Automatically building a database of gender/noun class/classifiers from digitized grammatical descriptions

Harald Hammarström
Uppsala University
One-Soon Her
Tunghai University & National Chengchi University
Marc Tang
University Lumière Lyon 2
Olof Lundgren
Lund University
Hilda Appelgren
Lund University
William Zetterberg
Lund University

Background

- We want to investigate the heritability and diffusability of various nominal classification systems
  - (M/F/N) Gender
  - Noun class
  - Classifiers
- For this, a **dense** database is required
- Available (Virk et al. 2020):
  - **13 002** digitized OCRed grammatical descriptions describing
  - **4 535** languages of the world

Manually Curating Databases: Time/Cost

- A single subdomain (equivalent to, say, 20 features) covering 200-300 languages is typically the size of a PhD thesis
  - 297 lgs: Matti Miestamo (2003) *Clausal negation: A typological study*. University of Helsinki PhD Thesis.
  - 172 lgs: Veselinova, Ljuba. (2003) *Suppletion in Verb Paradigms: bits and pieces of a puzzle*. Stockholm University PhD Thesis.
  - 100 lgs: Di Garbo, Francesca. (2014) *Gender and its interaction with number and evaluative morphology: An intra- and intergenealogical typological survey of Africa*. Stockholm University doctoral dissertation.
  - ...
- With a fixed questionnaire of 200 features student assistants can collect data from reading grammars at a rate no faster than 20 datapoints per hour. With 13 EUR per hour, one datapoint costs 1.53 EUR

Challenge

- Can we instead machine-read the same grammars with accuracy comparable to human collection?
- Is there a combined human-machine approach that saves time/money?
- Little previous work on machines reading grammars (Hammarström 2013, Macklin-Cordes et al. 2017, Virk et al. 2017, 2019, Wichmann and Rama 2019)
- It is of value that the machine-reading of grammar can explain its results, i.e., no black box neural network

Example Descriptive Grammars in Database

- Very extensive grammar in English

  Campbell, Lyle. (1985) *The Pipil Language of El Salvador* (Mouton Grammar Library 1). Berlin: Mouton de Gruyter. xiv + 957pp.

- Grammar in German

  Vorbichler, Anton. (1971) *Die Sprache der Mamvu* (Afrikanistische Forschungen V). Glückstadt: J. J. Augustin. 356pp.

- Not so large grammar in Mandarin Chinese

  Yu, Cuirong 喻翠容. (1980) *Buyiyu jianzhi* 布依语简志. Beijing: Minzu Chubanshe. 113pp.

OCRed Grammar Collection

Spans 4 535 (target-)languages written in 76 (meta-)languages

Meta-language		# lgs	# Doc:s	# Tokens
English	eng	3 640	7 856	499 405 292
French	fra	864	1 397	95 730 890
Spanish	spa	405	877	49 291 641
German	deu	643	866	59 425 731
Russian	rus	311	555	38 763 081
Portuguese	por	143	287	14 483 473
Chinese	cmn	197	278	25 383 169
Indonesian	ind	130	211	5 573 867
Dutch	nld	113	177	10 665 158
Italian	ita	92	146	9 093 823
Japanese	jpn	33	38	1 646 876
...	...	...	...	...

English accounts for a larger share than all the other ones together!

OCR Quality Example (Though Quality Varies)

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)

$\Downarrow$

```
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;,
```

Keyword Spotting

Simplest possible approach is to look for keywords that signal the presence of the features in question, e.g., **preposition(s)**, **dual**, **tone(s|me)**, *...*

- Does not work when the feature is expressed in a myriad of different ways across grammars, e.g., whether the verb agrees with the agent in person
- Simple but not completely trivial because of spurious occurrences:
  - Explicit absense: "there is no X"
  - Disparate target: "another relevant language/temporal stage has X"
  - Sample occurrence: X occurs in an example, a reference title etc.
  - ...
- Genuine occurrences should be more frequent than spurious occurrences, but **how frequent is frequent enough**?

Terms in a Grammatical Description

Genuine keywords: Terms that describe the language in question

Noise keywords: Descriptive terms that do not accurately describe the language in question

$\Updownarrow$ rarely overlap

Meta-language words: Words in the meta-language, e.g., *the*, *a*, *run*, 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, e.g., morphemes of the language, place names in the language area, etc.

Terms in a Grammatical Description: Model

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

- $G(t)$: Frequency distribution of the keywords of a descriptive grammar composed 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

Estimating Noise Via Multiple Grammars

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

- If we have many grammars for the **same** language we can estimate the noise levels α_i

$$\alpha_i = \frac{\sum_t g_L^i(t)}{\sum_t G_i(t)}$$

- where $g_L^i(t)$ is the *generality* of the term t

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

Estimating Noise: Example

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 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 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

- So terms like `cojocaru`, `ghe`, … have poor generality vs `triphthongs`, `gender`, `stress`, … have better generality
- $\alpha_i = \frac{\sum_t g_L^i(t)}{\sum_t G_i(t)}$ is the average generality of all the terms of a grammar

How frequent is frequent enough?

- Does the frequency of a term in a grammar exceed its noise level (1-α)?
- Removing the (1-α_i) of least frequent tokens effectively generates a threshold $\bar{t}$
- Example: Does Romanian [ron] have m/f/n gender?

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				**TRUE**	**TRUE**	**TRUE**

Database of gender/noun class/classifiers

- Search over 7000+ grammars written in English spanning
- spanning 3000+ languages
- For languages with only one grammar, the threshold was set to an average threshold for similar-size grammars

Feature		Search Regexp
Gender		`\W[Gg]ender`
	M	`[Mm]asculine|[Mm]asc\W`
	F	`[Ff]eminine|[Ff]em\W`
	N	`[Nn]euter|[Nn]eut\W`
Classifiers		`[Cc]lassifier`
Noun class		`[Nn]oun class[^i]|[Nn]ominal class[^i]|[Nn]ominal concord`

Example Output

file:///C:/Python38/gender2020oct.html

Chimakum [xch]

Source	bibtype	t	# tokens	Classifiers	Gender	Noun class
Boas 1892	S	1	2716	0	5	0
Majority				False	True	False

Boas, Franz. (1892) Notes on the Chemakum Language. *American Anthropologist* 5(1). 37-44. [boas_chemakum1892v2.pdf boas_chemakum1892.pdf boas_chemakum1892_o.pdf]

Show hits

- Classifiers
- Gender

  -It seems that nouns have two **gender**s, masculine and feminine, which have separate articles

  The plural article is the same for both **gender**s : ho tsitsqa'll'e, my cousins

  -It appears from the examples given above that the noun has two **gender**s

  It is of interest to note that pronominal **gender**, by means of which male and female are distinguished; is found in all Salishan dialects spoken west of the Cascade range and on the coast of B

  male and female are distinguished; is found in all Salishan dialects spoken west of the Cascade range and on the coast of British Columbia, while real **gender** occurs in all dialects of the Chinook

- Noun class

Chilcotin [clc]

Source	bibtype	t	# tokens	Classifiers	Gender	Noun class
Cook 2013	G	57	195161	236	16	1
Majority				True	False	False

Cook, Eung-Do. (2013) *A Tsilhqút'in Grammar* (First Nations Languages Series). Vancouver: UBC Press. [cook_tsilhqutin2013_o.pdf cook_tsilhqutin2013.pdf]

Show hits

Chiga [cgg]

Source	bibtype	t	# tokens	Classifiers	Gender	Noun class
Taylor 1985	G	11	86775	0	7	14
Majority				False	False	True

Evaluation: Classifiers

- Grammars in English for 3 220 languages keyword-spotted for `classifier(s)`
- Evaluated against Gold Standard by Marc Tang and One-Soon Her (Her et al. 2021)

Gold Standard	Keyword-Spotting	# lgs	
False	False	2 357	73.2%
True	True	512	15.9%
True	False	317	9.8%
False	True	34	1.1%
		3 220	

- Overall accuracy is **89.1%**

Manually Curated Databases: Accuracy

- On the WALS database
  - Wälchli 2005 checked every Latvian feature and found $102/112 \approx \mathbf{91}.\mathbf{1}\%$ correct
  - Donohue 2006 checked every Tukang Besi feature and found $122/142 \approx \mathbf{85}.\mathbf{9}\%$ correct
  - Plank 2009:67-68 checked every German feature and found

… *for over a quarter, perhaps almost a third of the features mapped, the values assigned are erroneous, arbitrary, or uncertain in view of analytic alternatives, or would have been different if one or the other variety of the language summarily located at 52°N 10°E had been chosen for coding*

  - Hammarström 2013 checked every language for one feature (basic word order in the transitive clause, WALS 81A) and found $1028/1228 \approx \mathbf{83}.\mathbf{7}\%$ correct
- On the Grambank database (checking 3x20 languages in 2016)
  - Average % two coders use the same sign on the same source document: $\frac{2203}{3116} = \mathbf{70.7\%}$
  - Average % two agree when both are non-? on the same source document: $\frac{1514}{1682} = \mathbf{90.0\%}$

Some Tweaks Evaluated on Classifiers

- The comparison revealed a certain amount of source errors (misattached files etc) and OCR errors (file looks OCRed on the surface, but is actually garbage)
  - Curiously, fixing them yielded a lower accuracy (**87%**) because a number of cases of possessive classifiers then rose above threshold
- Negative polarity mentions (= presence of `no|not|absent|absence|absense|lack|neither|nor|cannot` in the same sentence as the keyword) discounted
  - No discernable impact on accuracy
- Negative polarity mentions (= presence of `no|not|absent|absence|absense|lack|neither|nor|cannot` in the same sentence as the keyword) discounted
  - No discernable impact on accuracy
- Using the temporally latest description only
  - Accuracy down from **87%** to **86%**
- Using the `Most Extensive Grammar` only (= highest category, longest)
  - Accuracy down from **87%** to **79%**

Evaluation: Gender/Noun Class

- Manually checked in order of priority by Olof Lundgren, Hilda Appelgren and William Zetterberg
- Average pace: 22 languages per day and person

Prio	# lgs	Status	Selection
1	365	Checked by OL	Keyword-signalled as NC
2	928	Checked by OL + HA + WZ	Keyword-signalled as not NC + Some language in the family known to have NC
3	971	Checked by HA + WZ	Keyword-signalled as not NC + No language in the family has earlier been coded for NC
	2 264		
4	*813*	*Not Checked*	*Keyword-signalled as not NC + not Gender + No language in the family known to have NC*
	3 077		

Evaluation: Gender

Gold Standard	Keyword-Spotting	# lgs	
False	False	1 346	59.6%
True	True	500	22.2%
True	False	132	5.8%
False	True	279	12.3%
		2 257	

- Overall accuracy is **81.8%**

Evaluation: Noun Class

Gold Standard	Keyword-Spotting	# lgs	
False	False	1 829	81.2%
True	True	186	8.3%
True	False	129	5.7%
False	True	109	4.8%
		2 257	

- Overall accuracy is **89.4%**

Errors Analysis: Noun Class

Error Type	# lgs	Description
context	108	keyword found, but in another context, either referring to, e.g., "inflectional classes" or referencing other languages
high threshold	52	keywords found but not enough times or not consistently across different sources for the same language.
negative mention	13	keyword used in a negative context, e.g., "has no noun classes".
wrong keyword	44	Another keyword was used, e.g., "male/female" instead of "masculine/feminine".
gender	40	The language has "noun classes" which was detected, but gender was part of the NC system, so NC = FALSE according to our definition.
wrong hit	2	The hit was not the desired keyword, e.g., an in-language word "fem".
not english	3	Source is not in English

Summary & Conclusion

Feature	Accuracy	
Classifier	**87.1%**	Machine
M/F/N Gender	**81.8%**	Machine
Noun class	**89.4%**	Machine
Basic Constituent Order WALS	**83.7%**	Human
Overall WALS	**85.9%-91.1%**	Human
Overall Grambank	**90.0%**	Human

What is the actual time/accuracy trade-off for a combined machine-human checking approach?

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

Cojocaru, D. (2004). *Romanian Grammar*. Durham: SEELRC.

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

Gönczöl-Davies, R. (2008). *Romanian: an essential grammar*. New York: Routledge, New York.

Hammarström, H. (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).

Her, O.-S., Hammarström, H., and Allassonnière-Tang, M. (2021). Introducing wacl: The world atlas of classifier languages. *Submitted*, page 15pp.

Macklin-Cordes, J. L., Blackbourne, N. L., Bott, T. J., Cook, J., Ellison, T. M., Hollis, J., Kirlew, E. E., Richards, G. C., Zhao, S., and Round, E. R. (2017). Robots who read grammars. Poster presented at CoEDL Fest 2017, Alexandra Park Conference Centre, Alexandra Headlands, QLD.

Mallinson, G. (1986). *Rumanian*. Croom Helm Descriptive Grammars. London: Croom Helm.

Mallinson, G. (1988). Rumanian. In Harris, M. and Vincent, N., editors, *The Romance Languages*, pages 391–419. London: Croom Helm.

Murrell, M. and Ştefănescu Drăgăneşti, V. (1970). *Romanian*. Teach Yourself Books. London: English Universities Press.

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

Virk, S. M., Borin, L., Saxena, A., and Hammarström, H. (2017). Automatic extraction of typological linguistic features from descriptive grammars. In Ekštein, K. and Matoušek, V., 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, S. M., Hammarström, H., Forsberg, M., and Wichmann, S. (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, S. M., Muhammad, A. S., Borin, L., Aslam, M. I., Iqbal, S., and Khurram, N. (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*.

Wichmann, S. and Rama, T. (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).