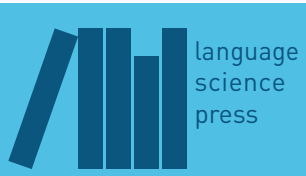
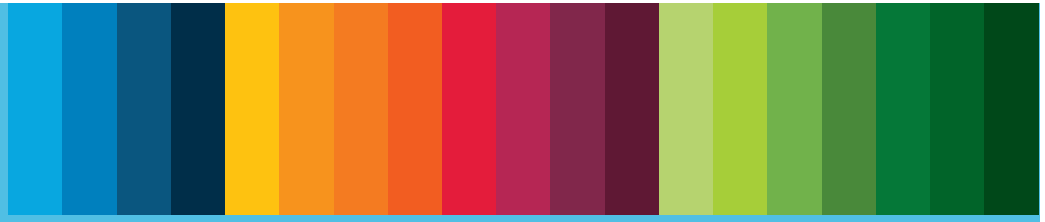


# Retrieving entities from publications in linguistics

Sebastian Nordhoff

2018-09-04, HIRMEOS Workshop, SUB Göttingen

Language Science Press

language science press

# Outline

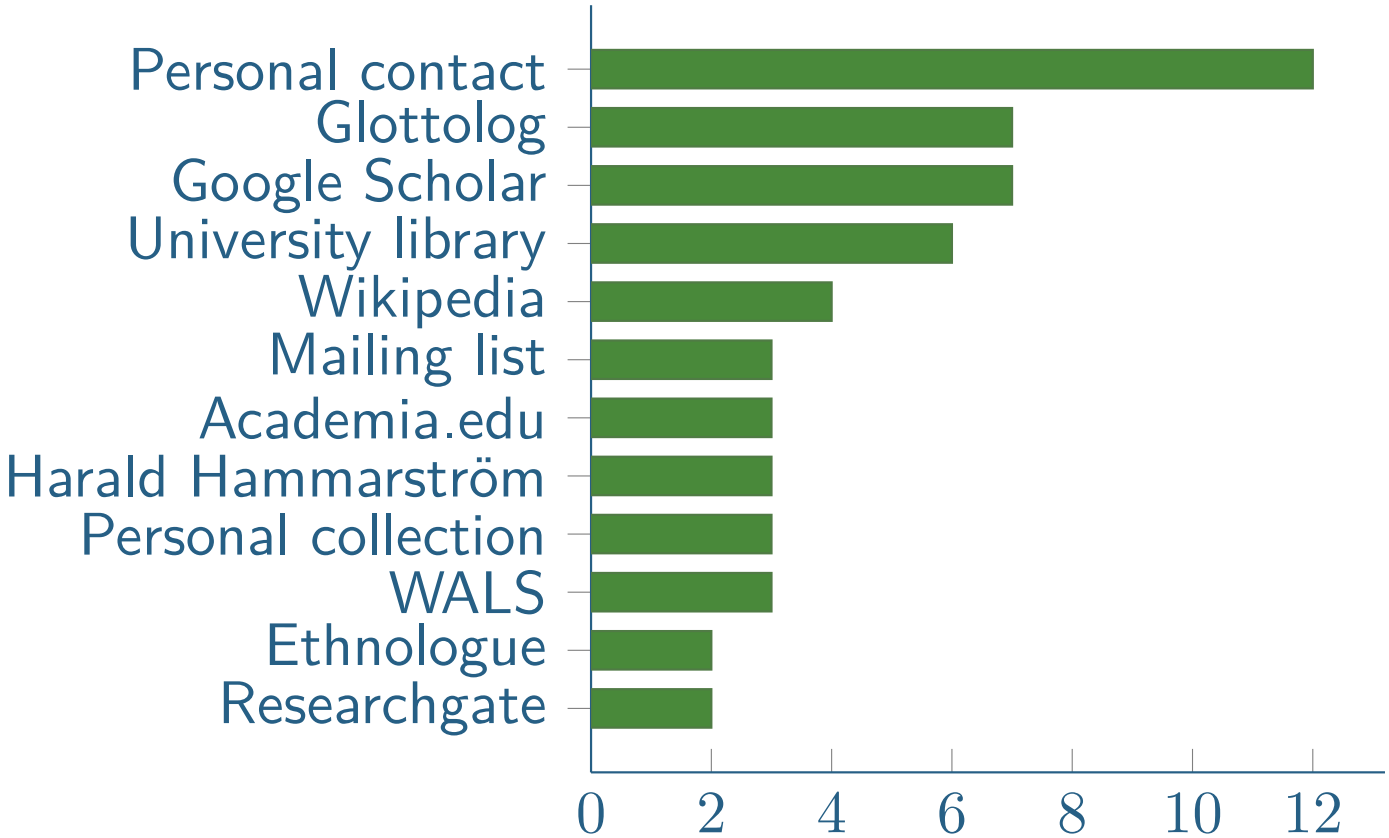
# Linguistics



- ca. 25,000 linguists worldwide
- both monographs and articles
- rather long publication cycles
- less output than for instance biology
  - possibility to keep track
- less sifting

# How do linguists search for literature?

⟩ For a domain you have little expertise of, how do you find relevant literature?




question asked on list *Linguistic Typology* on 2018-08-29, no predefined answers

n=18, multiple answers possible

# Language Science Press

## ⟩ start 2014; 75 books; 22 series

# Language Science Press

- Available formats for books
  - pdf
  - tex
  - bib
- Indexes in books
  - Language index
  - Subject index
  - Name index
  - Indexes are a discovery tool, similar to NERD.

# Bootstrapping index with sketchengine



⟩ A recent book on film subtitles and eyetracking had the following index candidates generated by sketchengine

image composition
eye tracking
speaking direction
typographic identity
fixation duration
audiovisual translation
aesthetic experience
title area
film material
speaker identification
film title
natural focus
text element
image track
information intake
graphical translation
title placement
bottom-centre area
typographic film
german image
split attention
gaze behaviour
reading speed
film identity
typographic film identity
tracking research
first fixation
additional language
narrative text
eye tracking research
visual attention
individual placement

# NERD and linguistics

# Goals

- higher level goals of text and data mining:
  - provide better tools for exploration:


  - automated reasoning:
    - gene $\longleftrightarrow$ protein
    - protein $\longleftrightarrow$ disease
    - gene $\longleftrightarrow$ protein $\longleftrightarrow$ disease

# Goals

- stated goals (Hirmeos):
  1. enhance discoverability
  2. aggregation (word clouds)
  3. generate collections
  4. highlighting

# Relevant knowledge bases

- The following knowledge bases can be seen as resources for disambiguation
  - **authority** (= Name Index)
    - GND
    - ORCID
  - **languoids** [languages, dialects, families] (= Language Index)
    - Glottolog
  - **concepts** (= Subject Index)
    - GOLD
    - concepticon

# Keywords

⟩ Several platforms have fields for “keywords” (OMP, Zenodo)

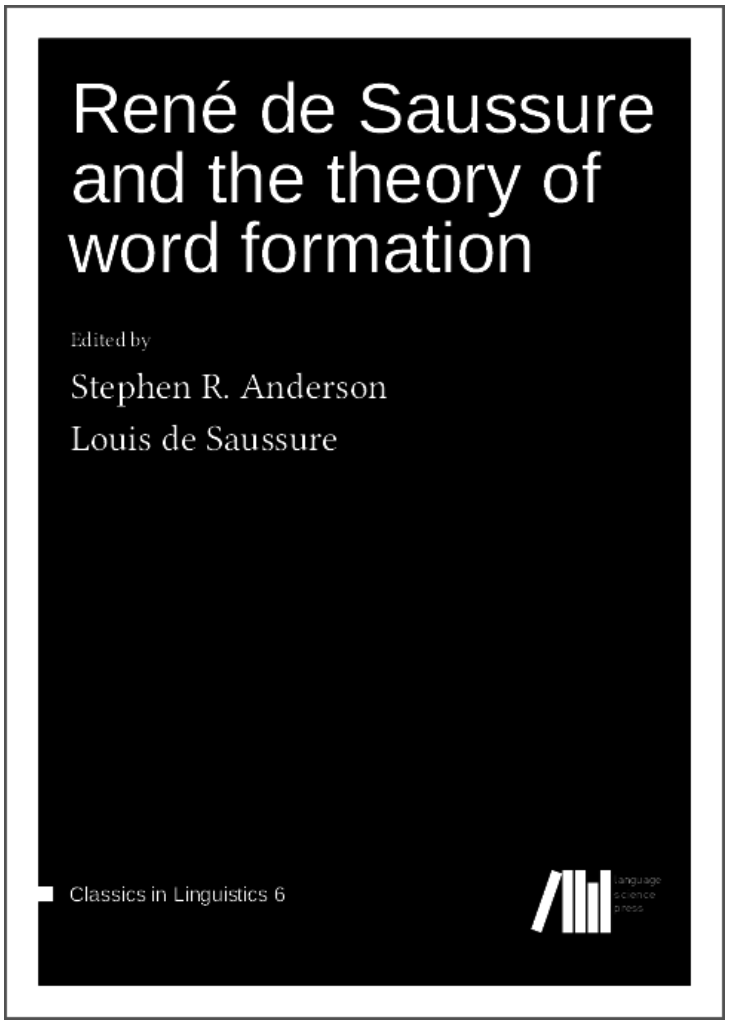
**Keywords**

⟩ But should I really enter strings there?

Alternative names

**hhbib_lgcode:**
B. 2009)
Meson de Guadalupe
Mixtec-Mixtepec
San Juan Mixtepec
San Juan Mixtepec-Oaxaca
**lexvo:**
Mixtepec Mixtec [en]
**multitree:**
Eastern Juxtlahuaca Mixtec
Mixtec, Mixtepec
Mixteco de San Juan Mixtepec
Northern Misteko
Northern Mixteco

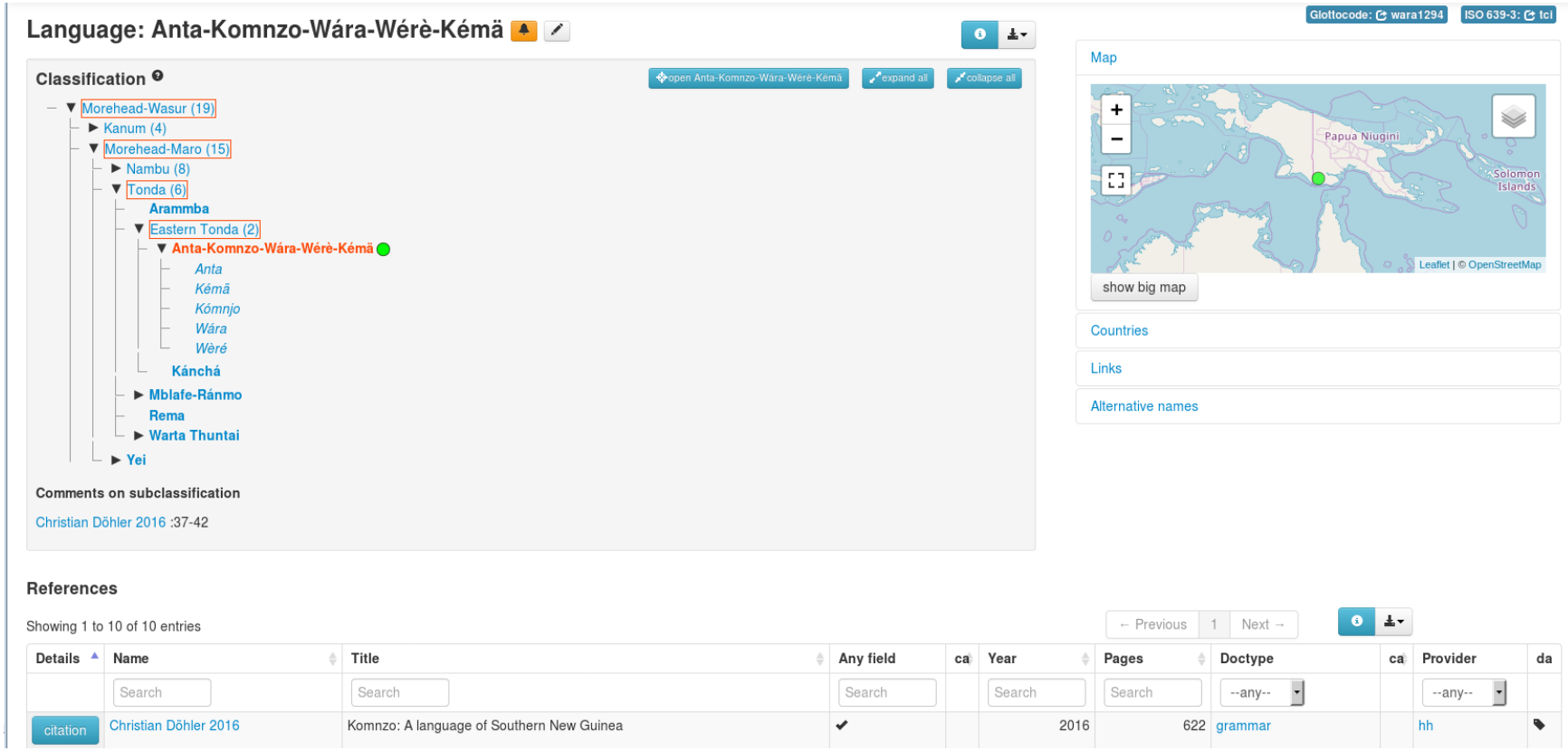
# Person names




- GND
- ORCID

# Language names: Glottolog

## ⟩ *A grammar of Komnzo*



# Linguistic concepts

- cross-linguistic categories don't exist

# Linguistic concepts

- cross-linguistic categories don't exist
- cross-linguistic categories don't exist

# Linguistic concepts

- cross-linguistic categories don't exist
- cross-linguistic categories don't exist
- cross-linguistic categories don't exist

# Linguistic concepts

- cross-linguistic categories don't exist
- cross-linguistic categories don't exist
- cross-linguistic categories don't exist
- something called "dative" in language X cannot be equated with something called "dative" in language Y
  - General Ontology for Linguistic Description (GOLD) tried and failed

# Inallative as "cross-linguistic category"

GOLD 2010 | issues | versions | xml | owl/rdf | gold community | help

top definition usage examples properties issues

## Inallative Case ( Concept )

**http://purl.org/linguistics/gold/InallativeCase**

Thing
|_ Abstract
|_ Linguistic Property
|_ Morphosyntactic Property
|_ Case Property
|_ Inallative Case

**Definition:**

InallativeCase expresses that something is moving toward the region that is inside the referent of the noun it marks. It has the meaning 'towards in(side)'. Kibrik says that Archi (aqc) possesses a nominal spatial form expressing InallativeCase, namely -aši [Kibrik 1998: 470].

**Usage Notes**

**Examples**

| Properties | Values | Definition |
|---|---|---|

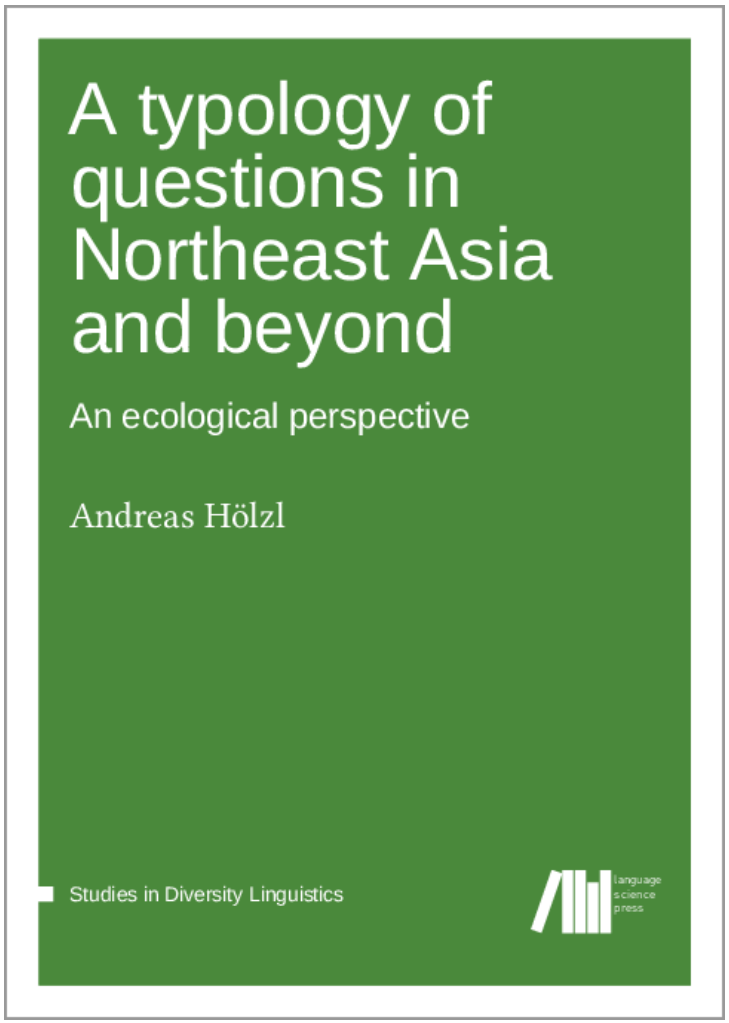
**User Submitted Issues**

# What have the Romans ever done for us?



- Is there an "inallative" in *Romani ite domu**m***?
- Is there an "inallative" in ***au*** *foyer*?
- Is there an "inallative" in ***t****huis*?
- Can you equate the usages in the three examples?
- take-home-message: it's complicated, and automated reasoning will not work.

# Testing NERD




- *A typology of questions in Northeast Asia and beyond*
- Book chosen as the most recent publication
- Variety of countries, languages, ethnic groups, concepts, etc.
- 546 pages
- NERD running on local machine

# Test section: Mongolic



Table 5.85: Spatial deictics in Mongolian according to Janhunen (2012b: 131), slightly reduced

| | PROX (hearer) | DIST | INT |
|---|---|---|---|
| LOC | naa-n' | tzaa-n' | xaa-(n') |
| LOC ABL | naa-n-aas | tzaa-n-aas | xaa-n-aas |
| LAT | naa-sh | tzaa-sh | xaa-sh |
| PROL | naa-g.oor | tzaa-g.oor | xaa-g.oor |

Table 5.86 shows five of the interrogatives that can be found in most modern Mongolic languages.

Table 5.86: Five Proto-Mongolic interrogatives and their modern representatives

| | *ken 'who' | *yaxun 'what' | *alin 'which' | *kejixe 'when' | *kaxana 'where' |
|---|---|---|---|---|---|
| Dagur | xeng | yoon | aly | xejee | xaan |
| Mongolian | xen | yuun | alyn | xedzee | xaan' |
| Buryat | xen | yüün | ali | xezee | xaana |
| Khamnigan Mongol | ken | yɛɛn | ali | kejie | kaana |
| Ordos | ken | yüün | ali | kejee | kaa |
| Written Oirat | ken | you/n | ali | kezee | **xamigha(a)** |
| Oirat | ken | yuu/n | äl ~ äl-k | keze | **xamaa** |
| Kalmyk | ken | yuun | aly(-k) | kezā | **xama**, **aly-d** |
| Shira Yughur | ken | **yima** | aali | kejee | xana |
| Santa | kien | yang | ali | giezhe | khala |
| Bonan | kang | yang | ane | kece(-) | hala |
| Kangjia | kɔ | jɔ ~ jaŋ | **ani(ye)** | gədʒe | χana |
| Huzhu Mongghul | ken | ya(a)n | ali | kijee | **an-ji(i)** |
| Minhe Mangghuer | kan | ya, yang | **a(yi)ge** | kejie | **ang(-ji)** |
| Moghol | ken ~ kiyan | **emah** ~ **imas** etc. | ? | keja | ? |

According to Janhunen (2003d: 20) the stem **ke-* originally had the meaning 'who' as well as 'what', which is an unlikely scenario from a cross-lingusitic point of view. As has been shown by Cysouw (2005), the only place worldwide where this pattern is not extremely rare or altogether absent is South America.

Proto-Mongolic had two resonances (submorphemes), one in **k~* that is still present in most Mongolic languages but changed to *x~* in Dagur, Buryat and Mongolian, and one in **y~* that has survived up to today. Similar changes from **k~* to > **x~* can be seen in Turkic

# Results

- NERD retrieved some pretty specialized concepts
  - Recall is good
- NERD also retrieved a lot of irrelevant concepts (“South America”) or lookalikes (business names, radio stations)
- NERD was rather aggressive and colored whole pages.
- the system seems to have understood that the book is about linguistics and often selects a linguistic concept. However, sometimes, the concept chosen is off the mark (Australia).
  - Precision is low.

# Results

- Installation procedure was OK
- Loading the book in the browser worked out of the box
- Loading the book in the browser takes several minutes

# Questions from a publisher

- in how far does NERD help the readers/authors?
  - Exploration/Discovery
    - currently, discoverability of content via series, e.g. *Contemporary African Linguistics*
    - linguists seem to prefer personal/social interaction to automated recommender systems
  - Automated reasoning
    - limited potential given the fuzzy nature of cross-linguistic concepts

# Goals (reprise)

⟩ stated goals (Hirmeos):

1. enhance discoverability: better than curated series?
2. aggregation (word clouds): better than index?
3. generate collections: better than curated series?
4. highlighting: is color a value in itself?