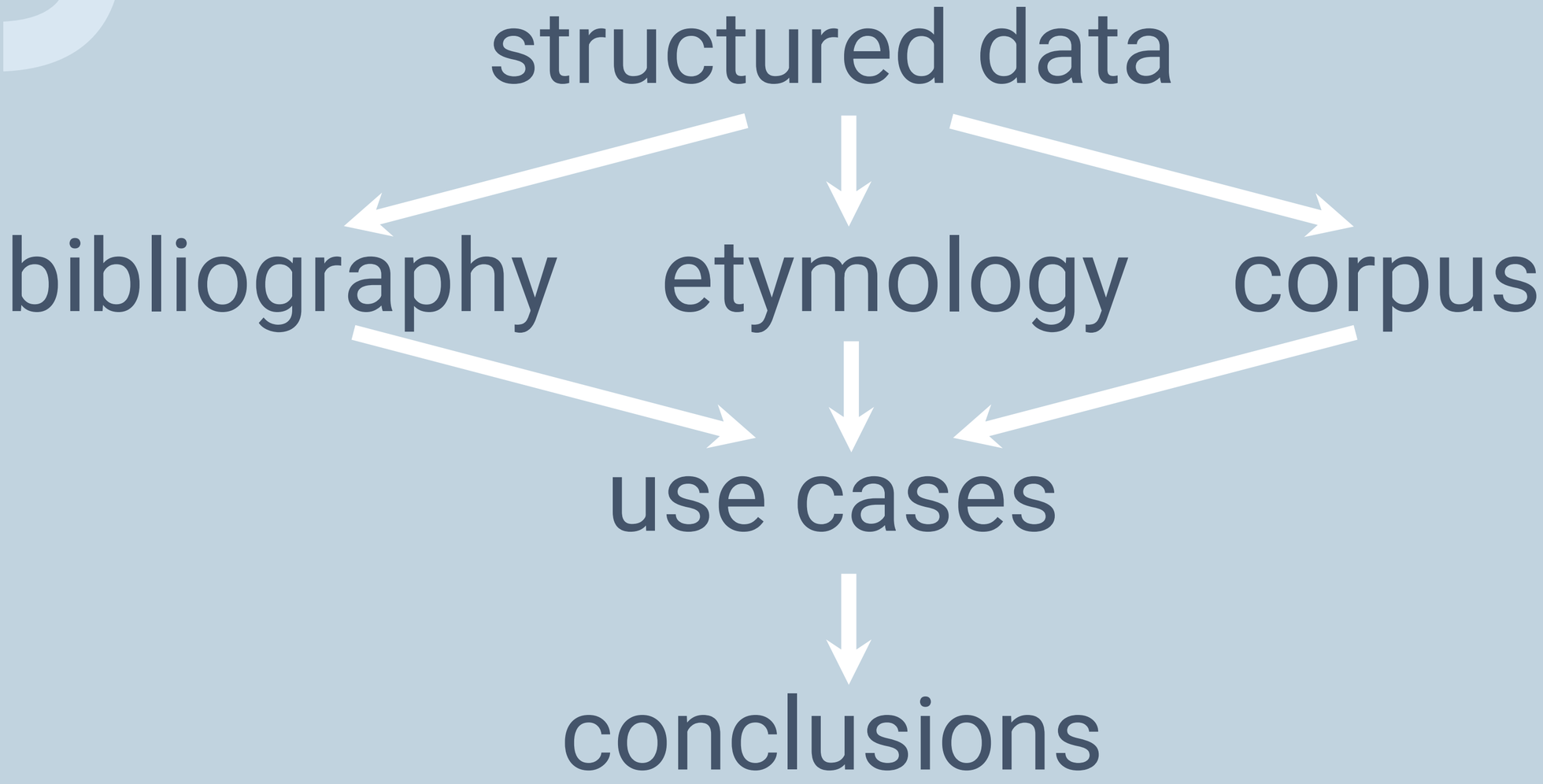
EESTI KEELE INSTITUUT

Variations on the theme of structured data

Arvi Tavast
Eesti Keele Instituut | EKI

Digital Dictionary Database for Slovenian: unstructured, semi-structured and structured data in modern lexicography

Simon Krek, “Jožef Stefan” Institute, Slovenia


structured data
bibliography
etymology
corpus
use cases
conclusions

EESTI
KEELE
INSTITUUT

STRUCTURED DATA

For every data element, we know what it is

- Each field contains a single piece of data
- Each piece of data is located on a single field

STRUCTURED DATA

piano – 1. *adj* flat, level, 2. *adv* quietly, 3. *n* plane (*geometry*)

piano – 1. *adj* flat, level, 2. *adv* quietly, 3. *n* plane (*geometry*)

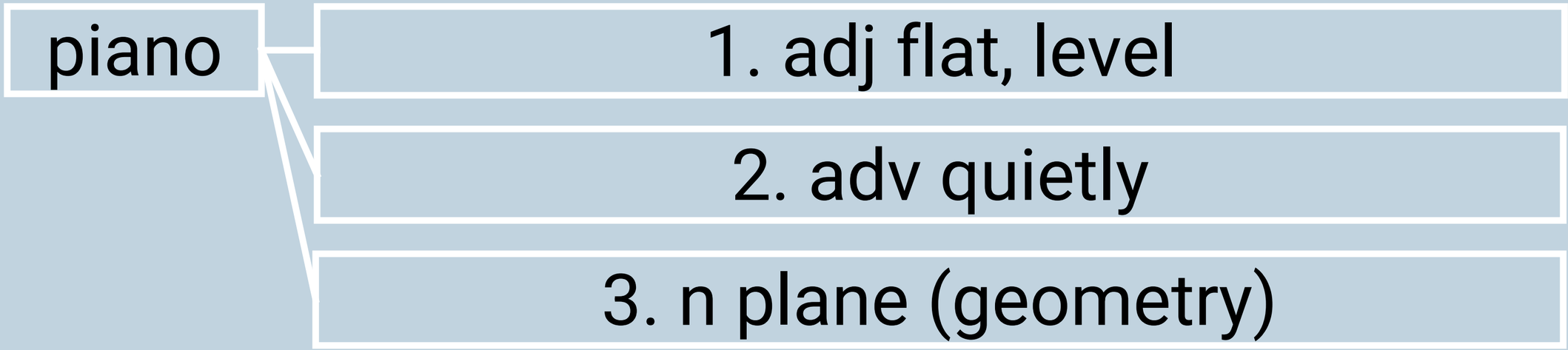
piano	1. adj flat, level, 2. adv quietly, 3. n plane (geometry)

piano – 1. *adj* flat, level, 2. *adv* quietly, 3. *n* plane (*geometry*)

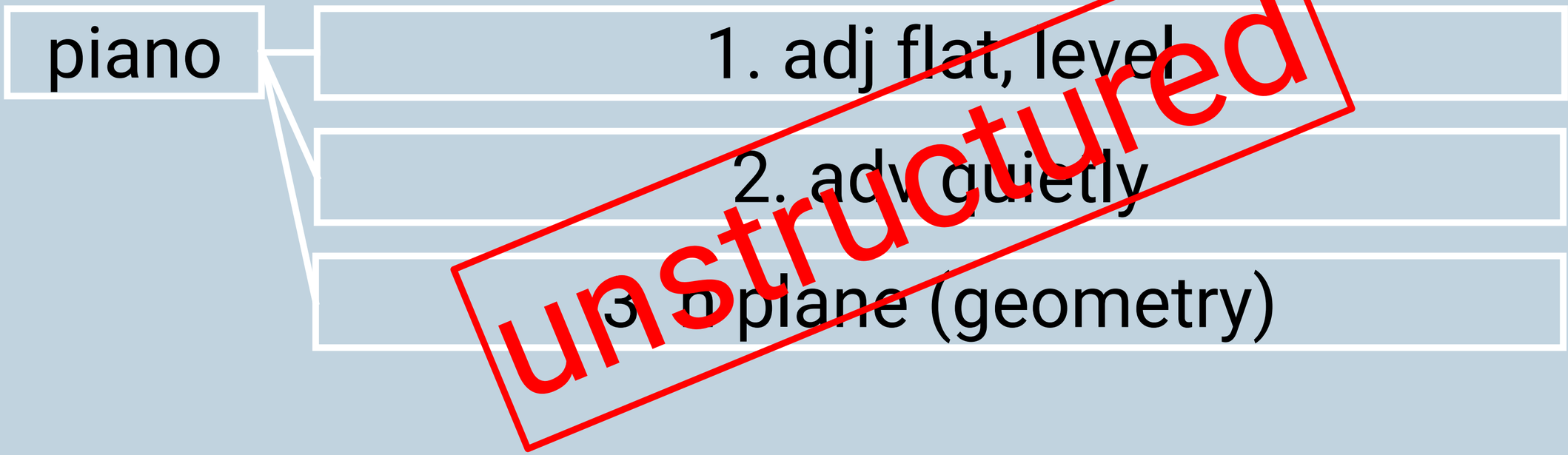
piano	1. adj flat, level, 2. adv quietly, 3. n plane (geometry)

unstructured

piano – 1. *adj* flat, level, 2. *adv* quietly, 3. *n* plane (*geometry*)



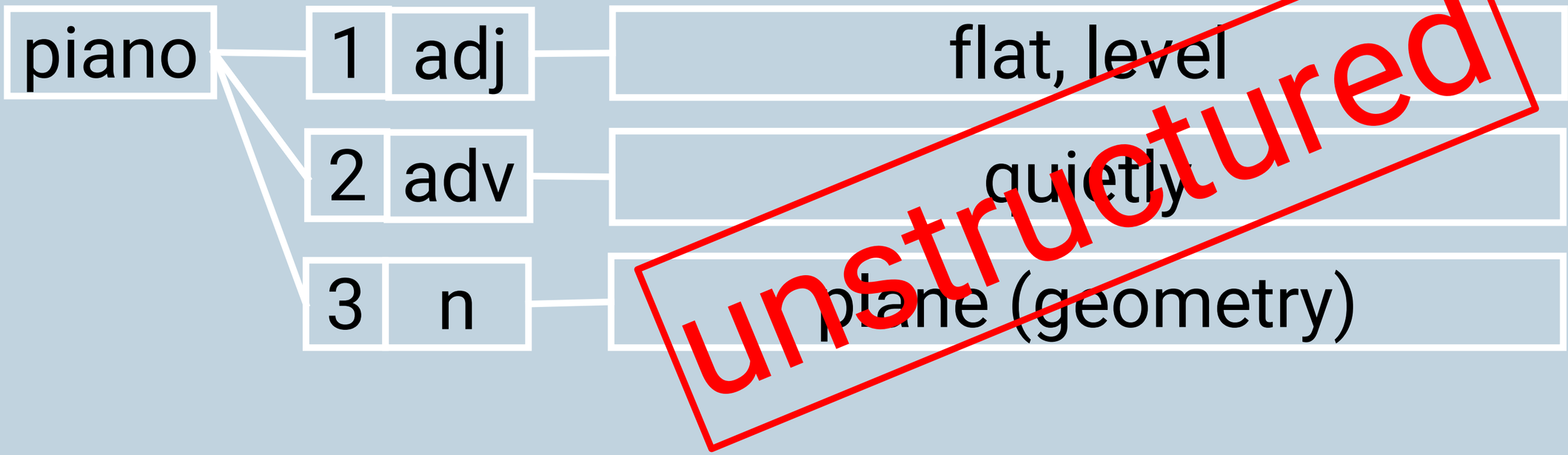
piano – 1. *adj* flat, level, 2. *adv* quietly, 3. *n* plane (*geometry*)



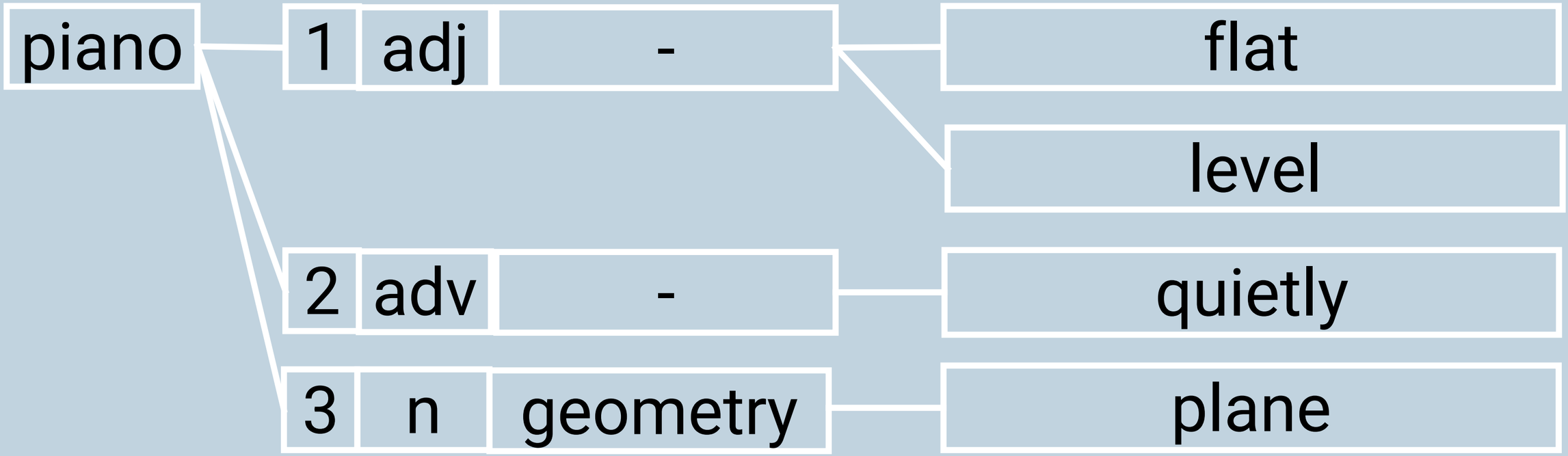
piano – 1. *adj* flat, level, 2. *adv* quietly, 3. *n* plane (*geometry*)

EESTI KEELE INSTITUUT

piano – 1. *adj* flat, level, 2. *adv* quietly, 3. *n* plane (*geometry*)



piano – 1. *adj* flat, level, 2. *adv* quietly, 3. *n* plane (*geometry*)



EESTI
KEELE
INSTITUUT

BIBLIOGRAPHIC REFERENCES

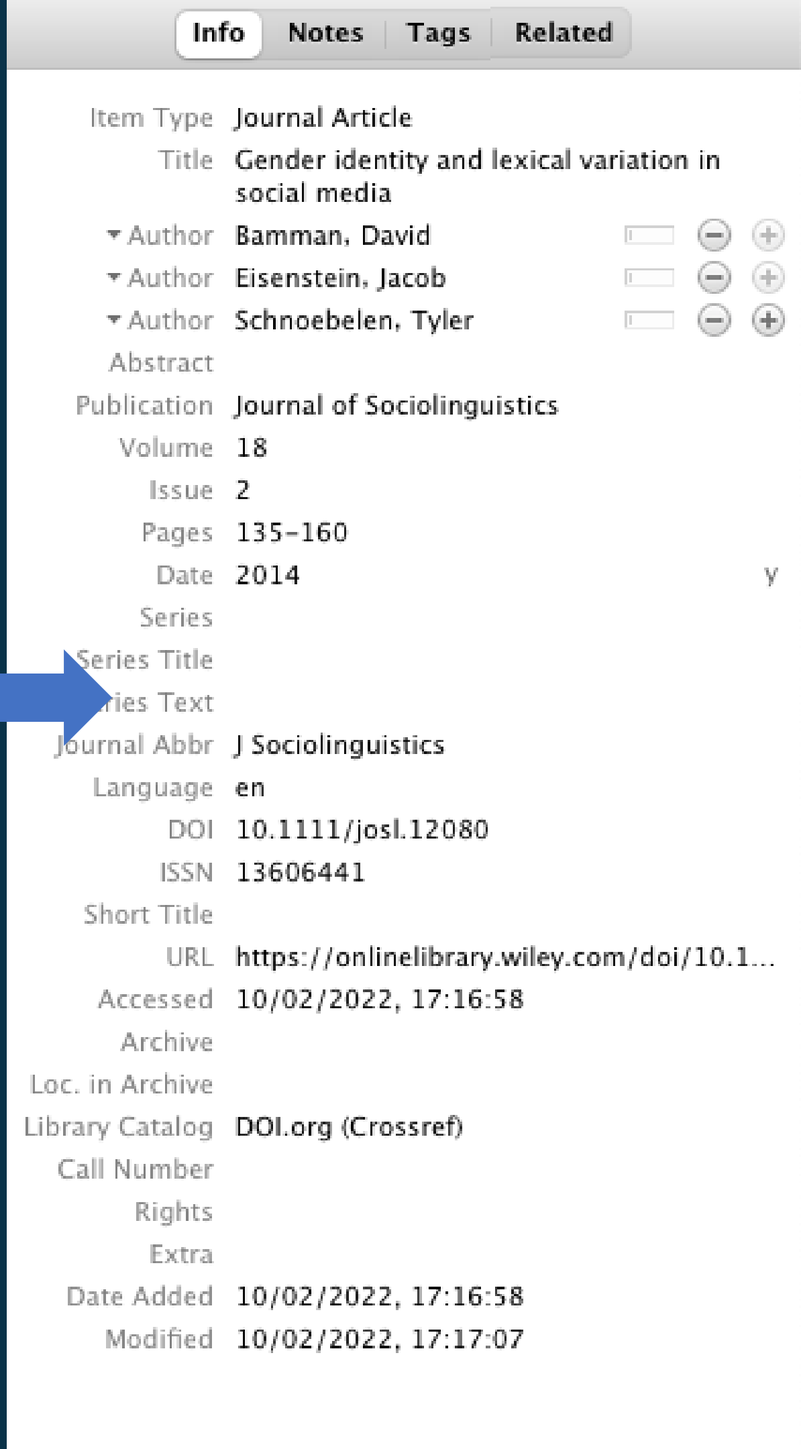
Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen. 2014. ‘Gender Identity and Lexical Variation in Social Media’. *Journal of Sociolinguistics* 18 (2): 135–60.

Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen. 2014. ‘Gender Identity and Lexical Variation in Social Media’. *Journal of Sociolinguistics* 18 (2): 135–60.

Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen. 2014. ‘Gender Identity and Lexical Variation in Social Media’. *Journal of Sociolinguistics* 18 (2): 135–60.

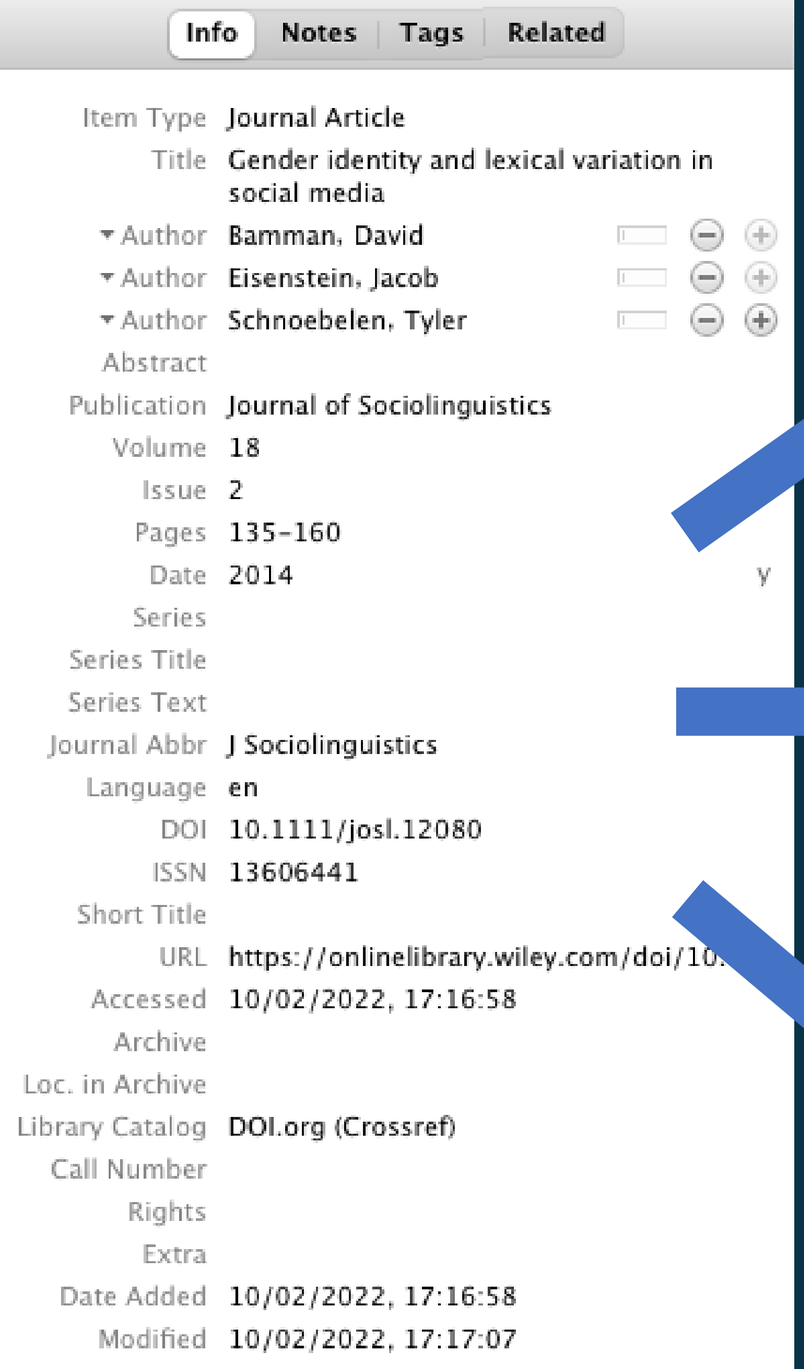
Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen. 2014. ‘Gender Identity and Lexical Variation in Social Media’. *Journal of Sociolinguistics* 18 (2): 135–60

Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen. 2014. ‘Gender Identity and Lexical Variation in Social Media’. *Journal of Sociolinguistics* 18 (2): 135–60.



Info | Notes | Tags | Related

Item Type	Journal Article
Title	Gender identity and lexical variation in social media
Author	Bamman, David
Author	Eisenstein, Jacob
Author	Schnoebelen, Tyler
Abstract	
Publication	Journal of Sociolinguistics
Volume	18
Issue	2
Pages	135–160
Date	2014
Series	
Series Title	
Series Text	
Journal Abbr	J Sociolinguistics
Language	en
DOI	10.1111/josl.12080
ISSN	13606441
Short Title	
URL	https://onlinelibrary.wiley.com/doi/10.1...
Accessed	10/02/2022, 17:16:58
Archive	
Loc. in Archive	
Library Catalog	DOI.org (Crossref)
Call Number	
Rights	
Extra	
Date Added	10/02/2022, 17:16:58
Modified	10/02/2022, 17:17:07



Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen. 2014. 'Gender Identity and Lexical Variation in Social Media'. *Journal of Sociolinguistics* 18 (2): 135–60.

Chicago Manual of Style 17th edition (author-date)

Bamman, David, et al. 'Gender Identity and Lexical Variation in Social Media'. *Journal of Sociolinguistics*, vol. 18, no. 2, 2014, pp. 135–60.

Modern Language Association 9th edition

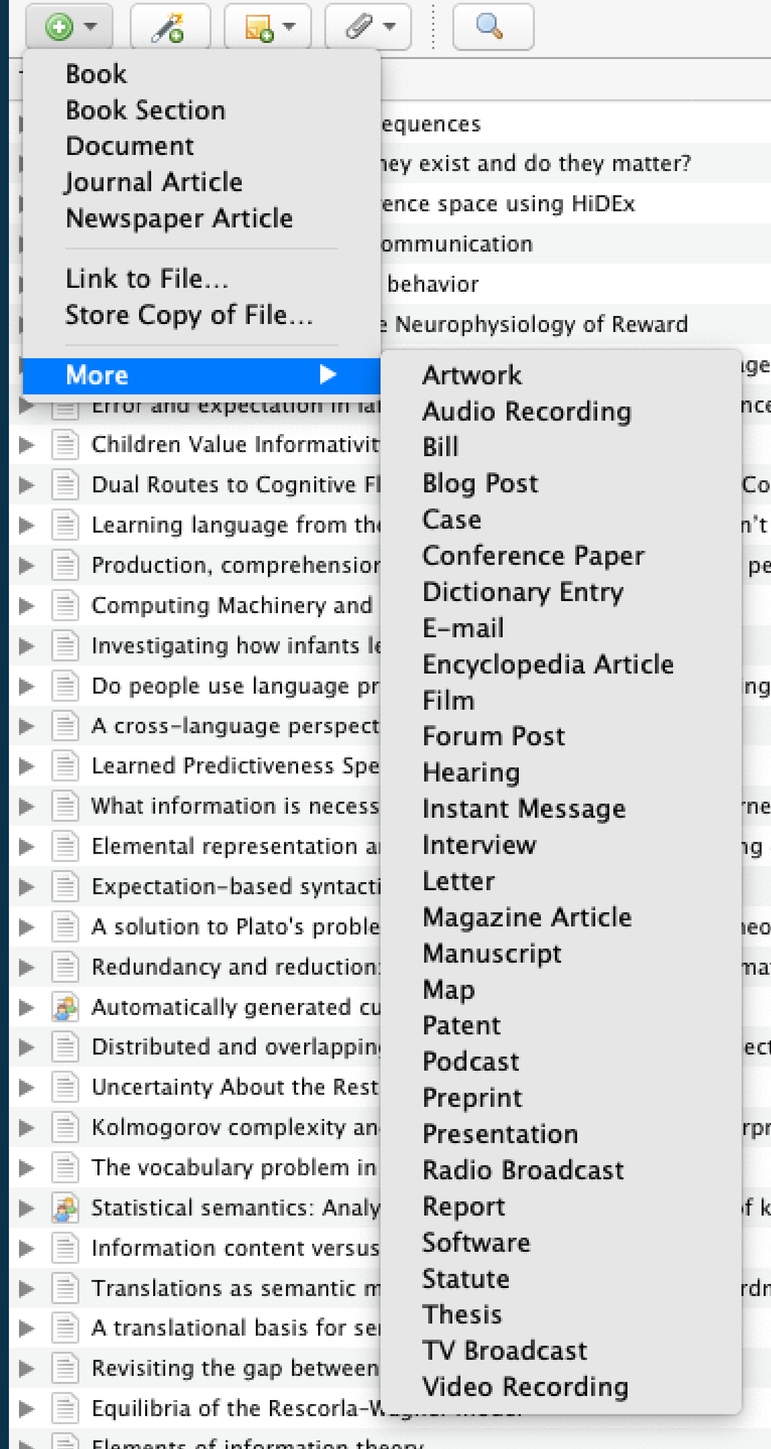
Bamman, D., Eisenstein, J. & Schnoebelen, T. Gender identity and lexical variation in social media. *J Sociolinguistics* **18**, 135–160 (2014)

Nature

Multiple document types

Multiple contributor types:

- author
- editor
- editor-in-chief
- but no publisher-in-charge (vastutav väljaandja)



EESTI
KEELE
INSTITUUT

ETYMOLOGY

Etymology module

DMLex standard (ver 1.0 working draft 01), as well as Ekilex: unstructured AND structured

Päritolu LAENSÕNA

et *duodeenum*
qat *duodenum* 'kaksteistsõrmik' lühenenud kuju keskladina meditsiinikeelendist *duodenum* digitorium 'kaheteistkümne sõrme ruum', mis on tõlkelaen kreeka sõnast *dodekadaktylon* 'kaksteist sõrme pikk'. Keskladina väljendi tõi käibele *Canon* Avicennae tõlkija Gerard Cremonast (u 1114-1187)

Päritolu LAENSÕNA

et *safiir* Piiblisõnavara hulka kuuluv varane kirjakeelne laen
de *Saphir* 'safiir'
la *sapphīrus* 'safiir'
el *sappheiros* 'safiir'
he *sappīr* 'safiir'

Päritolu LAENSÕNA

et *kohv*
de *Koffee* 'kohv'
en *coffee* 'kohv'
it *caffè* 'kohv'
tr *kahve* 'kohv'
ar *qahwah* 'kohv'

EESTI
KEELE
INSTITUUT

CORPUS DATA

**People don’t like the corpus,
because it contains real language**

Ondřej Matuška, Lexical Computing

Corpus queries: “real” language?

Veebilauseid

⚠ Veebilaused on automaatselt valitud ning võivad sisaldada vigu.

Must vari lume valgel **tasutal** koputab uksele.

Vesi ja **tasutal** ilusad Alpid, mida rohkemat tahta?

Paslik on meelde tuletada ja kuulata tänases maailmas toimuva **tasutal** spetsialisti soovitusi käitumiseks antud olukorras.

Põõsad lisavad kodusele **tasutale** rohelise ja looduslähedase tausta.

Eestisse püütakse lasta vaid kristliku **tasutaga** pagulasi, kel oleks lihtsam meie ühiskonda sulanduda.

Tegime veel kiriku **tasutal** pilti, ning lõpuks keerasime auto kodu poole.

Kokkuvõtteks, kui soovid endale sõpra, kellega rääkida aktiivselt ja huvitava **tasutaga**, siis kirjuta.

Peamat ise oli armastusest, moraalist, ihast

EESTI
KEELE
INSTITUUT

WHAT IS A DICTIONARY FOR?

EESTI KEELE INSTITUUT

Possible use cases

Use case	Unstructured	Structured
What is the source of this example	✓	✓
Where does this word come from	✓	✓
Show examples with sources published earlier than 1960	✗	✓
Show words derived from Greek via French	✗	✓
Help for avoiding mistakes	✗	✓
Help for avoiding duplicates, conflicts and inconsistencies	✗	✓
Easy to enter non-standard values	✓	✗
Can be automatically converted to the opposite	✗	✓
Data model complexity is manageable	✓	✗

EESTI KEELE INSTITUUT

Two types of users

Human

- Tolerates ambiguity
- Can easily make inferences
- Can correct mistakes on the fly
- Does it need to be correct?

Machine

- Needs to be told explicitly
- Believes everything verbatim
- Does it need to know?

EESTI KEELE INSTITUUT

Two types of users

Human

- Tolerates ambiguity
- Can easily make inferences
- Can correct mistakes on the fly
- Does it need to be correct?

Machine

- Needs to be told explicitly
- Believes everything verbatim
- Does it need to know?

Structured or not? The answer depends on the use case