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

Harald Hammarström, One-Soon Her, Marc Allassonnière-Tang

Term 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

Abstract

Starting from a large collection of digitized raw-text descriptions of languages of the world, we address the problem of extracting information of interest to linguists from these. We describe a general technique to extract properties of the described languages associated with a specific term. The technique is simple to implement, simple to explain, requires no training data or annotation, and requires no manual tuning of thresholds. The results are evaluated on a large gold standard database on classifiers with accuracy results that match or supersede human inter-coder agreement on similar tasks. Although accuracy is competitive, the method may still be enhanced by a more rigorous probabilistic background theory and usage of extant NLP tools for morphological variants, collocations and vector-space semantics.

1 Introduction

The present paper addresses extraction of information about languages of the world from digitized full-text grammatical descriptions. For example, the below reference describes a language called Kagulu, whose grammatical properties are of interest for various linguistic predicaments.

Petzell, Malin. (2008) *The Kagulu language of Tanzania: grammar, text and vocabulary* (East African languages and dialects 19). Köln: Rüdiger Köppe Verlag. 234pp.

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 are word-based and some combine words 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 term spotting, but in a way that obviates the need for either manual tuning of thresholds or supervised training data. However, this approach is limited to the features for which a (small set of) specific terms frequently signal the presence thereof, e.g., `classifier`, `suffix(es)`, `preposition(s)`, `rounded vowel(s)` or `inverse`. Term spottting 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-term-signalled features unless there is a specific formula such as SOV or N-Adj gaining sufficient popularity in grammatical descriptions. Term-signalled features are, of course, far simpler to extract, but not completely trivial, and hence the focus the present study.

The general-form premises to the problem addressed here are as follows. There is a set D of raw-text descriptions of entities from a set S, such that each $d \in D$ mainly describes exactly one $s \in S$. If a term k describing a property of objects in S occurs in a document d to a significant degree, the object s described in d actually has the property signalled by k. These premises apply to other domains and texts, e.g., ethnographic descriptions, than the linguistic descriptions in the present study. Judging from the surveys of Nasar et al. (2018) and Firoozeh et al. (2020), the premise that each

$d \in D$ mainly describes exactly one $s \in S$ is not dominant across scientific domains. Consequently most work has focussed on the broader tasks of extracting key-insights and salient keywords from scientific documents. We are not aware of any work in other domains on the specific task addressed in this paper.

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, (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), and (3) texts under publisher copyright where quotations of short extracts are legal. For each document, we know the language it is written in (the meta-language, usually English, French, German, Spanish, Russian or Mandarin Chinese, see Table 1), 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 feature, phonological description, grammar sketch, full grammar etc). The collection can be enumerated using the bibliographical- and metadata 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, contemporary 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 meta-language words has good prospects in spite of the noise, while extraction of accurately spelled vernacular data is not possible at present.

Meta-language		**# lgs**	**# documents**
English	eng	3497	7284
French	fra	826	1323
German	deu	620	813
Spanish	spa	394	808
Russian	rus	288	498
Chinese	cmn	180	234
Portuguese	por	141	274
Indonesian	ind	130	210
Dutch	nld	113	171
Italian	ita	92	141
...	...	...	...

Table 1: Meta-languages of the grammatical descriptions in the present collection.

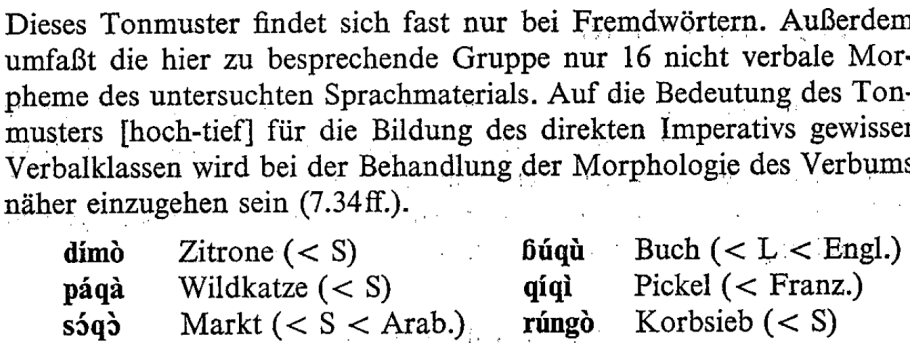
3 Model

At first blush, the problem might seem trivial: simply look for the existence of the term and/or its relative frequency in a document, and infer the feature associated with the term. Unfortunately, to simply look for the existence of a term is too naive. In many grammars, terms 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 is always the possibility that a term occurs in an example sentence, the title of a reference or the like. However, such "spurious" occurrences will not likely be frequent, at least not as frequent

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

> "It should be noted that unlike Proto-Uto-Aztecan (Langacker 1977:92-3) Pipil has no productive postpositions. However, it has reflexes of former postpositions both in the relational nouns (cf. 3.5.2) and in certain of the locative suffixes (cf. 3.1.3)" (Campbell, 1985, 61).

[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.

as a term 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 descriptive terms: Terms that describe the language in question.

Noise terms: Descriptive terms that do not accurately describe the language in question (i.e., through remarks on other languages 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 terms. Of the terms that genuinely describe a language, we would expect their frequency distribution in a grammar to mirror their functional load (Meyerstein, 1970), i.e., their relative importance, in the language being described. Thus we assume each language has a theoretical distribution $L(t)$ of terms t which is our object of interest. However, as noted, grammars typically also contain "noise" terms which distort the reflection of $L(t)$. A simple model for the frequency distribution of the terms of a grammar $G(t)$ is that it is composed merely of a sample of the "true" underlying descriptive terms $L(t)$ 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 there are only a handful of nouns with dual forms, but higher if there are dual pronouns, dual agreement, special dual case forms and so on. For most languages, we expect the functional load of verbs to be rather high. 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:

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

If we had infinitely many independent grammars accurately describing a language (and nothing else), their combined distribution would converge to $L(t)$ in the limit. Without the luxury of so many representative grammars, we can still attempt the simpler task of estimating the purity levels α_i of each grammar. That is, given actual distributions $G_1(t), \ldots, G_n(t)$ how can we make a heuristic estimate of α_i? 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 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, $G_j(dual) = 0.01$ and $G_k(dual) = 0.00$, this term barely occurs. The term *dual* would then have poor generality $g^i_L(dual) = \frac{\frac{1}{2}\cdot(0.00+0.01)}{0.1} = 0.05$. Some real examples of the generality of a few terms are found in Cojocaru (2004)'s grammar of Romanian given five other Romanian grammars are shown in Table 2. Terms like `triphthongs`, `gender` and `stress` have a role in describing the language and consequently show a generality close to 1.0, while "noise" terms like `cojocaru` and `ghe` are less common as items of description of the Romanian language.

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:

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

To remove outliers and speed up the calculation by removing hapax terms, in the experiments below, we measure all frequencies by logarithm.

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 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 3 has a list of grammars/grammar sketches of Romanian. Each grammar has a corresponding α_i purity level as described above, the total number of tokens, and the frequency threshold $\bar{t}$ induced by α and the token distribution. The last three columns concern the terms `masculine`, `feminine` and `neuter` respectively. The cells contain the frequency of the corresponding term, as well as the fraction of pages on which it occurs. The fraction of page occurrences is, of course, similar to, and highly correlated with the fraction of tokens but is often easier to interpret intuitively. We show it here for reference, although it is not advantageous to make use of in any of the above calculations. 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 term in question is probably genuinely describing 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.

t	cojocaru	triphthongs	gender	stress	ghe	...
Cojocaru 2004	0.00002	0.00004	0.00052	0.00025	0.00006	...
Agard 1958	0.00000	0.00002	0.00012	0.00078	0.00000	...
Gönczöl-Davies 2008	0.00002	0.00015	0.00046	0.00013	0.00002	...
Mallinson 1986	0.00000	0.00000	0.00103	0.00036	0.00000	...
Mallinson 1988	0.00000	0.00000	0.00055	0.00036	0.00000	...
Murrell and Ştefănescu Drăgăneşti 1970	0.00000	0.00004	0.00042	0.00027	0.00000	...
$g_{ron}^{\text{Cojocaru 2004}}(t)$	0.18	1.20	0.99	1.51	0.07	

Table 2: Some example terms from Cojocaru (2004) and their generality $g_{ron}^{\text{Cojocaru 2004}}(t)$ given five other Romanian grammars.

Romanian [ron]

Grammar	α_i	# tokens	$\bar{t}$	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				True	True	True

Table 3: Example grammars of Romanian and the frequencies of the terms masculine, feminine and neuter.

4 Evaluation

Thanks to a large manually elaborated database of languages with classifiers[3] (Her et al., 2021) 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 (as measured by number of tokens).

Gold Standard	Term-Spotting	# languages
False	False	2 357
True	True	512
True	False	317
False	True	34
		3 220

Table 4: Evaluation of term-spotting against a Gold Standard database of classifier languages.

A comparison between the Gold Standard database and the extracted data is shown in Table 4. The overall accuracy is 89.1%, to be compared with human inter-coder agreement on similar tasks, i.e., 85.9% or lower (as per Donohue 2006 and Plank 2009, 67-68). Not surprisingly, the method has better precision ($\frac{512}{512+34} \approx 0.94$) than it has recall ($\frac{512}{512+317} \approx 0.62$). The majority of errors are languages with classifiers which are not recognized as such by the term-spotting technique. Simple inspection reveals that in the majority of these cases, a different term, e.g., "enumerative" is used in place of the term in question. There are also errors where the automatic technique infers a slightly too high threshold for languages which have grammars from a large temporal range. The fact that descriptive tradition changes over time may be reason to refine the procedure for calculating reciprocated proportions.

We may add a few remarks on some obvious refinements. Excluding negative polarity mentions, by which we mean mentions where `no|not|absent|absence|absense|lack|neither|nor|cannot` occurs in the same sentence as the sought-after term, make no significant change to the overall accuracy. Using the temporally latest description only (instead of a majority vote) to assess the status for a language with several grammars, also made no significant change to the overall accuracy (in fact, it decreased by 2 percentage points). Furthermore, using the most extensive description only, i.e., the longest grammar or longest grammar sketch if there are no full grammars, had a negative impact on overall accuracy (down by 8 percentage points). These results seem to speak in favour of making use of multiple witnesses for each language if they are available, even if they are of different lengths and ages. If these impressions generalize, length and age differences between grammars — which are real — need to be addressed in a more sophisticated manner than simply excluding the old and short.

The above evaluation is relevant for the case when there is a specific term (or an enumerable set thereof) associated with the desired feature. It then shows what accuracy one may expect without supplying a threshold or any other information than the keyword itself. Choosing the right term(s) for a given linguistic feature requires knowledge of the feature and the way is it often (not) manifested in the literature (cf. Kilarski 2013 on classifiers versus other kinds of nominal classification).

5 Conclusion

We have described a novel approach to the extraction of linguistic information from descriptive grammars. The method requires only a term of interest, 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 terms. When this is the case, a broad evaluation shows that the results match or exceed the far more time-consuming manual curation by humans. Future work includes automated handling of collocations and morphological variants, vector-space lexical semantics, automated multi-lingual extraction and establishing the method on more rigorous probabilistic theory.

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 Wallen-

[3] See Aikhenvald (2000) and references therein for issues surrounding the definition of this feature.

bergs 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. The second author O.-H. Her gratefully acknowledges the financial support of Taiwan's MOST Research Grants 106-2410-H-029-077-MY3 and 108-2410-H-029-062-MY3. The last author is thankful for the support of grants from the Université de Lyon (ANR-10-LABX-0081, NSCO ED 476), the IDEXLYON Fellowship (2018-2021, 16-IDEX-0005), and the French National Research Agency (ANR-11-IDEX-0007).

References

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

Aikhenvald, Alexandra Y. 2000. *Classifiers: A Typology of Noun Categorization Devices*. Oxford Studies in Typology and Linguistic Theory. Oxford: Oxford University Press, Oxford.

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

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

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

Dryer, Matthew S. 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.).

Firoozeh, Nazanin, Adeline Nazarenko, Fabrice Alizon, and Béatrice Daille. 2020. Keyword extraction: Issues and methods. *Natural Language Engineering*, 26:259–291.

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

Hammarström, Harald. 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).

Hammarström, Harald, 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.

Hammarström, Harald, 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.

Her, One-Soon, Harald Hammarström, and Marc Allassonnière-Tang. 2021. Introducing WACL: The world atlas of classifier languages. *Submitted*, page 15pp.

Kilarski, Marcin. 2013. *Nominal classification: A history of its study from the classical period to the present*. Amsterdam: John Benjamins.

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

Macklin-Cordes, Jayden L., 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.

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

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

Meyerstein, R. S. 1970. *Functional Load: Descriptive Limitations Alternatives of Assessment and Extensions of Application*. The Hague: Mouton.

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

Nasar, Zara, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2018. Information extraction from scientific articles: a survey. *Scientometrics*, 117:1931–1990.

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

Virk, Shafqat Mumtaz, 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.

Virk, Shafqat Mumtaz, 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.

Virk, Shafqat Mumtaz, 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)*, page 1247–1256. Varna, Bulagaria: NCOMA Ltd.

Wichmann, Søren 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).