

APC detector

BERT-based token classification for nominal person in German

Georg F.K. Höhn

4 July 2025

Ironhack

Project overview

Adnominal pronoun constructions (APCs)

(1) a. **we** linguists

b. **wir** Linguisten
we linguists

German

Challenges

- hard to detect automatically
- limited empirical data on distribution
- available tagged dataset not publishable due to copyright

Aim

- use tagged dataset to train BERT-transformer for high-accuracy automatic detection (for German)

Outline

Background

Practical challenges

Methodology

Model evaluation

Quick demo

Conclusion and outlook

Background

Nominal person

- expressions marking whether a speech act participant (author, addressee) is part of the reference set of a nominal expression
- APCs are one type, other examples:

(2) a. clitic person marking

yima-nēm

Alamblak

person-1PL

“we people” (Bruce 1984:96, (158))

- b. aka [malav e] roa-ru kiu-la-m. Lavukaleve
then people 1PL.EXCL one.SG.M-none die-NEG-SG.M

“And we, the people [lit: the people we] didn’t die. [i.e. None of us people died.]” (Terrill 2003:171)

Crosslinguistic variation

- presence of definite article, possibly related to “unagreement” (ability to drop the pronoun)

- (3) a. [(**Nosotras**) **las** **mujeres**] denuncia-mos Castilian
we.F DET.PL.F women denounced-1PL
- las injusticias.
DET.PL injustices
- “We women denounced the injustices.” (after Hurtado 1985:187, (1))

- singular contexts restricted in English, fine in German

- (4) [**Ich** **Vulkanier**] habe Dinge gesehen, die German
1SG.NOM Vulcan have things seen REL.ACC.PL
- [**du** **Mensch**] dir nicht vorstellen kannst.
2SG.NOM human 2SG.DAT NEG imagine can.2SG
- “*I Vulcan have seen things that you human cannot imagine.”

- no adnominal third person APCs in many (European) languages

(5) a. ***They linguists** are concerned with strange issues.

b. Kipu mae [**ria nikana Japani**] de Hoava
NEG come 3PL man Japanese PURP

[**yami nikana hupa**] mae ŋani=i [**yami**].
1EXCL.PL man be.black come kill=ACC 1EXCL.PL

“The Japanese men did not come to come and kill us black men.” (Palmer 2017:426)

- a possible correlation between third person adnominal pronouns and definite articles? (Höhn 2020)

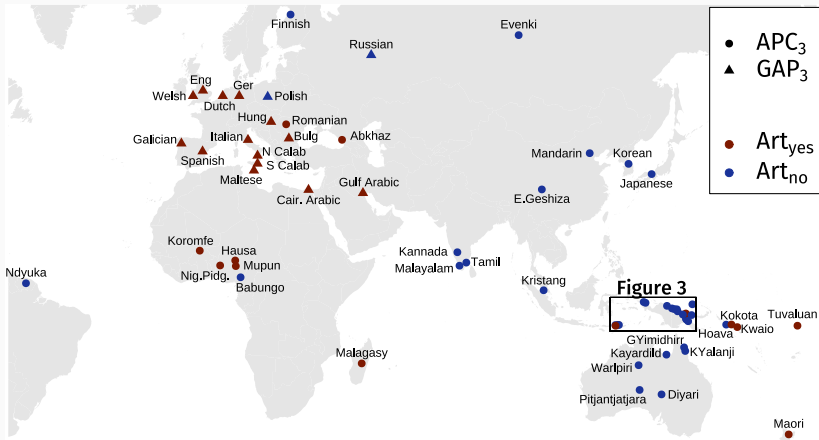


Figure 1: Availability of third APCs and definite articles

Practical challenges

- few corpus investigations on nominal person/APCs (Keizer 2016 uses data from BNC), although lots of potential:
 - language-internal distribution (genre, information structure)
 - relationship to definiteness
 - relationship between different types of nominal person (APCs and unagreement)
 - crosslinguistic differences in frequency of use
- identifying APCs in POS-tagged corpora is time/work-intensive
 - many false positives
 - capturing intervening modifiers requires more complex queries
→ even more false hits

Modifiers

- (6) a. **ihr** armen kleinen **Deutschen**
you.PL poor little Germans
- b. **du** von großer Sorge geplagter **Anführer** der
you.SG by great sorrow tormented leader DET.GEN.PL
Achaier
Achaeans
“you leader of the Achaeans who are(is?) tormented by great
sorrow”

Some potential false positives

- (7) a. I told [you] [**linguists** are great].
- b. Gestern haben [**wir**] [**Linguisten** gesehen].
yesterday have.1PL we linguists seen
“Yesterday we saw linguists.”

Methodology

Manually tagged datasets

- manual pattern searches in POS-tagged corpora
 - English: British National Corpus (BNC)
96,986,707 tokens
 - German: Digitales Wörterbuch der deutschen Sprache (DWDS),
Kernkorpus 1900–1999
121,494,429 tokens
- hits collated into tables with metainformation (search pattern, year, bibliographic reference etc.) + context
- annotation for APC + further linguistic properties
 - many thanks to Andrea Schröter, Maya Galvez Wimmelmann and Carolin Kuna for their work!
- unfortunately strict licensing conditions on the corpora preclude publication of the annotated datasets

Training dataset

45,539 hits:

- no APC: 39,239
- APC: 6,188

Table 1: Precision for APC search by pronoun

	ich	du	wir	ihr	Sie
total hits	14,771	5,285	17,939	2,623	4,756
APCs identified	327	1,753	3,690	202	208
Precision (in %)	2.214	33.17	20.57	7.7	4.37

- (8) Und er wollt dir einen Groschen schenken, und er
and he wanted you.SG.DAT a twopence gift and he
spielte mit dir Räuber und Gendarm.
played with you robber and cop
“and he wanted to give you a twopence and he played cops and
coppers with you.” (ID 16413)

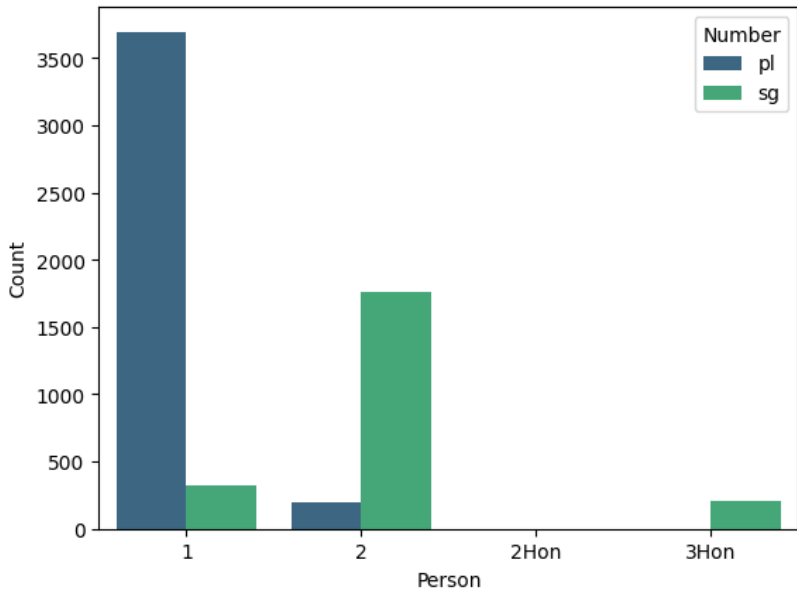


Figure 2: Absolute frequency of APCs by person/number

Relative frequency

(9) $n \text{ of APC}_{\text{PersNum}} * 100 / \text{number of pronouns}_{\text{PersNum}}$

- counts of pronouns contain only those that were not blocked from display for copyright reasons, since APC checking was necessarily restricted to the accessible examples

Table 2: Relative frequency of APCs by pronoun

	ich	du	wir	ihr	Sie
pronoun count	587,893	119,295	322,350	120,512	N/A
APCs identified	327	1,753	3,690	202	208
APC freq (in %)	0.056	1.47	1.447	0.168	N/A

Approach

- using huggingface's Trainer with pre-trained base model:
bert-base-german-cased

Token-level classification

- prominently used in named entity recognition
- tokenized items get one of three labels
 - B-APC: **beginning** of the construction
 - I-APC: **inside** the construction
 - O: **outside** the construction (majority of tokens)
- model learns to apply these labels to new input

(10) AMD Radeon is a brand name by Advanced Micro
B-BRAND I-BRAND O O O O O B-COMP I-COMP
Devices (AMD)
I-COMP I-COMP

Training data structure

- reduced version of annotated dataset
- retaining
 - ContextBefore, Hit, ContextAfter: strings
 - instance: marking whether row contains an APC (1) or not (0)
 - simplified externally in EDA (original annotation has several other levels)
 - APC: string of the concrete APC
- one row per (potential) APC instance → sentences with more than one (potential) APC occur in multiple rows

Data pre-/post-processing (APCData class)

- organises raw data into appropriate format for the trainer
- input: structured csv or raw text
 - raw text gets split into a table with triples of (previous sentence, focus sentence, following sentence)
- methods for `training=True` dataset
 - generate BIO-labels
 - merge BIO-labels in duplicate sentences to ensure full annotation, remove duplicates
 - create train-val-test split and BERT-tokenize to new dataset
- methods for `training=False` dataset
 - sentence tokenize input and split into overlapping triples
 - import predictions from inference and align with tokenization
 - de-tokenize texts and filter for APCs (and optionally pronouns)
- shared method: BERT-tokenization
 - chunking to avoid hitting maximum length of 512 tokens

Model evaluation

	train	validation				
epoch	loss	loss	accuracy	F1	precision	recall
1	0.0110	0.00996	0.99699	0.99699	0.99699	0.99699
2	0.0049	0.01134	0.99716	0.99716	0.99716	0.99716
3	0.0022	0.01420	0.99718	0.99718	0.99718	0.99718
		test				
		0.00955	0.99667	0.99667	0.99667	0.99667

Table 3: Evaluation metrics for training

- model size: about 415 MB
- almost 99.7% accuracy
- no detailed analysis of possible partial mis-/matches (e.g. Tjong Kim Sang & De Meulder 2003)

Quick demo

- choose input
 - direct text input
 - text file
- choose output
 - only detected APCs
 - detected APCs and personal pronouns
 - allows identification of false negatives and calculation of rel.freq.
- display results as DataFrame
 - works cumulatively
- save results as csv

Conclusion and outlook

- transformers offer a promising avenue for relatively reliable identification of APCs
- trade-off: large-ish model size at 415 MB

Notes

- beware of interactions of batched tokenisation and chunking of long datarows
 - batching of huggingface Dataset.map() does *not* work if rows are added during batch processing due to chunking

Further evaluation

- create systematic test set of difficult data types
- comparing up- and downwards
 - export test set as csv, compare performance of LLMs
 - build smaller models and compare performance

Some extensions

- extend APCData to allow import of csv files for testing as well
- train and include model for English data
- more options for data output (bracketed APCs, full dataset as plain text)
- inferencing larger dataset (including difficult cases)
 - check manually for gold-standard → further model training

Thanks for your attention and a pleasant
and productive 9 weeks!

Bruce, Les. 1984. *The Alamblak language of Papua New Guinea (East Sepik)*. Canberra: The Australian National University.

Höhn, Georg F. K. 2020. The third person gap in adnominal pronoun constructions. *Glossa* 5(1). 69. doi:10.5334/gjgl.1121.

Hurtado, Alfredo. 1985. The unagreement hypothesis. (Ed. by.) L. King & C. Maley. *Selected papers from the thirteenth linguistic symposium on Romance languages*. Amsterdam: John Benjamins. doi:10.1075/cilt.36.12hur.

Keizer, Evelien. 2016. We teachers, you fools: Pro+N(P) constructions in Functional Discourse Grammar. *Language Sciences* 53. 177–192. doi:10.1016/j.langsci.2015.05.006.

Palmer, Bill. 2017. Categorical flexibility as an artefact of the analysis. Pronouns, articles and the DP in Hoava and Standard Fijian. *Studies in Language* 41(2). 408–444. doi:10.1075/sl.41.2.05pal.

Rauh, Gisa. 2003. Warum wir Linguisten ‘euch Linguisten’, aber nicht ‘sie Linguisten’ akzeptieren können. Eine personendeiktische Erklärung. *Linguistische Berichte* 196. 390–424.

Terrill, Angela. 2003. *A grammar of Lavukaleve*. Berlin: Mouton de Gruyter. doi:10.1515/9783110923964.

Tjong Kim Sang, Erik F. & Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003*.
<https://aclanthology.org/W03-0419/>.