



# Retrieving entities from publications in linguistics

Sebastian Nordhoff

2018-09-04, HIRMEOS Workshop, SUB Göttingen

Language Science Press

---

Linguistics

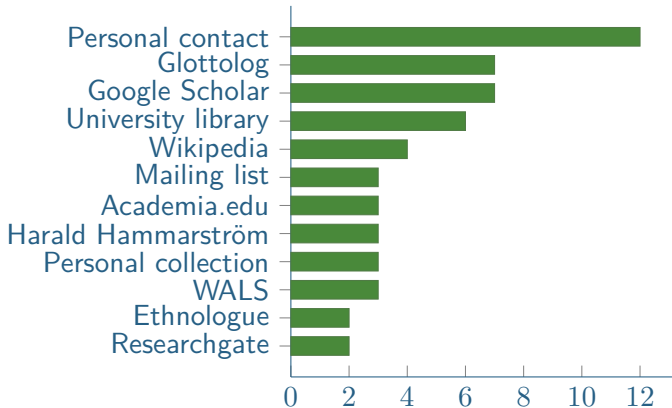
Language Science Press

NERD and linguistics

Testing NERD

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

> 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

## > start 2014; 75 books; 22 series

<p><b>Nature of language: Frames, forms, and functional transmission</b> N. J. Enfield (Author)</p>	<p><b>A typology of marked languages</b> Corina Finken (Author)</p>	<p><b>The Talking Heads experiment: Origins of words and meanings</b> Luc Steels (Author)</p>	<p><b>Grammaticalization in the Nordic: When phoneme changes in Scandinavian vernaculars</b> Olov Dahl (Author)</p>	<p><b>A grammar of Fula Sereer</b> Jo-Paar Wilbur (Author)</p>	<p><b>Prosodic detail in Neapolitan Italian</b> Francesca Gargano (Author)</p>
<p><b>Linguistic variation, identity, socialization and cognition</b> Karin E. Döpker (Author)</p>	<p><b>Language strategies for the study of</b> Jon Bay (Author)</p>	<p><b>A grammar of Tulu</b> Diana Schwan (Author)</p>	<p><b>Thoughts on grammaticalization</b> Dietmar Lehmann (Author)</p>	<p><b>New methods for self-organizing</b> Paul Fung (Author)</p>	<p><b>A grammar of Macedonian</b> Ljiljana Filipović (Author)</p>
<p><b>Roots of language</b> Derek Bickerton (Author)</p>	<p><b>The general base of linguistics: Grammatical judgments and linguistic methodology</b> Carsten Schwan (Author)</p>	<p><b>A grammar of Fula</b> Frank S. Egger (Author)</p>	<p><b>New directions in corpus-based translation studies</b> Claudia Finkenauer, Ingrid Isenhardt (Editors)</p>	<p><b>Grammatical theory: From traditional grammar to conceptualist approaches</b> Søren Madsen (Author)</p>	<p><b>Advances in the study of African languages and linguistics</b> Catherine Nkomo, Ruth James Gordon (Editors)</p>
<p><b>The evolution of grounded spatial language</b> Michael Torgerson (Author)</p>	<p><b>The evolution of language</b> René van der Lely (Author)</p>	<p><b>The future of Fula: Selected papers from the 10th Fula Conference</b> Mama Haidara, Aminata Diallo, Aminata Diallo (Editors)</p>	<p><b>Typical and atypical: Zur Modellierung von atypischen und atypischen Sprachverläufen</b> Thomas Lohmann (Author)</p>	<p><b>Adjective inflection</b> Michael Heller (Author)</p>	<p><b>A grammar of Tswana</b> Alicia Steinberg (Author)</p>
<p><b>Sprachliche Variation: Jiddisch in der deutschsprachigen Literatur (18.-20. Jahrhundert)</b> Leo Schaller (Author)</p>	<p><b>Forthcoming: Language technologies for computational grammar</b> George Bejken, Daniel G. Jones, Julia Saulis, Andrew Wilks (Editors)</p>	<p><b>Eye-tracking and Applied Linguistics</b> Miguel Ángel Sánchez, Sandra García (Editors)</p>	<p><b>Forthcoming: Grammatik und Grammatikalisierung: Die Rolle der Grammatik in der Sprachentwicklung</b> Günther Müller (Author)</p>	<p><b>Die Einführung in die Grammatik</b> Reinhold Köpcke (Author)</p>	<p><b>New perspectives on cohesion and coherence</b> Klaus R. Müller, Klaus R. Müller, Klaus R. Müller (Editors)</p>
<p><b>Translating N. J. Enfield's work and methodology</b> Alexandra M. (Author)</p>	<p><b>A grammar of Fula</b> Frank S. Egger (Author)</p>	<p><b>Forthcoming: African languages and grammaticalization</b> Aminata Diallo, Aminata Diallo (Editors)</p>	<p><b>Forthcoming: Grammatical modeling of translation and interpreting</b> Sandra Müller, Sandra Müller, Sandra Müller (Editors)</p>	<p><b>Forthcoming: Reflections on</b> Celia C. (Author)</p>	<p><b>Diversity in African languages: Selected papers from the 10th Annual Conference on African Linguistics</b> Doreen L. (Editors)</p>
<p><b>Annotation, exploitation and evaluation of parallel corpora (L33)</b> Sandra Müller, Sandra Müller, Sandra Müller (Editors)</p>	<p><b>Forthcoming: Beiträge zur deutschen Grammatik</b> Thomas Lohmann (Author)</p>	<p><b>The Language Dictionary and grammar class</b> Toni Schwan (Author)</p>	<p><b>A grammar of Fula</b> Frank S. Egger (Author)</p>	<p><b>Forthcoming: Unity and diversity in grammaticalization studies</b> Walter D. (Editors)</p>	<p><b>Text placement in Tschingel: How an innovation system accommodates its own technological environment</b> Thomas Lohmann (Author)</p>
<p><b>Forthcoming: Order and structure in syntax</b> Lorenz (Author)</p>	<p><b>Forthcoming: Order and structure in syntax</b> Lorenz (Author)</p>	<p><b>Forthcoming: A grammar of Fula</b> Frank S. Egger (Author)</p>	<p><b>Forthcoming: Further investigations in the study of Fula</b> Frank S. Egger (Author)</p>	<p><b>Dependencies in language: On the neural encoding of linguistic systems</b> N. J. Enfield (Author)</p>	<p><b>On looking into words (and beyond): Structures, relations, functions</b> Celia C. (Editors)</p>

## } Available formats for books

- } pdf
- } tex
- } bib

## } Indexes in books

- } Language index
- } Subject index
- } Name index

} Indexes are a discovery tool, similar to NERD.

› 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

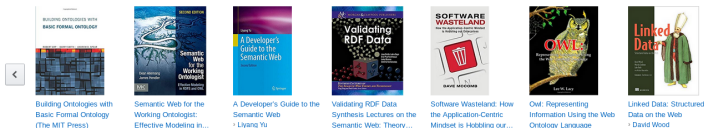




› higher level goals of text and data mining:

› provide better tools for exploration:

Customers who viewed this item also viewed



› automated reasoning:

› gene  $\longleftrightarrow$  protein  
 › protein  $\longleftrightarrow$  disease  
 › gene  $\longleftrightarrow$  protein  $\longleftrightarrow$  disease

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

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

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

Keywords

## } But should I really enter strings there?

Alternative names

**hhbib\_lgcode:**

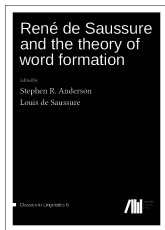
B. 2009)  
Meson de Guadalupe  
Mixtec-Mixtepec  
San Juan Mixtepec  
San Juan Mixtepec-Oaxaca

**lexvo:**

Mixtepec Mixtec [en]

**multitree:**

Eastern Juxtlahuaca Mixtec  
Mixtec, Mixtepec  
Mixteco de San Juan Mixtepec  
Northern Misteko  
Northern Mixteco



**(BnF) Catalogue général**

Accueil > Nom de personne

**Notice de personne**

Sélectionner le format voulu

Personne: René de (1868-1942) (nom international)

Pages : 2 pages

Langue(s) : Français

Sexe : Masculin

Responsable(s) étendue(s) sur les documents : Auteur

Naissance : 1868-03-17, Genève, Suisse

Mort : 1942-12-02, Rome

Docteur en philosophie - Expérimental - Professeur à l'Université de Genève, Suisse

Forme(s) abrégée(s) :

+ Arrière (1868-1942) pseudonyme

Source(s) :

La direction linguistique des mots dans les langues naturelles considérée au point de vue de son application aux langues artificielles : par René de Saussure, 1918 - Nouv. formes de la langue latine "linguistique" / rené de Saussure, 1918 - Les mots

BN Cat. gén.

Identifiant international : ISBN 0000-0003-2051-4706, cf. <http://nrcid.org/0000-0003-2051-4706>

Notice n° : 1789013003556

Création : 14/01/25 Mise à jour : 16/05/21

Actions

Créer la notice : <https://catalogue.bnf.fr/notice/>

Télécharger l'impression

Envoyer par e-mail

Ajouter à mes notices

Signaler une erreur sur cette notice

NOTICES BIBLIOGRAPHIQUES LIÉES

Voici les notices liées en tant que :

- auteur (2)
- sujet (7)

Voici toutes les notices liées (5)

ORCID ID	First/given name	Last/family name
<a href="https://orcid.org/0000-0003-2051-4706">https://orcid.org/0000-0003-2051-4706</a>	Wei	Wang
<a href="https://orcid.org/0000-0001-9168-3297">https://orcid.org/0000-0001-9168-3297</a>	Wei	Wang
<a href="https://orcid.org/0000-0003-2428-8515">https://orcid.org/0000-0003-2428-8515</a>	wei	wang
<a href="https://orcid.org/0000-0002-7025-9435">https://orcid.org/0000-0002-7025-9435</a>	Wei	Wang
<a href="https://orcid.org/0000-0003-1726-5120">https://orcid.org/0000-0003-1726-5120</a>	Wei	Wang
<a href="https://orcid.org/0000-0003-0248-6094">https://orcid.org/0000-0003-0248-6094</a>	Wei	Wang
<a href="https://orcid.org/0000-0002-8598-0831">https://orcid.org/0000-0002-8598-0831</a>	Wei	Wang
<a href="https://orcid.org/0000-0003-1666-7531">https://orcid.org/0000-0003-1666-7531</a>	Wei	Wang
<a href="https://orcid.org/0000-0002-6776-0528">https://orcid.org/0000-0002-6776-0528</a>	Wei	Wang
<a href="https://orcid.org/0000-0002-1568-2396">https://orcid.org/0000-0002-1568-2396</a>	Wei	Wang

## > A grammar of Komnzo

Language: Anta-Komnzo-Wára-Wéré-Kémä



### Classification

- Morehead-Wasur (19)
  - Kanum (4)
  - Morehead-Maró (15)
    - Nambu (8)
    - Tonda (6)
      - Arammba
        - Eastem Tonda (2)
          - Anta-Komnzo-Wára-Wéré-Kémä
          - Anta
          - Kémä
          - Kómnyo
          - Wára
          - Wéré
        - Káncchá
      - Mblale-Ránmo
      - Rema
      - Warta Thuntai
    - Yei

Comments on subclassification

Christian Döhler 2016 :37-42

### References

Showing 1 to 10 of 10 entries

Details	Name	Title	Any field	ca	Year	Pages	Doctype	ca	Provider	da
	<input type="text"/>	<input type="text"/>	<input type="text"/>		<input type="text"/>	<input type="text"/>	--any--		--any--	
<a href="#">citation</a>	Christian Döhler 2016	Komnzo: A language of Southern New Guinea	✓		2016	622	grammar		hh	

Glottocode: wara1294 ISO 639-3: wra

### Map




show big map

Countries

Links


Alternative names

## } A grammar of Komnzo



Participating Archives • OLAC • Delivered by

### OLAC Language Resource Catalog

Search for language resources   

**▼ Navigating the Catalog**

- » Catalog Home
- » Search Strategies
- » Advanced Search
- » New: Records recently added or modified

**▼ Quick Links**


- » Browse by Language
- » Browse by Country
- » Browse by Linguistic Field
- » Browse by Linguistic Type
- » Browse by Language Family


**▼ Contacts**

- » Email Us

**▼ More Information**

- » OLAC Homepage
- » OLAC FAQ
- » Participating Archives

 Powered by the DLA

**Results:**  Showing hits **1 - 50** out of **56**

« First • Previous • Next • Last »

**Wáɿra (Wáɿra) (tci) -- Wáɿra (Wáɿra) (tci)**  
n.a. n.d. The Language Archive at the MPI for Psycholinguistics.

**Wáɿrá (tci) -- Wáɿrá (tci)**  
n.a. n.d. The Language Archive at the MPI for Psycholinguistics.

**Gaída yén nù ní a lam (The widow and her son)**  
Kelta Kurabe (compiler); Kelta Kurabe (depositor); H. Pri (speaker). 2016. Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC).

**Jaugawng a lam (The hunter)**  
Kelta Kurabe (compiler); Kelta Kurabe (depositor); W. Awng (speaker). 2016. Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC).

**Sumpyi a lam (The origin of the flute)**  
Kelta Kurabe (compiler); Kelta Kurabe (depositor); S. Tu (speaker). 2016. Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC).

**Hkra pu n rawng ai lam (Why cicadas do not have intestines)**  
Kelta Kurabe (compiler); Kelta Kurabe (depositor); W. La Tawng (speaker). 2017. Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC).

**Galawng ní u hkai hpe hta sha ai lam (The eagle that ate young chickens)**  
Kelta Kurabe (compiler); Kelta Kurabe (depositor); S. Lu Bu (speaker). 2017. Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC).

**Kasha hkan nu ní a maumwi (The bad child)**  
Kelta Kurabe (compiler); Kelta Kurabe (depositor); L. Roi (speaker). 2017. Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC).

**▼ Currently Used Filters**

- ✓ Query: Wáɿra

**Sort Results By:**

**▼ Possible Sorts:**

- Title
- Id
- Date

**Narrow Results By:**

**▼ Archive**

- Pacific And Regional Archive for Digital Source Cultures (PARADISEC)
- The Language Archive at the MPI for Psycholinguistics

**▼ Online**

- Yes
- No

**▼ Subject language**

- Kachin

**▼ Language family**

- Sino-Tibetan
- Tibeto-Burman

- › 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



**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



- > Is there an “inallative” in *Roma**ni** ite domu**m***?
- > Is there an “inallative” in ***au** foyer*?
- > Is there an “inallative” in ***thuis***?
- > Can you equate the usages in the three examples?
- > take-home-message: it’s complicated, and automated reasoning will not work.

# A typology of questions in Northeast Asia and beyond

An ecological perspective

Andreas Hölzl

Studies in Diversity Linguistics



- › *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

## 5.8 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 ABI	naa-n-aas	tzaa-n-aas	xaa-n-aas
LAT	naa-sh	tzaa-sh	xaa-sh
PROI	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	seng	yoon	aly	sejer	xan
Mongolian	sen	youn	alyh	sejee	xam
Buryat	sen	yūn	ali	sejee	xama
Khannigan Mongol	ken	yeen	ali	kejee	kama
Ordos	ken	yūn	ali	kejee	kai
Written <i>Qirai</i>	ken	yau/n	ali	kejee	xamig(h)a
Qirai	ken	yau/n	al - al-k	kene	xama
Kalmuk	ken	yūn	aly(-k)	keza	xama, aly-d
Shira Yuzhur	ken	yima	aali	kejee	xama
Santa	ken	yang	ali	giczi	khala
Bonan	hang	yang	ane	keet(-)	hala
Kanghis	ko	jo - jai	am(ve)	gadje	yana
Huzia Mongghul	ken	ya/n	ali	kijet	an-j(i)
Minhe Mangghur	kan	ya, yang	alyge	kejie	ang(ji)
Monghol	ken	kyan	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-linguistic 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**

### ALY

Normalized: Upper Aramite language

Domains: Astronomy, Biology, Geography, Sociology

cont: 0.4157

**Aramite** or **Aramite** or more specifically **Upper Aramite** (Upper Aramite), is a **Arabic** spoken in and around **Aramite** (Aramite in Aramite) in the **Southwest**, **Australia**. The name is sometimes spelled **Aramite** or **Aramite**.

Freebase ID	/m/59288
writing system	Latin script
number of speakers	[exact figure]
instance of	Dialect continuum
UNESCO Atlas of the World's Languages in Danger ID	168

References: [W](#) [B](#)

- › NERD retrieved some pretty specialized concepts
  - › Recall is good
- › NERD also retrieved a lot of irrelevant concepts (“South America”) or lookalikes (business names)
- › 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.

- › 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