (independent?) samples of the *same* $L(t)$, but there is no reason to suppose the noise level and the actual noise terms to be the same across different grammars. Thus we have:

$$\begin{aligned} G_1(t) &= \alpha_1 \cdot L(t) + (1 - \alpha_1) \cdot N_1(t) \\ G_2(t) &= \alpha_2 \cdot L(t) + (1 - \alpha_2) \cdot N_2(t) \\ &\cdots\cdots \\ G_n(t) &= \alpha_n \cdot L(t) + (1 - \alpha_n) \cdot N_n(t) \end{aligned}$$

Given actual distributions $G_1(t), \ldots, G_n(t)$ can we get at estimating the purity level, α_i of each one? The following procedure suggests itself. Take each term t for each grammar G_i and calculate the *generality* of its incidence $g^i_L(t)$ by comparing the

[1] One example is the Pipil grammar of Campbell (1985, 61) which says that Pipil has no productive postpositions.

[2] For example, Lorenzino (1998)'s description of Angolar Creole Portugues [aoa] contains a number of references to the fate of nouns that were masculine in Portuguese, yet the modern Angolar does not have masculine, or other, gender.



Dieses Tonmuster findet sich fast nur bei Fremdwörtern. Außerdem umfaßt die hier zu besprechende Gruppe nur 16 nicht verbale Morpheme des untersuchten Sprachmaterials. Auf die Bedeutung des Tonmusters [hoch-tief] für die Bildung des direkten Imperativs gewisser Verbalklassen wird bei der Behandlung der Morphologie des Verbums näher einzugehen sein (7.34ff.).

dímò	Zitrone (< S)	**ɓúqù**	Buch (< L < Engl.)
páqà	Wildkatze (< S)	**qíqì**	Pickel (< Franz.)
sɔ́qɔ̀	Markt (< S < Arab.)	**rúngò**	Korbsieb (< S)


```
Dieses Tonmuster findet sich fast nur bei Fremdwörtern. Außerdem
umfaßt die hier zu besprechende Gruppe m1r 16 nicht verbale
Morpheme des untersuchten Sprachmaterials. Auf die Bedeutun& des
Tonmus.ters [hoch-tief] für die Bildung des direkten Imperativs gewisser
Verbalklassen wird bei der Behandlung .der Morphologie des Verbums
nähereinzugehen sein (7.34ff.). Â·
Â·
Â·
dimo
paqa
s˜q;,
```


Figure 1: An example of OCR output.

fraction in $G_i(t)$ to the fraction of t in all other grammars for the language L.

$$g^i_L(t) = \frac{\frac{1}{n-1}\sum_{j \neq i} G_j(t)}{G_i(t)}$$

For example, suppose $G_i(dual) = 0.1$ for some grammar G_i. Maybe for two other grammars of the same language, $G_j(dual) = 0.01$ $G_k(dual) = 0.00$, this term barely occurs. The term *dual* would then have poor generality $g^i_L(dual) = \frac{1}{2} \cdot (0.00 + 0.01) = 0.005$. Grammars with lots of terms with poor generality have a high level of noise, and, conversely, grammars where all terms have a reciprocated proportion in other grammars are pure, devoid of noise. Thus, α_i can be gauged as:

$$\alpha_i = \frac{\sum g^i_L(t)}{\sum G_i(t)}$$

We now return to the question "how frequent is frequent enough?". We can now rephrase this as: does the frequency of a term in a grammar exceed its noise level (1-α)? Given that we know α_i for a grammar G_i, let us make the assumption that the fraction (1-α_i) of least frequent tokens are "noise". Simply subtracting the fraction (1-α_i) of tokens of the least least frequent types effectively generates a threshold $\bar{t}$ separating the tokens being retained versus those subtracted. For example, the grammar of Romanian by Cojocaru (2004) has an α_i of 0.81 and contains a total of 83 365 tokens. We wish to subtract $(1 - 0.81) \cdot 83365 \approx 15839$ tokens from the least frequent types. It turns out in this grammar that this removes all the types which have a frequency of 9 or less, rendering the frequency threshold $\bar{t} = 9$.

Let us look at an example. Table 1 has a list of grammars/grammar sketches of Romanian. Each grammar has a corresponding α purity level as described above, the total numbers of tokens, and the frequency threshold $\bar{t}$ induced by α and the token distribution. The last three columns concern the keywords `masculine`, `feminine` and `neuter` respectively. The cells contain the frequency of the corresponding keyword, as well as the fraction of pages on which it occurs. The fraction page occurrences is, of course, similar to, and highly correlated with the fraction of tokens but is often easier to interpret intuitively. Thus, for example, in Cojocaru (2004) the term `masculine` occurs 240 times in total, distributed onto 74 of the total 184 pages (≈ 0.40). The cells with a frequency that exceeds the threshold $\bar{t}$ for their corresponding grammar are shown in green, indicating that the keyword in question is probably genuinely describ-

Romanian [ron]

Grammar	α	$\sum G_i(t)$	$\bar{t}$	`masculine`	`feminine`	`neuter`
Cojocaru 2004	0.81	83365	9	240 0.40 (74/184)	259 0.46 (84/184)	124 0.23 (43/184)
Murrell and Ştefănescu Drăgăneşti 1970	0.72	95226	13	3 0.01 (3/424)	5 0.01 (5/424)	4 0.01 (3/424)
Gönczöl-Davies 2008	0.68	45423	9	63 0.13 (30/233)	75 0.15 (34/233)	23 0.06 (13/233)
Agard 1958	0.68	51239	9	23 0.08 (10/123)	28 0.08 (10/123)	0 0.00 (0/123)
Mallinson 1988	0.66	11019	4	18 0.30 (9/30)	18 0.23 (7/30)	18 0.17 (5/30)
Mallinson 1986	0.82	105018	6	119 0.15 (57/375)	110 0.12 (46/375)	25 0.03 (11/375)
Majority consensus				1	1	1

Table 1: Example grammars of Romanian and the frequencies of the keywords `masculine`, `feminine` and `neuter`.

ing the language. In this case, by majority consensus, we can infer that the language Romanian [ron] does have all three of masculine, feminine and neuter.

4 Evaluation

Thanks to a large manually elaborated database of classifier languages we were able to do a formal evaluation of extraction accuracy for this feature. We extracted the feature `classifier(s)` from 7 284 grammars/grammar sketches written in English spanning 3 220 languages. Each language was assessed as per the majority vote of the extraction result of each individual description, with ties broken in favour of a positive result. For languages where only one description exists, the noise-level was taken to be the average noise-level of grammars of other languages of similar size. A comparison between the Gold Standard database and the extracted data is shown in Table 2. The overall accuracy is 89.1%, to be compared with human inter-coder agreement on similar tasks (85.9% or lower, as per (Donohue, 2006) and (Plank, 2009, 67-68)).

Gold Standard	Keyword-Spotting	# languages
False	False	2357
True	True	512
True	False	317
False	True	34
		3 220

Table 2: Evaluation of keyword-spotting against a Gold Standard database of classifier languages.

5 Conclusion

We have described a novel approach to the extraction of linguistic information from descriptive grammars. The method requires only a keyword, but no manual tuning of thresholds or annotated training data. However, the approach can only address information that is associated with an enumerable set of specific keywords. When this is the case, a broad evaluation shows that the results match or exceed the far more time-consuming manual curation by humans.

Acknowledgments

This research was made possible thanks to the financial support of the From Dust to Dawn: Multilingual Grammar Extraction from Grammars project funded by Stiftelsen Marcus och Amalia Wallenbergs Minnesfond 2017.0105 awarded to Harald Hammarström (Uppsala University) and the Dictionary/Grammar Reading Machine: Computational Tools for Accessing the World's Linguistic Heritage (DReaM) Project awarded 2018-2020 by the Joint Programming Initiative in Cultural Heritage and Global Change, Digital Heritage and Riksantikvarieämbetet, Sweden.

References

Frederick B. Agard. 1958. A structural sketch of rumanian. *Language*, 34(3):7–127. Language Dissertation No. 26.

Lyle Campbell. 1985. *The Pipil Language of El Salvador*, volume 1 of *Mouton Grammar Library*. Berlin: Mouton de Gruyter.

Dana Cojocaru. 2004. *Romanian Grammar*. Durham: SEELRC.

Mark Donohue. 2006. Review of the the world atlas of language structures. *LINGUIST LIST*, 17(1055):1–20.

Matthew S. Dryer and Martin Haspelmath. 2013. The world atlas of language structures online. Leipzig: Max Planck Institute for Evolutionary Anthropology (Available online at http://wals.info, Accessed on 2015-10-01.).

Ramona Gönczöl-Davies. 2008. *Romanian: an essential grammar*. New York: Routledge, New York.

Harald Hammarström, Thom Castermans, Robert Forkel, Kevin Verbeek, Michel A. Westenberg, and Bettina Speckmann. 2018. Simultaneous visualization of language endangerment and language description. *Language Documentation & Conservation*, 12:359–392.

Harald Hammarström. 2013. Three approaches to prefix and suffix statistics in the languages of the world. Paper presented at the Workshop on Corpus-based Quantitative Typology (CoQuaT 2013).

Harald Hammarström, Shafqat Mumtaz Virk, and Markus Forsberg. 2017. Poor man's ocr post-correction: Unsupervised recognition of variant spelling applied to a multilingual document collection. In *Proceedings of the Digital Access to Textual Cultural Heritage (DATeCH) conference*, pages 71–75. Göttingen: ACM.

Gerardo A. Lorenzino. 1998. *The Angolar Creole Portuguese of São Tomé: Its Grammar and Sociolinguistic History*. Ph.D. thesis, City University of New York.

Jayden L. Macklin-Cordes, Nathaniel L. Blackbourne, Thomas J. Bott, Jacqueline Cook, T. Mark Ellison, Jordan Hollis, Edith E. Kirlew, Genevieve C. Richards, Sanle Zhao, and Erich R. Round. 2017. Robots who read grammars. Poster presented at Co-EDL Fest 2017, Alexandra Park Conference Centre, Alexandra Headlands, QLD.

Graham Mallinson. 1986. *Rumanian*. Croom Helm Descriptive Grammars. London: Croom Helm.

Graham Mallinson. 1988. Rumanian. In Martin Harris and Nigel Vincent, editors, *The Romance Languages*, pages 391–419. London: Croom Helm.

Martin Murrell and Virgiliu Ştefănescu Drăgăneşti. 1970. *Romanian*. Teach Yourself Books. London: English Universities Press.

Frank Plank. 2009. Wals values evaluated. *Linguistic Typology*, 13(1):41–75.

Shafqat Mumtaz Virk, Lars Borin, Anju Saxena, and Harald Hammarström. 2017. Automatic extraction of typological linguistic features from descriptive grammars. In Kamil Ekštein and Václav Matoušek, editors, *Text, Speech, and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings*, volume 10415 of *Lecture Notes in Computer Science*, pages 111–119. Berlin: Springer.

Shafqat Mumtaz Virk, Harald Hammarström, Markus Forsberg, and Søren Wichmann. 2020. The dream corpus: A multilingual annotated corpus of grammars for the world's languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 871–877. Marseille, France: European Language Resources Association, Marseille, France.

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 RANLP 2019*.

Søren Wichmann and Taraka Rama. 2019. Towards unsupervised extraction of linguistic typological features from language descriptions. First Workshop on Typology for Polyglot NLP, Florence, Aug. 1, 2019 (Co-located with ACL, July 28-Aug. 2, 2019).