



# NLP-Methoden in der Klassischen Philologie

## Word Embeddings

Daidalos-Kickoff-Workshop, HU Berlin, 14.6.2024

Dr. Andrea Beyer & Konstantin Schulz (Humboldt-Universität zu Berlin)



— Digital Research for All —

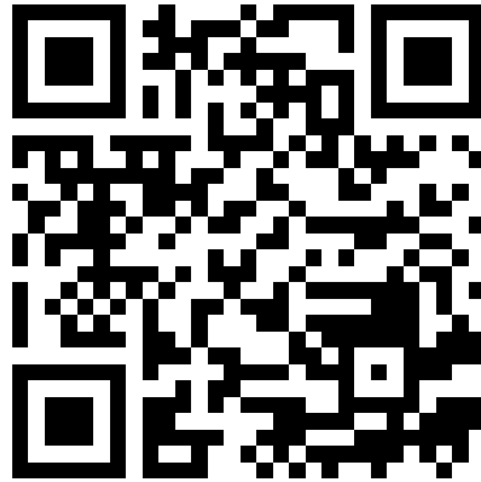


Gefördert durch  
Deutsche  
Forschungsgemeinschaft

# Folien

<https://kurzlinks.de/embeddings-klassphil>

Weitere Materialien in  
der Zenodo-  
Community Daidalos!





# DH-Methoden: Natural Language Processing (NLP)

01

DH: AI, ML, NLP

02

Digital Classics:  
Ausgewählte NLP-  
Methoden

03

Word Embeddings:  
Vektoren als  
Repräsentation

04

Embeddings in der  
Klassischen  
Philologie

# 01 | DH: AI, ML, NLP

Digital Humanities:  
Artificial Intelligence,  
Machine Learning,  
Natural Language Processing



— daidalos —  
— Digital Research for All —



Gefördert durch  
Deutsche  
Forschungsgemeinschaft



# Arbeitsfelder der Digital Humanities (DH)

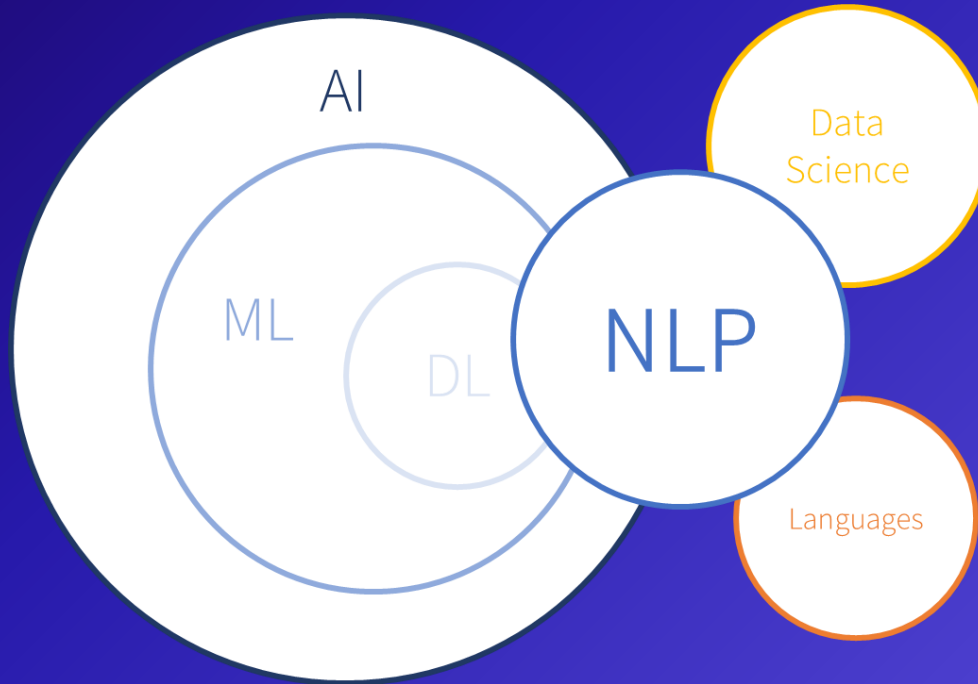


# Der Einsatz von digitalen Methoden umfasst ...

- ... einen Text oder ein Artefakt digital zu erfassen,
- ... die gewonnenen Daten zu organisieren und zu speichern,
- ... die Daten zu verarbeiten,
- ... die Ergebnisse der Verarbeitung darzustellen und
- ... die Daten für eine Nachnutzung zur Verfügung zu stellen.

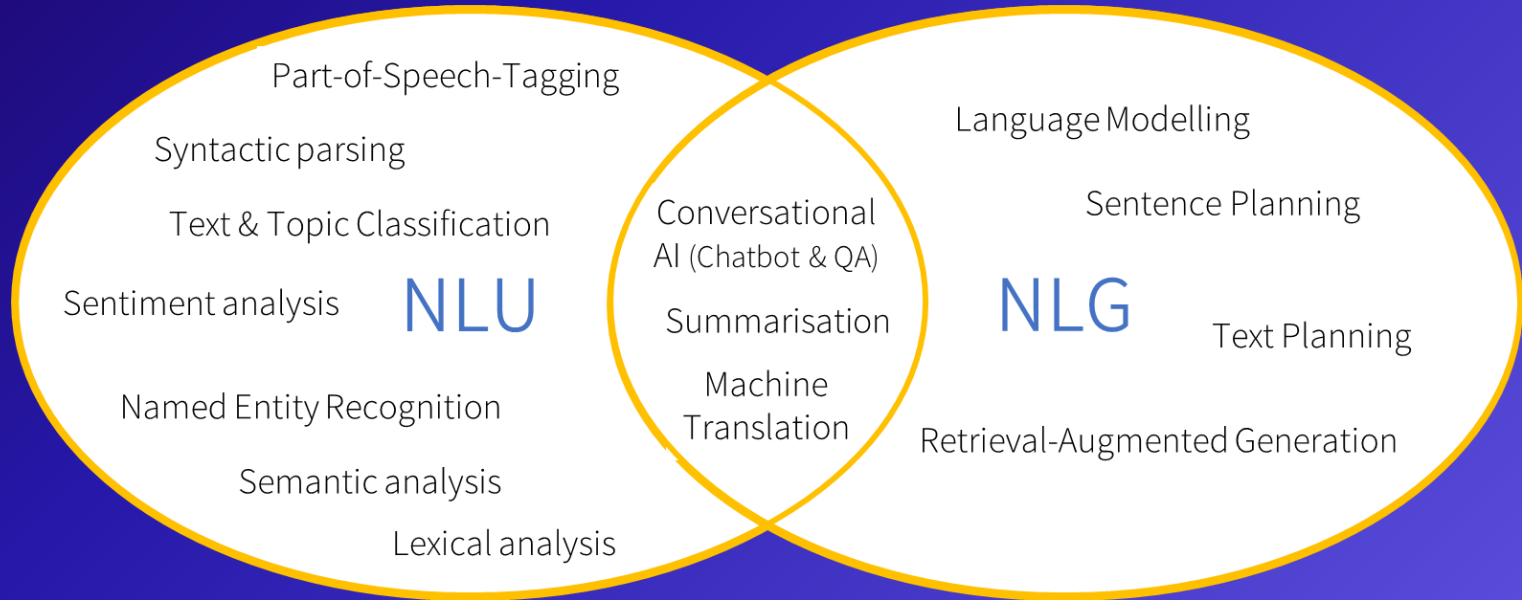


# NLP im Verhältnis zu KI



AI – Artificial Intelligence  
ML – Machine Learning  
DL – Deep Learning  
NLP – Natural Language Processing

# NLP = NLU + NLG



- NLP – Natural Language Processing
- NLU – Natural Language Understanding
- NLG – Natural Language Generation
- QA – Question Answering



# Wie versteht eine Maschine Sprache?

- Sprachdaten sind unstrukturiert, d.h. nicht mit zusätzlichen Informationen erklärt (annotiert).
- Sie müssen also strukturiert werden, damit sie von Maschinen analysiert werden können (Sprachvorverarbeitung).
- Die Annotationen können sich z. B. auf Wortstellung, Wortkombinationen, Semantik und Morphologie beziehen.
- Maschinelle Sprachverarbeitung (NLP) analysiert mithilfe dieser Merkmale Sprache, berechnet Zusammenhänge, erkennt Muster und erstellt Modelle.

# Vorverarbeitung: Korpusaufbau

Zerlegen und Annotieren der Texte im Korpus

1. Zusammenstellen des Inhalts (Rohtext)
2. Zerlegen des Rohtexts in Paragraphen
3. Zerlegen der Paragraphen in einzelne Sätze (Segmentation)
4. Zerlegen der Sätze in einzelne Wörter (Tokenisation)
5. Zuordnung der Tokens zu den passenden Lemmata (Lemmatisation)
6. Bestimmen der Wortarten (POS-Tagging)
7. Erste (oberflächliche) Korpus-Analysen

und anschließend Transformation des Korpus mit weiteren NLP-Methoden

# Segmentation (Zerlegen in Sätze)

```
1 <id n="02">
2 <sentence n="16">}. De duodecim signis</sentence>
3 <p n="1"> <sentence n="17">Signa sunt in caelo duodecim.
  </sentence><sentence n="18">Aries beneficio Liberi, quod is cum
  exercitum in Indiam per Libyam duceret per loca sicca et arenosa,
  qua aquae inopia esset et exercitus eius siti adfligeretur, aries
  eis aquam demonstravit; et ob id a Libero Iovis Ammon est
  appellatus, eique fanum magnificum fecit ad eum locum ubi aquam
  invenit;</sentence><sentence n="19">quod abest ab Aegypto et
  Alexandria milia passuum novem.</sentence><sentence n="20"> ob
  eam rem a Iove petiit ut in<ter> sidera reciperetur. alii putant
  eum esse qui Hellen et Phryxum vexerit.</sentence></p>
4 <p n="2"><sentence n="21">Taurus beneficio Iovis, quem Iuppiter a
  Neptuno fratre per gratiam abduxit; </sentence><sentence
  n="22">qui sensum humanum figura tauri continebat hisque Iovis
  iussu Europam Agenoris filiam Sidonia adludens decepit et eam
  Cretam deportavit. </sentence><sentence n="23">ob eam rem
  Iuppiter in sideribus eum dignatus est immortalis
  memoria.</sentence></p>
5 <p n="3"><sentence n="24">Gemini, qui <dii> Samothraces
  nominantur [esse]; </sentence><sentence n="25">quorum argumentum
  nefas est pronuntiare praeter eos qui mi<ste>riis
  prae[sto]sunt.</sentence><sentence n="26"> alii Castorem et
```



Beispiel: *Liber memorialis* des Ampelius (ca. 2. Jh.)

# Zuweisung von Token zu Lemma und Wortart

```

1 <id n="02">
2 <sentence n="16">
3 <w n="1" form="2" lemma="num._arab." pos="NUM"></w>
4 <w n="2" form="." lemma="." pos="PUNCT"></w>
5 <w n="1" form="De" lemma="de" pos="ADP"></w>
6 <w n="2" form="duodecim" lemma="duodecim" pos="NUM"> </w>
7 <w n="3" form="signis" lemma="signum" pos="NOUN"> </w></sentence>
8 <p n="1"> <sentence n="17">
9 <w n="5" form="Signa" lemma="signum" pos="NOUN"> </w>
10 <w n="6" form="sunt" lemma="sum" pos="VERB"> </w>
11 <w n="7" form="in" lemma="in" pos="ADP"> </w>
12 <w n="8" form="caelo" lemma="caelum" pos="NOUN"> </w>
13 <w n="9" form="duodecim" lemma="duodecim" pos="NUM"> </w>
14 <w n="10" form="." lemma="." pos="PUNCT"> </w></sentence>
15 <sentence n="18">
16 <w n="1" form="Aries" lemma="Aries" pos="NOUN"> </w>
17 <w n="2" form="beneficio" lemma="beneficium" pos="NOUN"> </w>
18 <w n="3" form="Liberi" lemma="libeor" pos="VERB"> </w>
19 <w n="4" form="," lemma="," pos="PUNCT"> </w>
20 <w n="5" form="quod" lemma="quod" pos="SCONJ"> </w>
21 <w n="6" form="is" lemma="is" pos="ADJ"> </w>
22 <w n="7" form="cum" lemma="cum" pos="SCONJ"> </w>
23 <w n="8" form="exercitum" lemma="exercitus" pos="VERB"> </w>
24 <w n="9" form="in" lemma="in" pos="ADP"> </w>

```



Beispiel: *Liber memorialis* des Ampelius (ca. 2. Jh.)

# Erste (oberflächliche) Korpus-Analysen

- Erstelle eine Liste der 100 häufigsten Wörter und sortiere sie absteigend nach Frequenz. (Frequenzanalyse)
- Wie oft kommt bei Ampelius das Wort *magnificus* vor? (Frequenzanalyse)
- Wie viele Adjektive verwendet Ampelius? (Frequenzanalyse)
- Gib alle Vorkommen von *atque* mit je 2 Wörtern davor und danach. (Key Word in Context)
- Berechne die lexikalische Variation / die durchschnittliche Satzlänge / die Textlänge / die durchschnittliche Länge der Kapitel.

# 02 | Digital Classics: NLP-Methoden

Automatische Erkennung von  
Wortarten, Eigennamen, Gefühlen und  
Themen



**dardalos**  
— Digital Research for All —



Gefördert durch  
**DFG**  
Deutsche  
Forschungsgemeinschaft

# Part-of-Speech-Tagging (POS-Tagging)

## Ziel

Bestimmung  
der Wortarten

## Einsatz

Datenannotation für komplexere  
Verfahren, z. B. Textklassifikation,  
Autorschaft

## Methoden

regelbasiert, statistisch, ML, DL,  
auf der Basis eines festgelegten  
Tagsets





	Dominus	Erasmus	plurimam	salutem	tibi	adscribere	iussit	.
<b>GS:</b>	NOUN	PROPN	ADJ	NOUN	PRON	VERB	VERB	PUNCT
<b>LC:</b>	NOUN	PROPN	ADJ	NOUN	PRON	NOUN	VERB	PUNCT
<b>RDR:</b>	NOUN	VERB	ADJ	NOUN	PRON	VERB	VERB	VERB
<b>GPT-4:</b>	NOUN	PROPN	ADJ	NOUN	PRON	VERB	VERB	PUNCT

# Tagger-Vergleich



getaggt mit Gold Standard (GS), LatinCy (LC), RDRPOSTagger (RDR) und GPT-4.  
Vgl. Stüssi & Ströbel, 2024: Part-of-Speech Tagging of 16th-Century Latin with GPT, 197.





# Named Entity Recognition (NER)

## Ziel

Informationsextraktion Texten:  
automatische Identifikation und  
Klassifikation von Eigennamen  
(Person, Ort, Volk)

## Einsatz

Datenannotation für  
Textklassifikation, Sentiment-  
Analyse, Social-Network-  
Analyse etc.

## Methoden

regelbasiert, statistisch, ML, DL  
sowie deren Kombination



## Suet. Iul. 24

sed cum **Lucius Domitius PERSON** consulatus candidatus palam minaretur consulem se effecturum quod praetor nequisset adempturumque ei exercitus, **Crassum PERSON** **Pompeiumque PERSON** in urbem prouinciae suae **Lucam LOC** extractos compulit, ut detrudendi **Domitii PERSON** causa consulatum alterum peterent, perfecitque tutrumque, ut in quinquennium sibi imperium prorogaretur. qua fiducia ad legiones, quas a re publica acceperat, alias priuato sumptu addidit, unam etiam ex **Transalpinis LOC** conscriptam, uocabulo quoque **Gallico LOC** — **Alauda LOC** enim appellabatur—, quam disciplina cultuque Romano institutam et ornatam postea uniuersam ciuitate donauit. nec deinde ulla belli occasione, ne iniusti quidem ac periculosi abstinuit, tam foederatis quam infestis ac feris gentibus ultro lacessitis, adeo ut senatus quondam legatos ad explorandum statum Galliarum mittendos decreuerit ac nonnulli dedendum eum hostibus censuerint. sed prospere de cedentibus rebus et saepius et plurium quam quisquam umquam dierum supplicationes impetrauit.

# Latein: LatinCy Personen & Orte



ὁ δὲ **Καίσαρ PER** ἔν τε **Κελτοῖς LOC** καὶ **Βρεττανοῖς MISC** πολλὰ καὶ λαμπρὰ εἰργασμένος, ὅσα μοι περὶ **Κελτῶν MISC** λέγοντι εἶρῃται, πλοῦτου γέμων ἐς τὴν ὁμορον τῆ **Ἰταλῖα LOC** **Γαλατῖαν LOC**, τὴν ἀμφὶ τὸν **Ἥριδανόν LOC** ποταμόν, ἦκεν, ἐκ συνεχοῦς πολέμου τὸν στρατὸν ἀναπαύσων ἐπ' ὀλίγον. ὅθεν αὐτῷ περιπέμποντι ἐς **Ῥώμην LOC** πολλὰ πολλοῖς χρήματα αἱ τε ἐτήσιοι ἀρχαὶ παρὰ μέρος ἀπήντων καὶ οἱ ἄλλως ἐπιφανεῖς ὅσοι τε ἐς ἡγεμονίας ἐθνῶν ἢ στρατοπέδων ἐξήεσαν, ὡς ἑκατὸν μὲν ποτε καὶ εἴκοσι ῥάβδους ἀμφ' αὐτὸν γενέσθαι, βουλευτὰς δὲ πλείους διακοσίων, τοὺς μὲν ἀμειβομένους ὑπὲρ τῶν ἤδη γεγονότων, τοὺς δὲ χρηματιουμένους, τοὺς δ' ἄλλο τι τοιοῦτότροπον αὐτοῖς ἐξεργασομένους. πάντα γὰρ ἤδη διὰ τούτου ἐπράσσετο στρατιάς τε πολλῆς οὐνεκα καὶ δυνάμειος χρημάτων καὶ σπουδῆς ἐς ἅπαντας φιλανθρώπου. ἀφίκοντο δ' αὐτῷ καὶ **Πομπήιος PER** καὶ **Κράσσοσ PER**, οἱ κοινωοὶ τῆς δυναστείας, καὶ αὐτοῖς βουλευομένοις ἔδοξε **Πομπηίων PER** μὲν καὶ **Κράσσον PER** αὐθις ὑπατεῦσαι, **Καίσαρι PER** δ' ἐς τὴν ἡγεμονίαν ὧν εἶχεν ἐθνῶν, ἄλλην ἐπιψηφισθῆναι πενταετίαν. ὧδε μὲν ἀπ' ἀλλήλων διεκρίθησαν, **Πομπηίω PER** δ' ἐς τὴν ὑπατείαν ἀντιπαρήγγελλε **Δομίτιος PER** Αἰνόβαρβος· καὶ τῆς κυρίας ἡμέρας ἄμφω κατήεσαν ἔτι νυκτὸς ἐς τὸ πεδίον ἐς τὴν χειροτονίαν. τῶν δ' ἀμφ' αὐτοῦς ἔριδες ἦσαν καὶ συνεπλέκοντο, μέχρι τις τὸν **Δομιτίου PER** δοξοῦχον ἐπάταξε ξίφει. καὶ φυγὴ μετὰ τούτου ἦν, **Δομιτίος PER** τε αὐτὸς ἐς τὴν οἰκίαν διεσώζετο μόλις, καὶ **Πομπηίου PER** τὴν ἐσθῆτά τινες ἡμαγμένην ἔφερον οἴκαδε. παρὰ τοσοῦτον ἐκάτερος ἤλθε κινδύνου.

# Griechisch: flair\_grc\_bert\_ner Personen, Orte & Völker



# Text Classification

## Ziel

Zuweisung von Kategorien an Text(e), z. B. zur Unterscheidung von Genres

## Einsatz

Merkmalszuweisung für Autorschaft, Sentiment-Analyse, Topic Modelling, Bewerten grammatischer Korrektheit

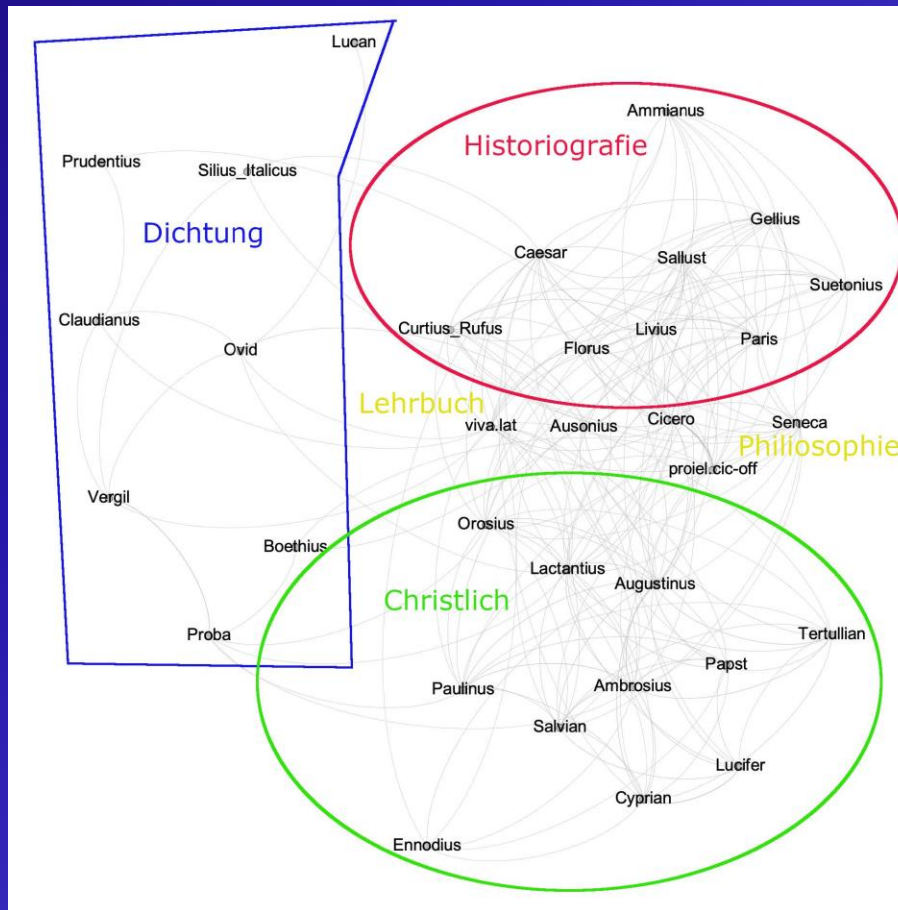
## Methoden

regelbasiert, statistisch, ML, DL  
sowie deren Kombination





# Autorencluster



# Topic Modelling

## Ziel

Exploration versteckter  
semantischer Strukturen in  
großen Textkorpora

## Einsatz

Vergleich von Texten,  
Merkmalszuweisung für  
Autorschaft, Textklassifikation,  
Motive bzw. Topik

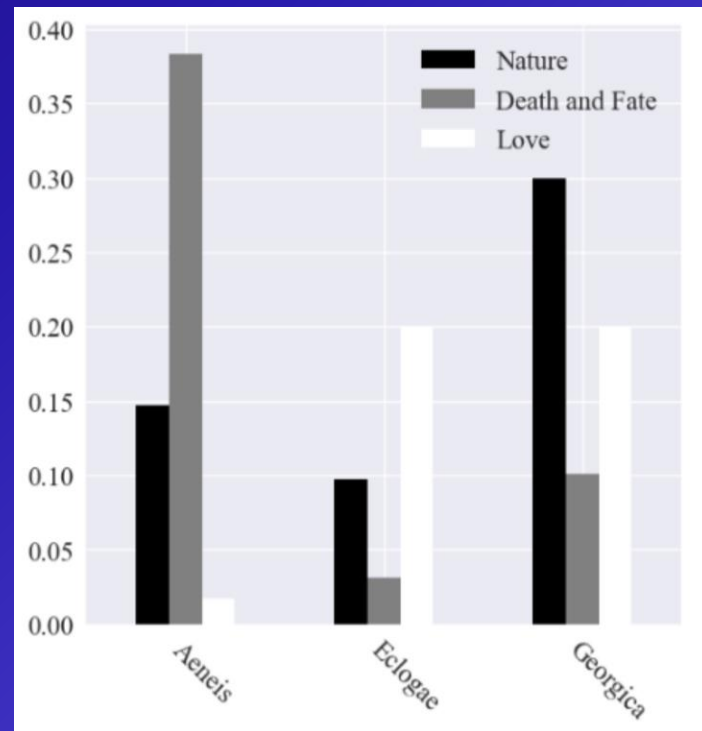
## Methoden

regelbasiert,  
statistisch





# Gewichtung von Themen in den Werken Vergils



# Sentiment Analysis

## Ziel

Analyse von dargestellten  
menschlichen Gefühlen,  
Empfindungen, Meinungen

## Einsatz

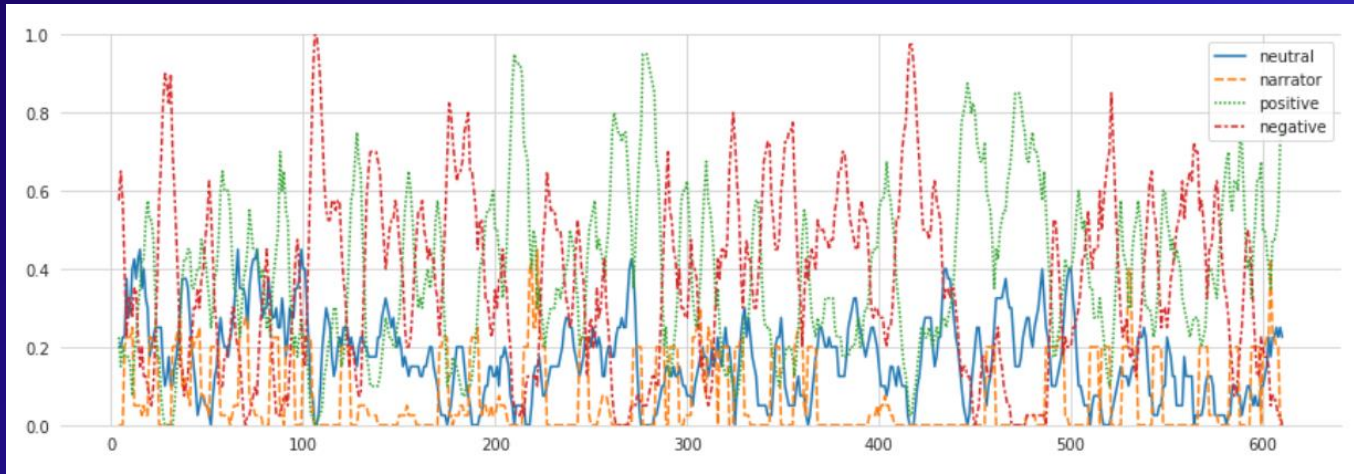
Bewertung von Meinungen oder  
Stimmungen im Verhältnis zum  
Handlungsverlauf,  
Figurencharakterisierung,  
Genreklassifikation

## Methoden

lexikonbasiert, manuell,  
statistisch, ML







# 1. Buch der Ilias: Einschätzung der Annotierenden pro Vers

(positiv, negativ, neutral, Erzähler)

# 03 | Word Embeddings: Vektoren als Repräsentation



# Word Embedding

## Ziel

Repräsentation eines Wortes  
(Satzes, Paragraphen,  
Dokuments) als Vektor

## Einsatz

Textanalysen

## Methoden

Wörter oder Phrasen aus dem  
Vokabular werden auf Vektoren  
mit reellen Zahlen abgebildet





Beispielsatz

In principio erat verbum et verbum erat  
apud deum et deus erat verbum

Auftrag

Definiere die Bedeutung eines jeden Worts durch seinen linken und rechten **Nachbarn**.

Schritt 1

Weise jeder einzigartigen Wortform einen Identifikator zu.

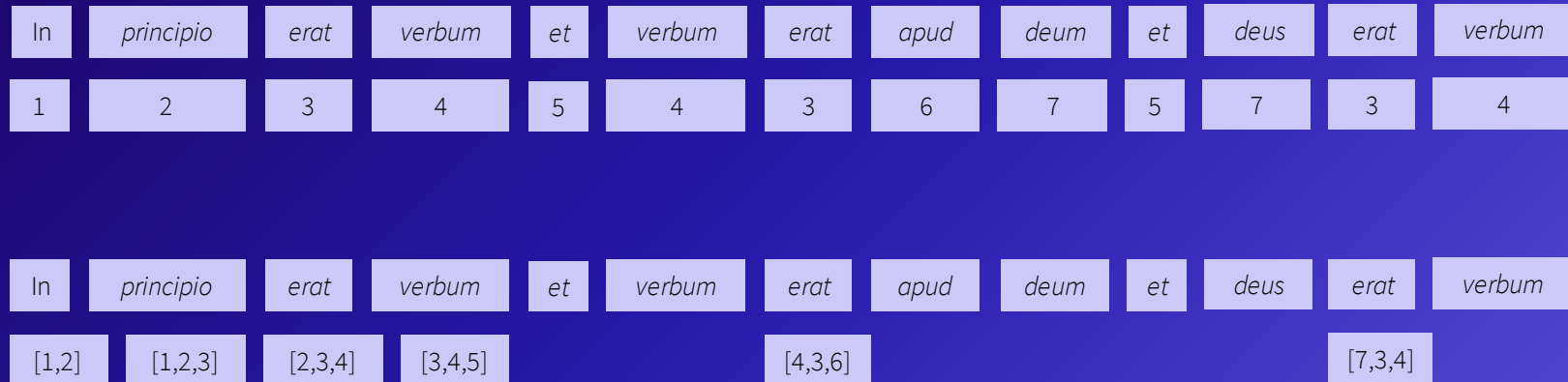
In	<i>principio</i>	<i>erat</i>	<i>verbum</i>	<i>et</i>	<i>verbum</i>	<i>erat</i>	<i>apud</i>	<i>deum</i>	<i>et</i>	<i>deus</i>	<i>erat</i>	<i>verbum</i>
1	2	3	4	5	4	3	6	7	5	7	3	4

# Wort → Vektor

## Schritt 1

## Schritt 2

Repräsentiere jedes Wort als eine Sequenz von Identifikatoren.



# Wort → Vektor

## Schritt 2

### Schritt 3

Kombiniere mehrfache Kodierungen derselben Wortform.

In	<i>principio</i>	<i>erat</i>	<i>verbum</i>	<i>et</i>	<i>verbum</i>	<i>erat</i>	<i>apud</i>	<i>deum</i>	<i>et</i>	<i>deus</i>	<i>erat</i>	<i>verbum</i>
		[2,3,4]			[4,3,6]					[7,3,4]		
<i>erat</i>	=	([2,3,4]	+	[4,3,6]	+	[7,3,4])	/3	=	7,266			
<i>verbum</i>	=	([3,4,5]	+	[5,4,3]	+	[3,4])	/3	=	6,381			
<i>deus</i>	=	([6,7,5]	+	[5,7,3])			/2	=	9,799			

Anfrage

Was ist ein typischer Kontext von „*erat*“?

Antwort

*verbum*

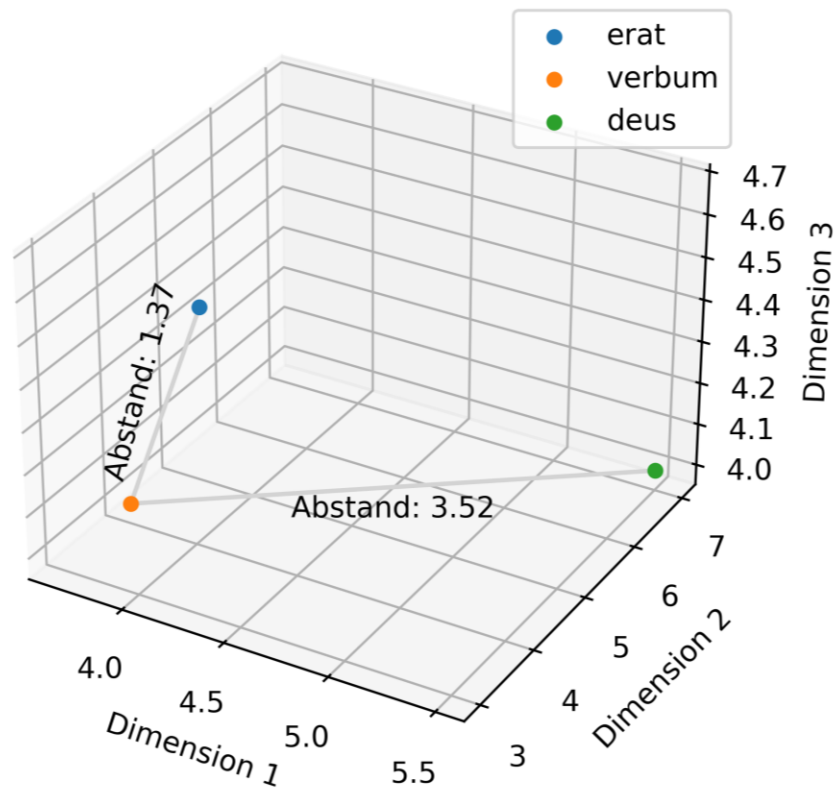
# Wort → Vektor

## Schritt 3



# Worte im Vektorraum

## 3-dimensionaler Vektorraum mit Euklidischem Abstand



# Vektoren und Semantik

- Distanz als Indiz für semantische Relationen
  - Stärke der Verbundenheit
  - Art der Relation (Synonym, Antonym ...)
- Nicht alle Wortarten gleichermaßen gut durch Vektoren abbildbar
- Rückschluss von Verteilung auf Funktion
- Mengentheoretische Zuordnungen: "Alle X sind Y"
- Globale vs. kontextualisierte Vektoren
- Qualität des Modells stark abhängig von der Konfiguration



Gries & Divjak 2009, Grefenstette & Sadrzadeh 2011, Baroni & Lenci 2011, Karan et al. 2012, Herbelot & Vecchi 2015, Rodda et al. 2019



dardalos  
— Digital Research for All —

Gefördert durch  
**DFG**  
Deutsche  
Forschungsgemeinschaft



# Forschungsfrage

Wie unterscheidet sich das Verständnis der griechischen Polis in der Grabrede aus Platons Menexenos von dem Verständnis in seinen Nomoi?

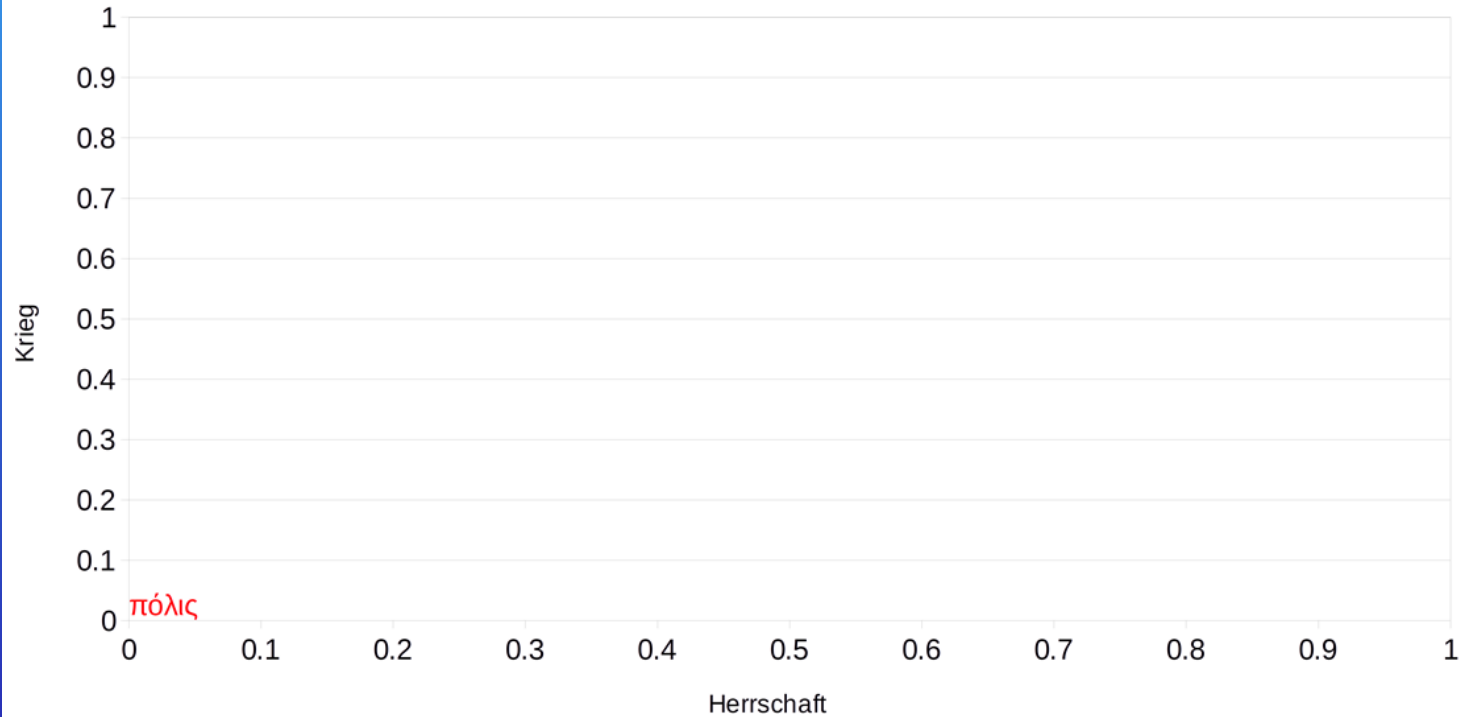


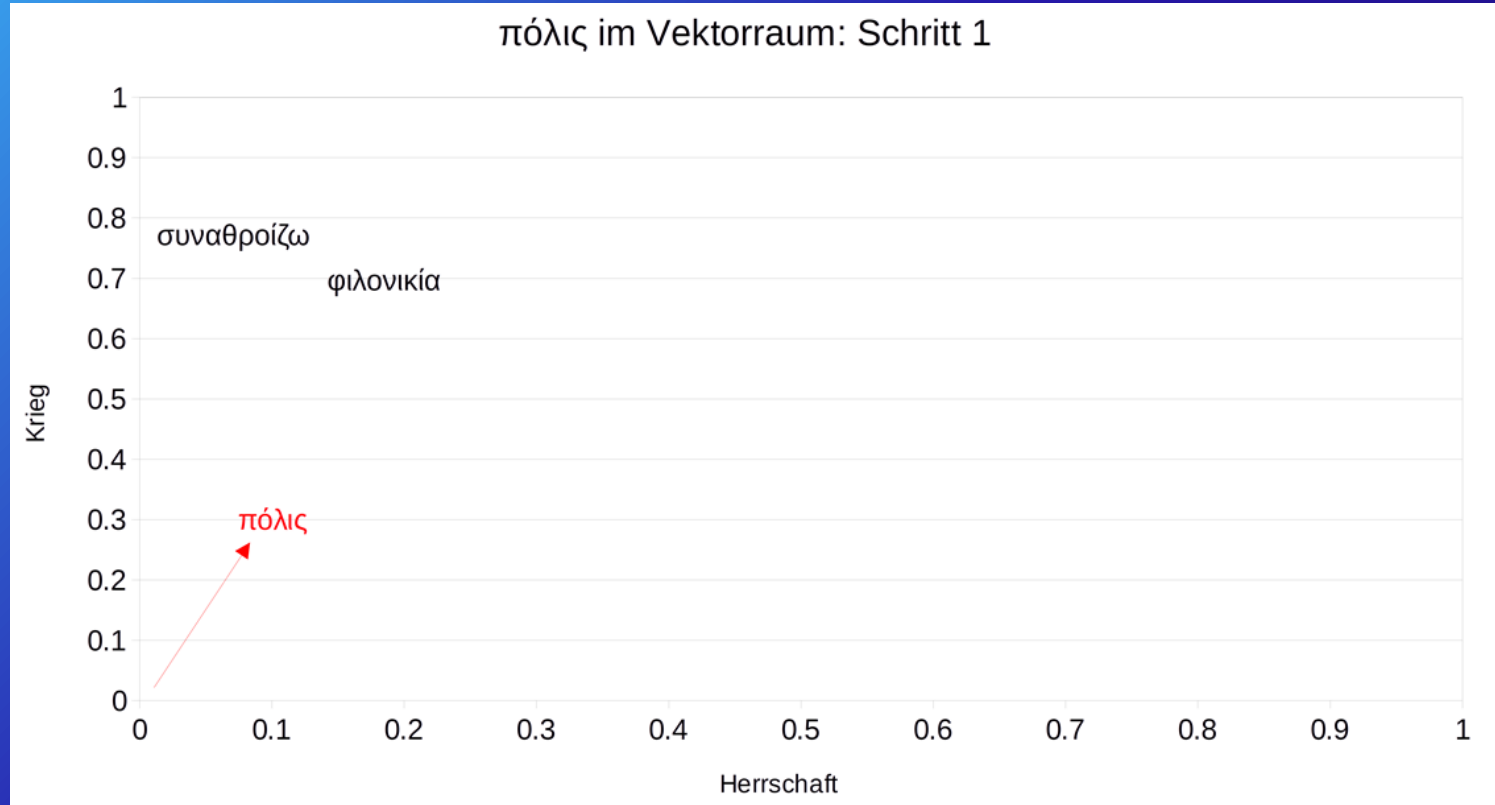
**daidalos**  
— Digital Research for All —



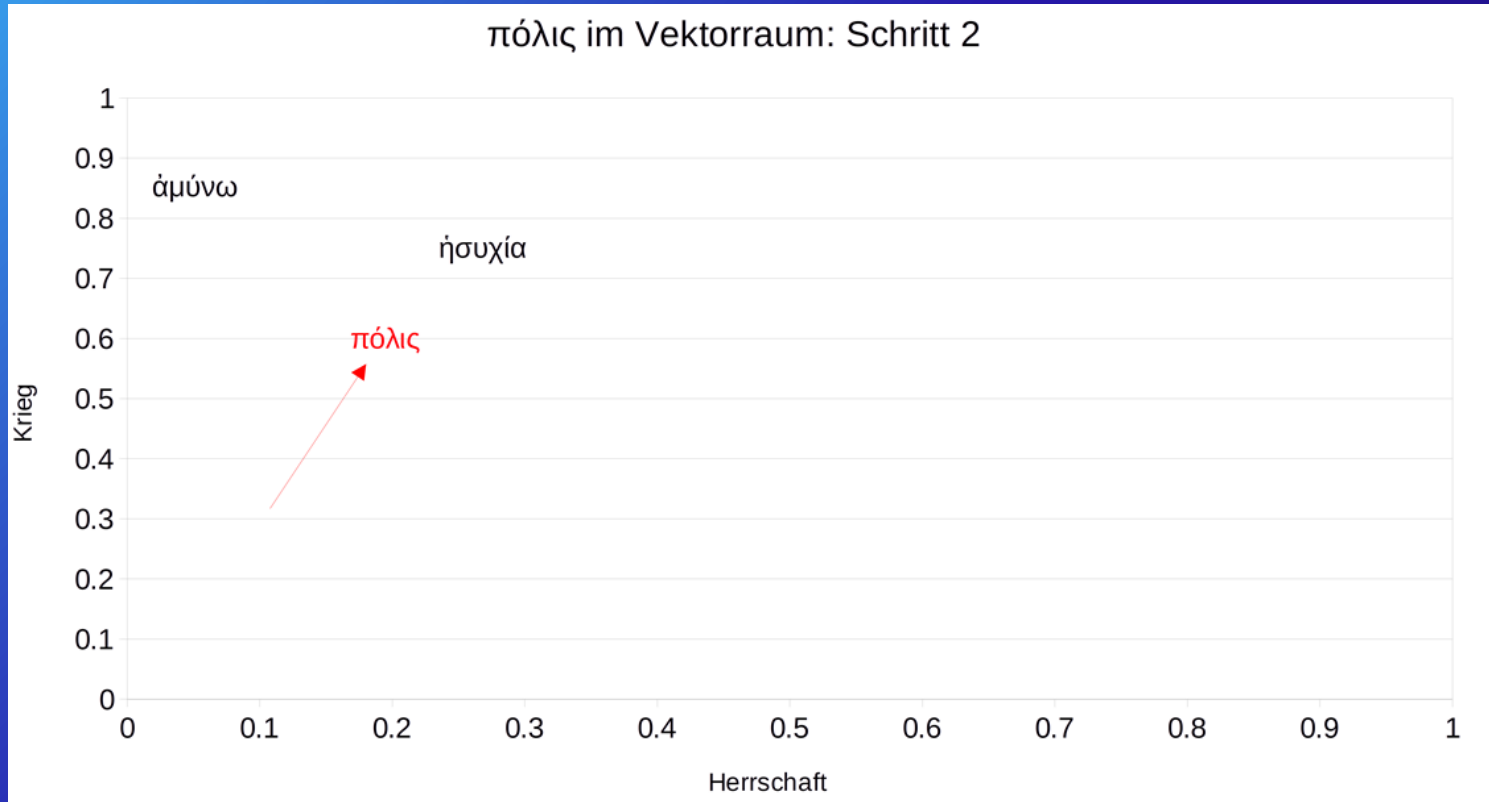
Gefördert durch  
**DFG**  
Deutsche  
Forschungsgemeinschaft

## πόλις im Vektorraum: Vor dem Training

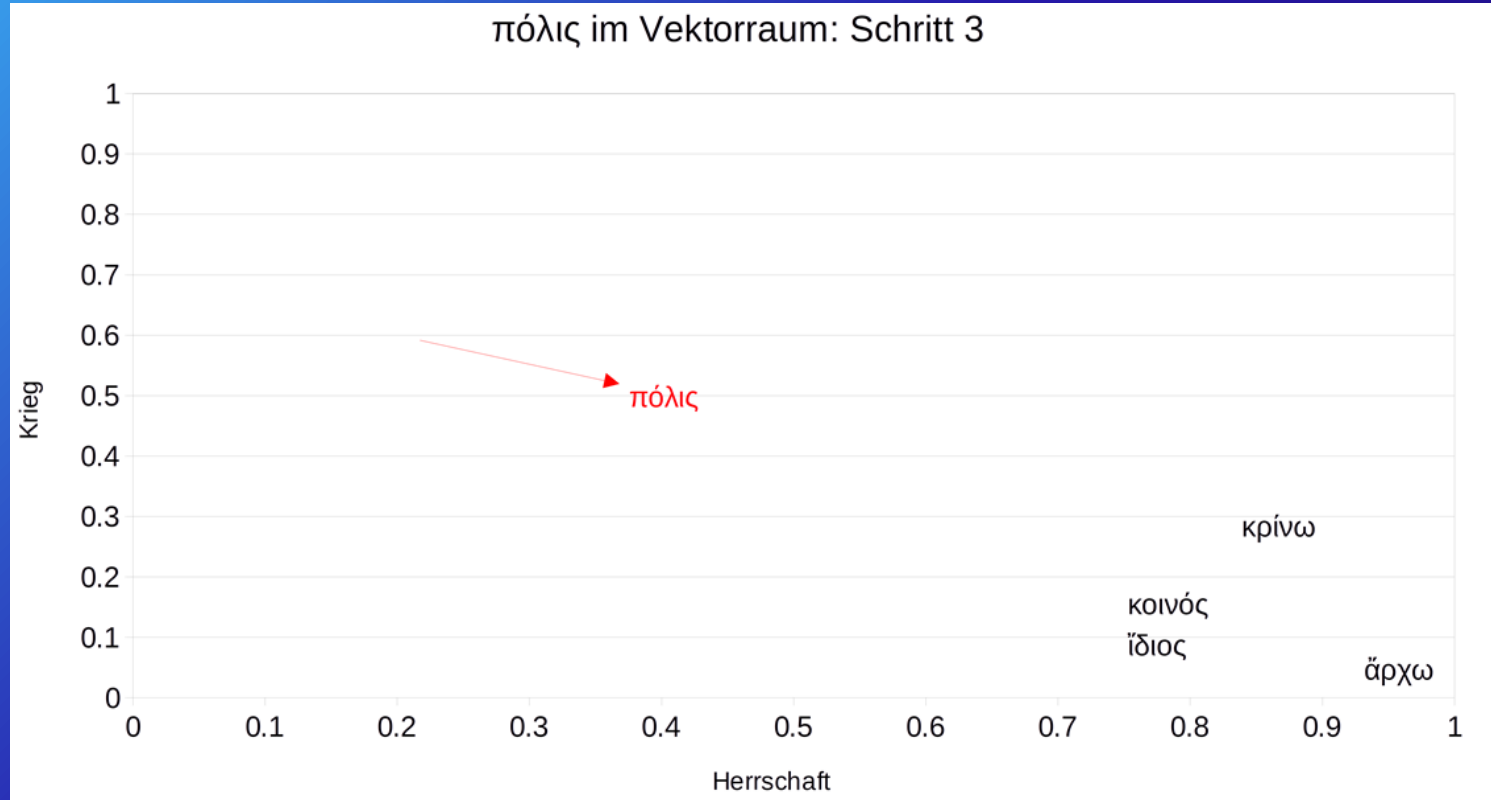




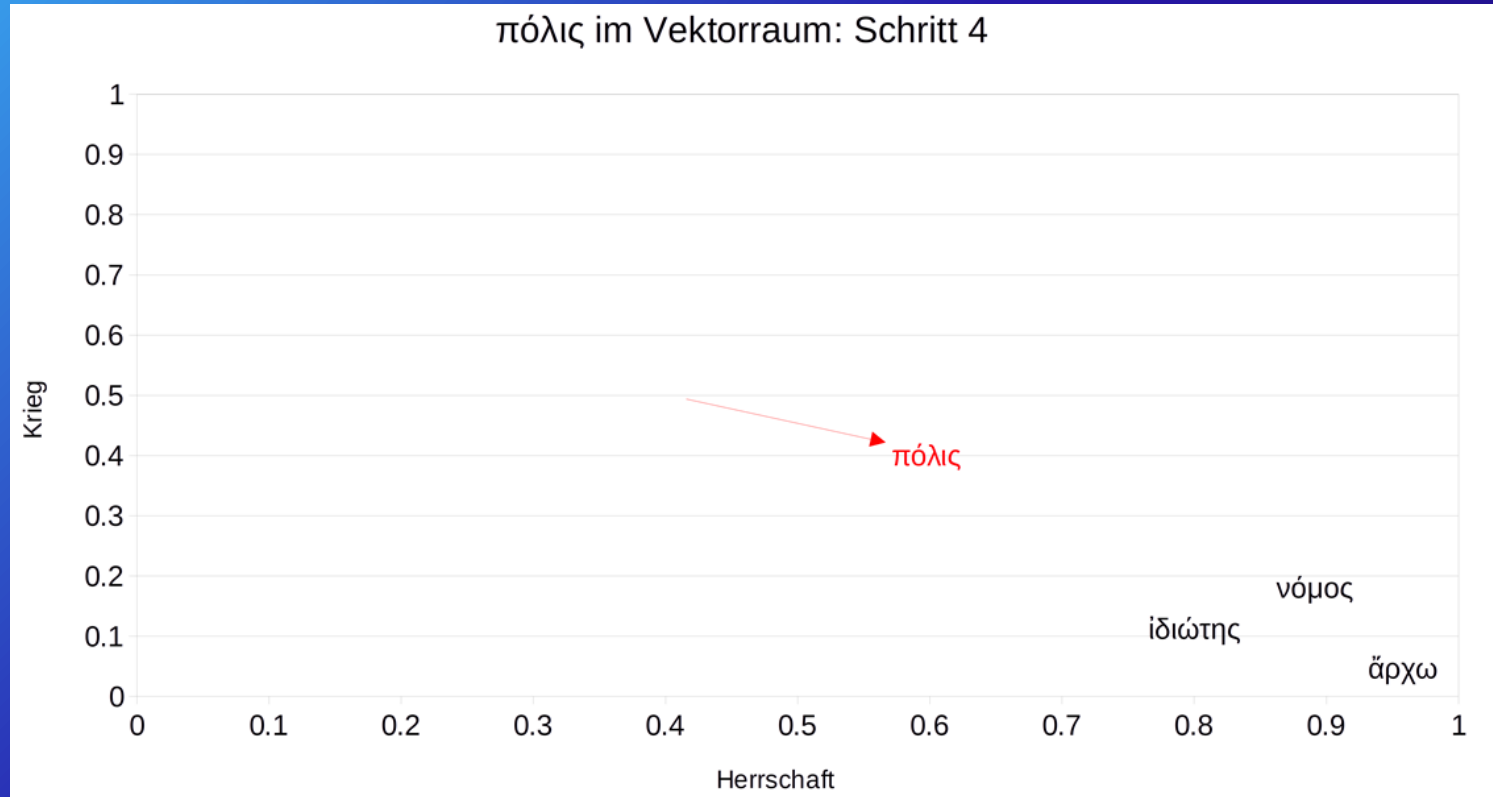
Plat. Μκ. 243b: τόδε λέγω τὸ εἰς τοσοῦτον **φιλονικίας** ἔλθεῖν πρὸς τὴν πόλιν τοὺς ἄλλους Ἕλληνας, ὥστε **συναθροῖσαι** ἐπὶ τὴν πόλιν πάντας Ἕλληνας τε καὶ βαρβάρους.



Plat. Μχ. 244b-c: μετὰ δὲ τοῦτο ἡσυχίαν ἦγεν ἡ πόλις, τοῖς μὲν βαρβάροις συγγινώσκουσα, ὅτι παθόντες ὑπ' αὐτῆς κακῶς ἰκανῶς οὐκ ἐνδεῶς ἡμύναντο.



Plat. leg. 946d: καὶ τὰ μὲν **ἴδια** ἕκαστος, τὰ δὲ καὶ **κοινῇ** μετ' ἀλλήλων **κρίναντες** τοὺς **ἄρξαντας** τῆ πόλει, ἀποφηνάντων ὅτι χρῆ παθεῖν ἢ ἀποτίνειν.



Plat. leg. 714a: εἰ δ' ἄνθρωπος εἷς ἄρξει δὴ πόλεως ἢ τινος **ιδιώτου** καταπατήσας ὁ τοιοῦτος τοὺς **νόμους**, ὁ νυνδὴ ἐλέγομεν, οὐκ ἔστι σωτηρίας μηχανή.

# πόλις im Vektorraum

## Zwischenergebnis

Spannungsfeld aus Krieg und  
Herrschaft

## Beobachtung

Wörter im Kontext beeinflussen  
die Position von πόλις

## Erkenntnis

Jeder weitere Satz führt zu einer  
weiteren Bewegung im  
Vektorraum.



# Methodendiskussion

Beispiel Wortanalogien:  
vir = fortis und mulier = ???



— **daidalos**  
— Digital Research for All —



Gefördert durch  
**DFG**  
Deutsche  
Forschungsgemeinschaft

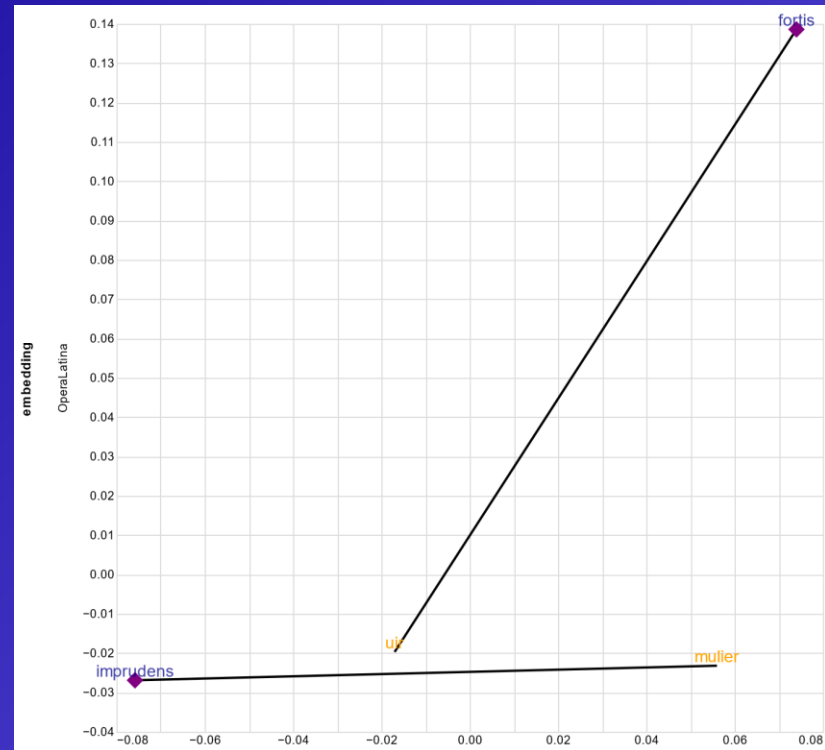


# Erwartung: *vir* = *fortis* und *mulier* = ???

Mann + positive  
Eigenschaft  
sowie  
Frau + positive  
Eigenschaft

→ erwartetes Ergebnis  
z.B. *pu-dicus*

→ Ergebnis: *im-prudens*



# Zusammenfassung

- Erwartetes vs. vektorbasiertes Ergebnis
    - a. Unterschiedliche Erschließungsmethoden
    - b. Historisches Wissen vs. Berechnung der Distanzen/Vektoren
  - Datenbias und User-Bias
  - deterministisch reproduzierbar: Visualisierung vs. Embeddings
  - Einfluss von Morphologie auf vektorbasierte Analogien: *fortis* vs. *imprudens* >> Adjektiv, oft attributiv
  - Analogien fallspezifisch, nur zwischen kohärenten Wortklassen
  - nicht alle Typen von linguistischen Analogien gleichermaßen abgebildet
- Wert und Mehrwert
- neue Perspektiven auf Bekanntes
  - Bearbeitung umfangreicher Korpora mit gleichbleibender Genauigkeit

# 04 | Embeddings in der Klassischen Philologie



— daidalos —  
— Digital Research for All —

Gefördert durch  
**DFG**  
Deutsche  
Forschungsgemeinschaft

# Beispiel 1: Autorschaft

Wer ist der Autor des *Commentariolum petitionis*, das Q. Cicero zugeschrieben wird?

NLP-Methoden: Text Classification, Part-of-Speech Tagging, **Word Embeddings**

Vainio et al. 2019: *Reconsidering Authorship in the Ciceronian Corpus through Computational Authorship Attribution*.

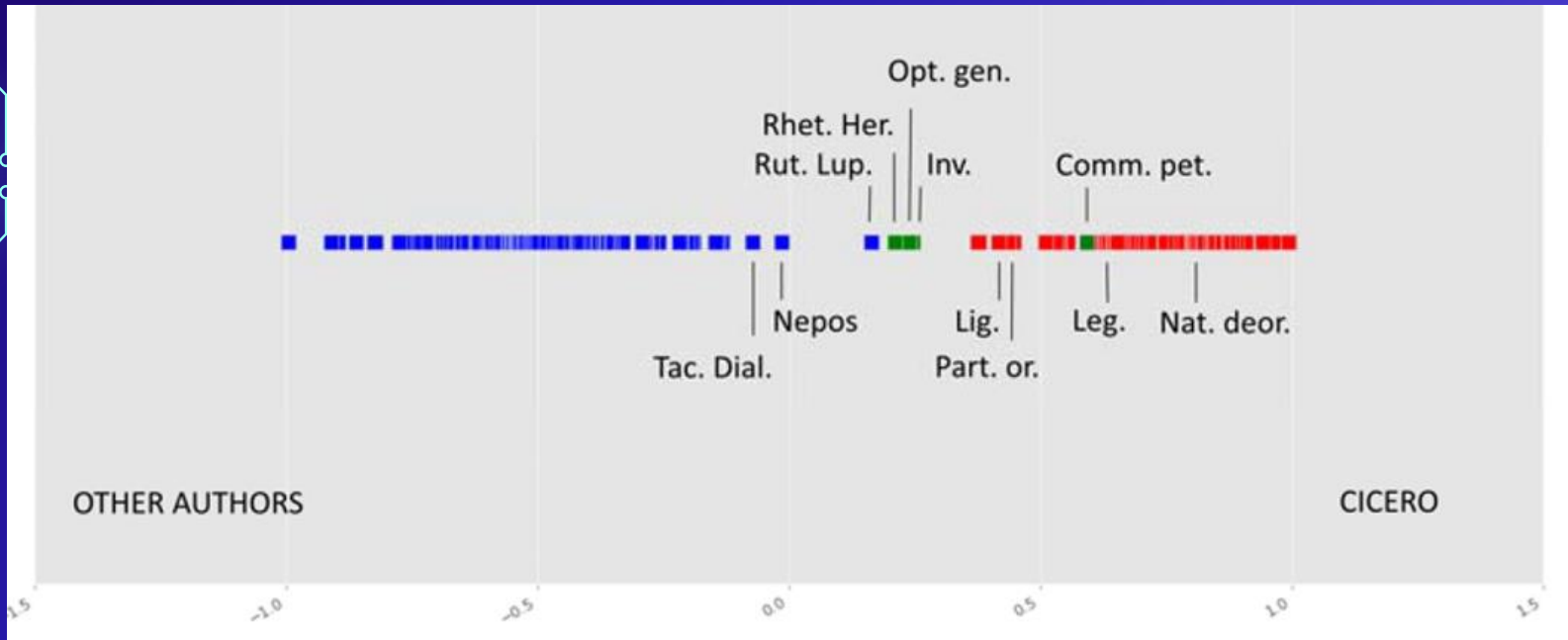


— Digital Research for All —



Gefördert durch  
Deutsche  
Forschungsgemeinschaft

# *Commentariolum* typisch für Stil von M. Cicero



Texte werden auf einer Skala nach ihrem Klassifizierungswert dargestellt: Positive Werte weisen auf eine ciceronische und negative auf eine nicht-ciceronische Autorschaft hin (Null markiert die Schwelle). Arbeit mit zwei Korpora (Cicero-Korpus + Referenzkorpus)

A decorative graphic on the left side of the slide, consisting of white and light blue lines and dots resembling a circuit board or digital data flow.

# Ergebnis

„Commentariolum petitionis **was written**, at least largely, by Marcus Cicero. It is possible that his brother Quintus participated in writing it, but we do not consider any major input by him probable, since **it would be unlikely that the brothers had such a uniform style as that in the text**, and since our analyses very strongly indicate a Ciceronian authorship.” (Vainio et al. 2019, 37)

# Beispiel 2: Analyse eines Konzepts

Was kann man über das Konzept Schmerz in der klassischen griech. Antike mittels maschineller Wortfeldanalysen erfahren?

NLP-Methoden: Lemmatisierung (Lexikon-basiert), **Word Embeddings**, Methodenmix mit manueller Wortklassifikation

Linka, V. & V. Kaše, 2023: Pain in Classical Greek Texts.

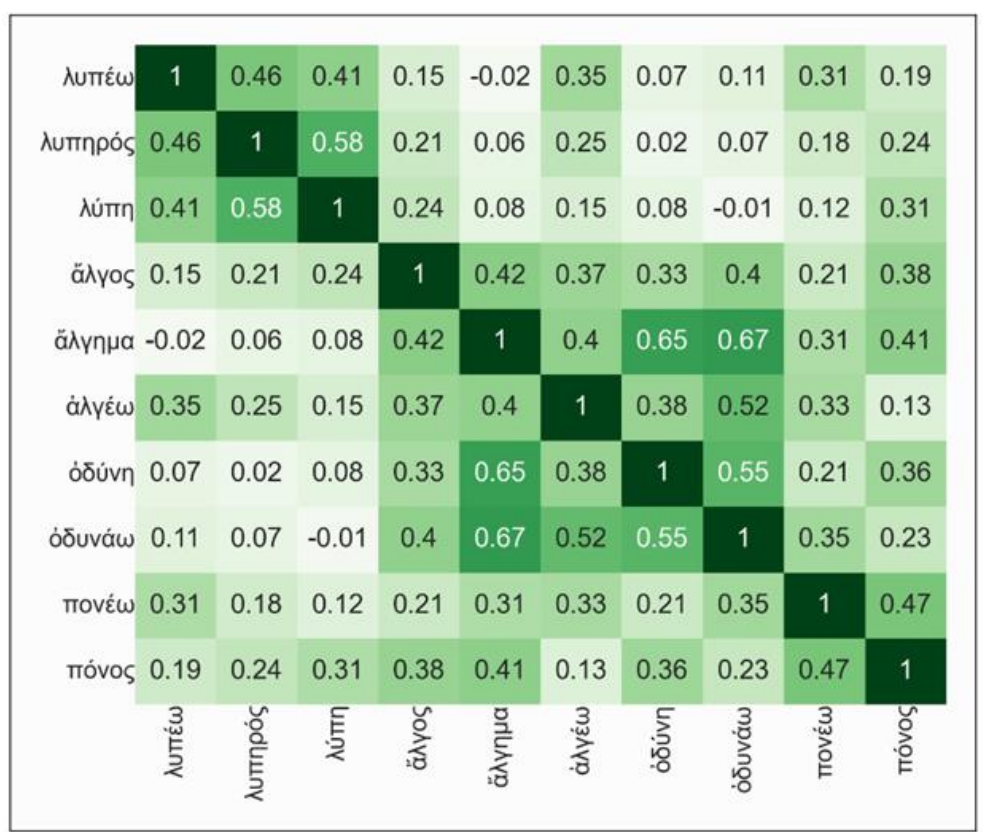


**dardalos**  
— Digital Research for All —

Gefördert durch  
**DFG**  
Deutsche  
Forschungsgemeinschaft



# Wortfamilien für „Schmerz“ entdecken

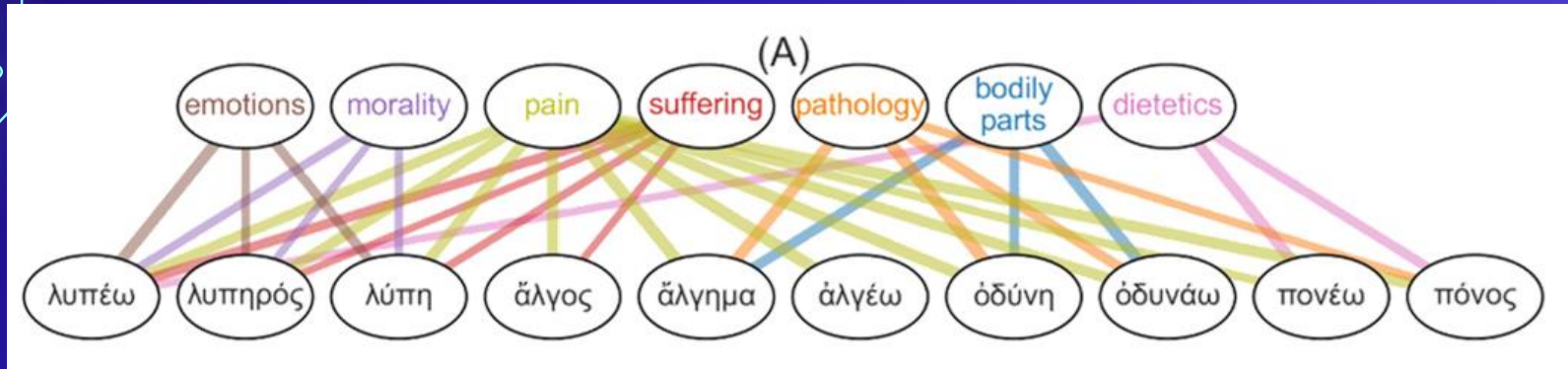


Wert 1 = identisch  
 Positive Werte = ähnlich  
 Negative Werte = unähnlich

Heatmap: Je dunkler das Grün, desto ähnlicher sind sich zwei Wörter.



# „Schmerz“ nach Kategorien



Assoziation zwischen Schmerzwörtern und ausgewählten semantischen Kategorien, berechnet mittels Kosinus-Ähnlichkeit, die ein Maß für die Ähnlichkeit zweier Vektoren ist. Stärkere Linien bedeuten ein höheres Ausmaß an Ähnlichkeit.



# Ergebnis

„Although all four pain word families were used for denoting pain, we have shown that their meanings differ significantly according to the features of pain that they associate with. [...] Then we saw how particular pain word families differ in their associations to semantic categories: λυπ\* is close to emotions and morality, ἀλγ\* and ὀδυν\* to bodily parts and pathologies, πον\* to dietetics.” (Linka & Kaše, 2023, 12)



# Beispiel 3: Intertextualität

Wie kann man in lateinischen Texten Anspielungen auf altgriechische Literatur maschinell finden?

NLP-Methoden: Multilingual **Sentence Embeddings**, Machine Translation

Riemenschneider, F. & A. Frank, 2023: *Graecia capta ferum victorem cepit. Detecting Latin Allusions to Ancient Greek Literature.*



dardalos  
— Digital Research for All —



Gefördert durch  
Deutsche  
Forschungsgemeinschaft

# Fallstudie: *Aeneis* und *Odyssee*



Entsprechende sprachübergreifende Begriffspaare wurden mit individuellen Farben markiert.

Query	Results
<p><i>Haec ubi dicta dedit, lacrimantem et multa volentem</i>  This speech uttered, while I wept and would have said many a thing.</p>	<p>τῆς δ' ἄρ' ἀκουούσῃς ῥέε δάκρυα, τήκετο δὲ χρῶς  and as she listened her tears flowed and her face melted  ὡς ῥάτο, τῆς δ' εὐνήσῃ γόον, σκέθε δ' ὄσσε γόοιο.  So she spoke, and lulled Penelope's laments, and made her eyes to cease from weeping.  ὡς ῥάτο, τῆ δ' ἄρα θυμὸν ἐνὶ στήθεσσιν ὄρνε.  So he spoke, and stirred the heart in her breast.</p>
<p><i>dicere deseruit, tenuisque recessit in auras.</i>  [...said], she left me and retreated into thin air.</p>	<p>ἡ μὲν ἄρ' ὡς ἔρξατο ἀπεβήκετο οἶα θεῶων.  Now when she had done this the fair goddess departed.  ἡ μὲν ἄρ' ὡς εἶποισα ἀπέβη πρὸς δόματα καλά,  So saying, she departed to the fair palace.  ἡ μὲν ἄρ' ἔς κρήνην κατεβήκετο καλλιρέεθρον  [She] had come down to the fair-flowing spring [Artacia].</p>
<p><i>Ter conatus ibi collo dare brachia circum:</i>  Thrice there was I fain to lay mine arms round her neck;</p>	<p>ὅππ' ἐν χερσὶν ἔλῳν, τὰ ῥά οἱ γέρε παρθεσαν αὐτῶ.  he took in his hands roast meat and set it before them, [...] which they had set before himself as a mess of honor.  τρίς μὲν μιν πελέμιζεν ἐρίσσεισθα μενεάων.  Thrice he made it quiver in his eagerness to draw it,  αὐτίκ' ἔπειτα τρίαναν ἔλῳν χερσὶ σίβηρσιν  straightway took his trident in his mighty hands,</p>
<p><i>ter frustra comprehensa manus effugit imago.</i>  thrice the vision I vainly clasped fled out of my hands.</p>	<p>τρίς δὲ μοι ἐκ χειρῶν σκιῇ εἰσελὼν ἦ καὶ ἄνετρον  and thrice [she flitted] from my arms like a shadow or a dream,  τρίς μὲν ἐρωαμῆθην, ἔλεσεν τέ με θυμὸς ἀνώγει,  Thrice I sprang towards her, and my heart bade me clasp her,  χερσὶ δὲ μή τι λίην προκαλίξω, μή με χολώσῃς,  But with thy hands do not provoke me overmuch,</p>
<p><i>par levibus ventis volucrique simillima somno.</i>  even as the light breezes, or most like to fluttering sleep.</p>	<p>ἡ δ' ἔθεεν βορρῇ ἀνέμῳ ἀκραεὶ καλῶ,  And she ran before the North Wind, blowing fresh and fair,  ἄρσας ἀργαλέων ἀνέμων ἀμέγαρτον αὐτιμν.  when he had roused a furious blast of cruel winds  ἐς πνοῖας ἀνέμων, ἡ δ' ἔξ ὕπνου ἀνόρουσε  into the breath of the winds. And [she] started up from sleep</p>



# Ergebnis

„[...] our case study demonstrates the proficiency of our models in recognizing sentence structures and translating them to a different language [...], and in identifying common topics or concepts across languages, even locating verses where multiple relevant concepts exist within the same verse [...], our SPHILBERTA model can serve as a useful tool for automatic first-pass exploration of potential cross-lingual intertextual allusions, [...].” (Riemenschneider & Frank 2023, 8)



# Bibliographie

- Baroni, Marco, and Alessandro Lenci. 2011. "How We BLESSed Distributional Semantic Evaluation." In Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, 1–10. Association for Computational Linguistics. <https://www.aclweb.org/anthology/W11-2501>
- Bengfort, B., T. Ojeda, & R. Bilbro, 2018: Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning. Boston: O'Reilly.
- Dale, R., 2010: Classical Approaches to Natural Language Processing. Handbook of Natural Language Processing. Hrsg. von N. Indurkha & F. J. Damerau. CRC Press, Taylor & Francis Group, 3–7.
- Forstall, C. & Scheirer, W. (2019): Quantitative Intertextuality. Cham.
- Gladkova, Anna, Aleksandr Drozd, and Satoshi Matsuoka. 2016. "Analogy-Based Detection of Morphological and Semantic Relations with Word Embeddings: What Works and What Doesn't." In Proceedings of the NAACL Student Research Workshop, 8–15. <https://www.aclweb.org/anthology/N16-2002>
- Grefenstette, Edward, and Mehrnoosh Sadzadeh. 2011. "Experimental Support for a Categorical Compositional Distributional Model of Meaning." arXiv Preprint arXiv:1106.4058. <https://arxiv.org/pdf/1106.4058.pdf>
- Gries, Stefan Th, and Dagmar Divjak. 2009. "Behavioral Profiles: A Corpus-Based Approach to Cognitive Semantic Analysis." New Directions in Cognitive Linguistics, 57–75. <https://pdfs.semanticscholar.org/bedf/34474970f22b87605659f939732cbf1b2b4b.pdf>
- Herbelot, Aurélie, and Eva Maria Vecchi. 2015. "Building a Shared World: Mapping Distributional to Model-Theoretic Semantic Spaces." In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 22–32. <https://www.aclweb.org/anthology/D15-1003>
- Karan, Mladen, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. "Distributional Semantics Approach to Detecting Synonyms in Croatian Language." Information Society, 111–16. <https://pdfs.semanticscholar.org/acc4/70ec10cddc2aa720fd2c02407332ddcf2fbf.pdf>
- Köntges, Th., 2020: Measuring Philosophy in the First Thousand Years of Greek Literature. Digital Classics Online 1–23. <https://doi.org/10.11588/dco.2020.2.73197>
- Linka, V. & V. Kaše, 2023: Pain in Classical Greek Texts. Digital Classics Online 1–14. <https://journals.ub.uni-heidelberg.de/index.php/dco/article/view/93792/90156>
- Martinelli, Ginevra, Paola Impicciché, Elisabetta Fersini, Francesco Mambrini, and Marco Passarotti. 2024. "Exploring Neural Topic Modeling on a Classical Latin Corpus." In, 6929–34
- Paviopoulos, J., Xenos, A., & Picca, D. (2022). Sentiment Analysis of Homeric Text: The 1st Book of Iliad. 7071–7077. <https://aclanthology.org/2022.lrec-1.765.pdf>
- Rivemenschneider, Frederick & Anette Frank. 2023: Graecia capta ferum victorem cepit. Detecting Latin Allusions to Ancient Greek Literature. <https://arxiv.org/pdf/2308.12008.pdf>
- Rodda, M, Philomen Probert, and Barbara McGillivray. 2019. "Vector Space Models of Ancient Greek Word Meaning, and a Case Study on Homer." Traitement Automatique Des Langues 60 (3)
- Rogers, Anna, Aleksandr Drozd, and Bofang Li. 2017. "The (Too Many) Problems of Analogical Reasoning with Word Vectors." In Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (\*SEM 2017), 135–48. <https://www.aclweb.org/anthology/S17-1017>
- Stüssi, E., & Ströbel, P. B. (2024). Part-of-Speech Tagging of 16th-Century Latin with GPT. Proceedings of LaTeCH-CLfL 2024, 196–206. <https://aclanthology.org/2024.latechclfl-1.18.pdf>
- Vainio, Raija, Reima Välimäki, Alekski Vesanto, Anni Hella, Marjo Kaartinen, & Teemu Immonen. 2019: Reconsidering Authorship in the Ciceronian Corpus through Computational Authorship Attribution. Ciceroniana online 3:15–48. <https://ojs.unito.it/index.php/COL/article/view/3518/3182>