SunoikisisDC 2021
Session 9

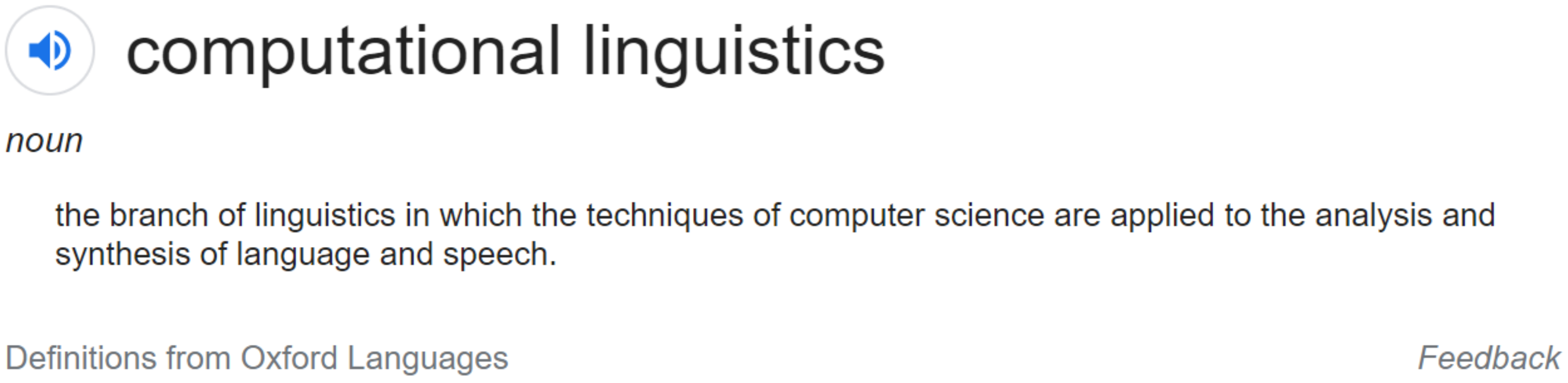
Computational Linguistics

Alek Keersmaekers (KU Leuven)
Martina Astrid Rodda (Oxford)

Classical Computational Linguistics: an introduction

What is computational linguistics?

- In a broad sense: any linguistic work with a computational methodology
- Computational analysis of linguistic data (i.e. corpus linguistics)
- Automated processing and/or annotation of linguistic data (i.e. natural language processing)



computational linguistics

noun

the branch of linguistics in which the techniques of computer science are applied to the analysis and synthesis of language and speech.

Definitions from Oxford Languages

Feedback

Corpus compilation (1)

- Most of the work in classical computational linguistics: focused on the compilation of corpus resources
- Full-text databases (see also *https://wiki.digitalclassicist.org/Greek_and_Latin_texts_in_digital_form*):
  - *TLG, Perseus* (literary texts)
  - *Packard Humanities Institute* (inscriptions)
  - *Duke Databank of Documentary Papyri* (papyri)
- **Automatically** annotated texts (mostly lemmas + morphology): e.g. *Perseus under Philologic*, *Diorisis, LatinISE, GLAUx*
- **Manually** annotated texts (treebanks: lemmas, morphology, syntax, sometimes semantics): see next slide
- Has led to several corpus-based studies (even though the classical languages are still lagging behind: see Jenset & McGillivray 2017)

Corpus compilation (2)

Greek project	Tokens	Texts
AGDT	560K	Archaic + Classical poetry/prose
Gorman	546K	Classical + post-classical prose
Pedalion	333K	Classical + post-classical poetry/prose
PROIEL	270K	Herodotus, NT, Sphrantzes
PapyGreek	42K	Papyri
Harrington	18K	Lucian, Septuagint, Life of Aesop
Aphthonius	7K	Aphthonius

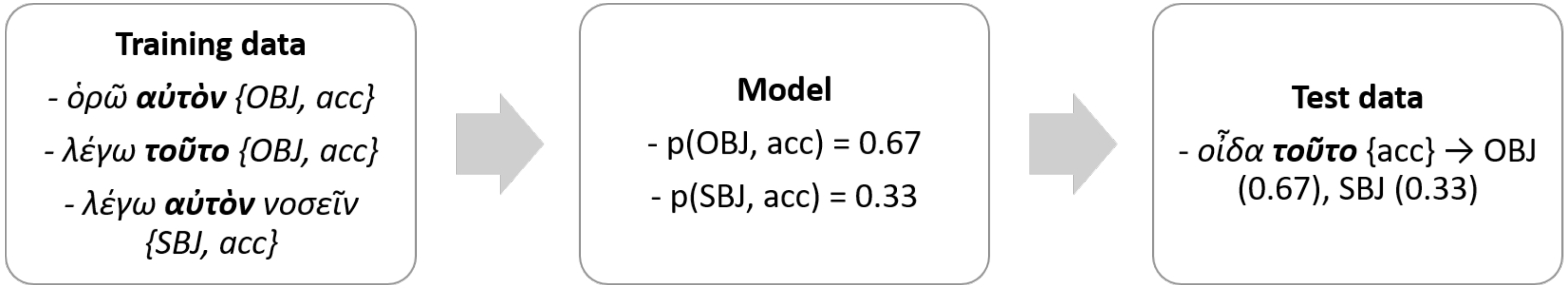
Latin project	Tokens	Texts
Index Thomisticus	450K	Thomas Aquinas
PROIEL	225K	Classical + post-classical prose
LLCT	200K	Medieval charter texts
Harrington	120K	Classical prose/poetry
ALDT	53K	Classical prose/poetry

NLP and Classical languages

- … in contrast, natural language processing of classical languages is still in its infancy (large amounts of data required)
- Natural language can be processed in several ways:
  - Morphological processing and lemmatization (McGillivray & Kilgariff 2013; Celano, Crane & Majidi 2016)
  - Syntactic parsing (e.g. Mambrini & Passarotti 2012; Ponti & Passarotti 2016)
  - Named Entity Recognition (e.g. Erdmann et al. 2016; Palladino, Karimi & Mathiak 2020)
  - Distributional lexical semantics (e.g. Rodda, Senaldi, and Lenci 2016; Bamman and Burns 2020)
  - …
- See the case studies in this session for more examples
- See also the Classical Language Toolkit (http://cltk.org/, see also https://github.com/SunoikisisDC/SunoikisisDC-2017-2018/wiki/The-Classical-Language-Toolkit-(CLTK) and https://www.digitalclassicist.org/wip/wip2018-09pb.html)

Machine learning: main concepts (1)

- Older computational work (for English): rule-based
- Most methods nowadays: machine learning
- **Supervised** machine learning:
    - To predict a certain class (e.g. OBJ vs. SBJ)
    - … one starts from a large corpus annotated for a number of **features** (e.g. case), the **training data**
    - The computer 'learns' patterns from the data and formalizes them in a mathematical **model**
    - This model can be used to predict class labels for new, unseen data, the **test data**



Training data

- *ὁρῶ **αὐτὸν** {OBJ, acc}*
- *λέγω **τοῦτο** {OBJ, acc}*
- *λέγω **αὐτὸν** νοσεῖν {SBJ, acc}*

Model

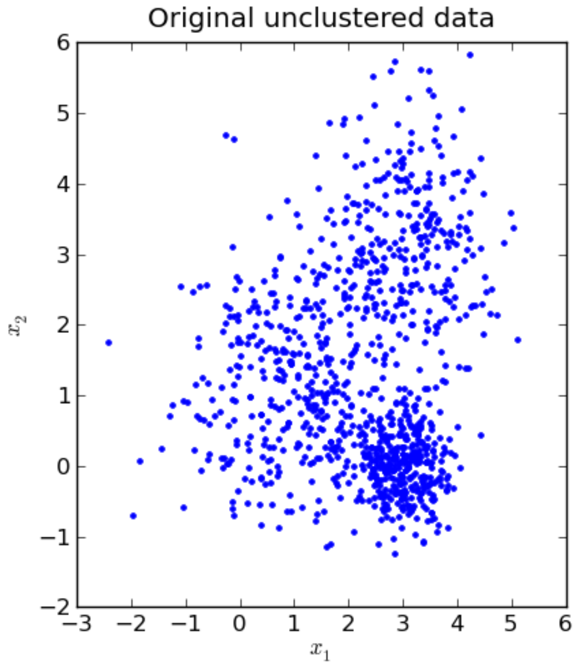
- p(OBJ, acc) = 0.67
- p(SBJ, acc) = 0.33

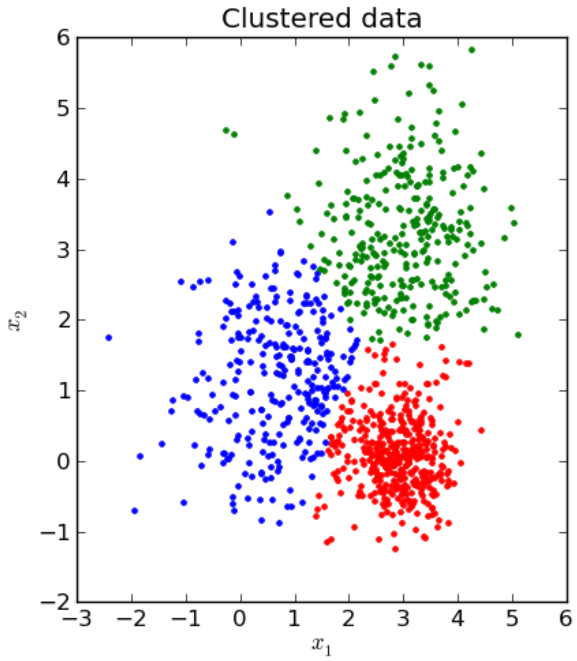
Test data

- *οἶδα **τοῦτο*** {acc} → OBJ (0.67), SBJ (0.33)

Machine learning: main concepts (2)

- **Unsupervised** machine learning (e.g. clustering): no pre-defined class labels
- Instead, the computer divides the data into groups with internally similar members





Classical computational linguistics: main challenges

- Typological characteristics of Greek and Latin (most NLP methods tailored to English): e.g. highly inflectional, free word order, large rate of ellipsis
- Large genre and diachronic variation
- Data sparseness

References

- Bamman, D. & Burns, P. J. 2020. Latin BERT: A Contextual Language Model for Classical Philology. *arXiv Preprint arXiv:2009.10053*.
- Celano, G. G. A., Crane, G. & Majidi, S. 2016. Part of Speech Tagging for Ancient Greek. *Open Linguistics* 2(1).
- Erdmann, A., Brown, C., Joseph, B., … Marneffe, M.-C. de. 2016. Challenges and Solutions for Latin Named Entity Recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. Osaka: The COLING 2016 Organizing Committee, pp. 85–93.
- Jenset, G. B. & McGillivray, B. 2017. *Quantitative historical linguistics: a corpus framework*. Oxford: Oxford University Press.
- Mambrini, F. & Passarotti, M. C. 2012. Will a parser overtake Achilles? First experiments on parsing the ancient Greek dependency treebank. In *Eleventh International Workshop on Treebanks and Linguistic Theories*. Edições Colibri, pp. 133–144.
- McGillivray, B. & Kilgarriff, A. 2013. Tools for historical corpus research, and a corpus of Latin. In Durrell, P., Scheible, M., Whitt, S. & Bennett, R. J. (eds.), *New Methods in Historical Corpus Linguistics*. pp. 247–257.
- Palladino, C., Karimi, F. & Mathiak, B. 2020. NER on Ancient Greek with minimal annotation. In *Digital Humanities 2020*. Ottawa: DH2020.
- Ponti, E. M. & Passarotti, M. 2016. Differentia compositionem facit. A Slower-Paced and Reliable Parser for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. Portorož: European Language Resources Association (ELRA), pp. 683–688.
- Rodda, M. A., Senaldi, M. S. & Lenci, A. 2016. Panta Rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek. In *CLiC-it/EVALITA*.

Distributional Semantics and Homeric formulae

Dr Martina Astrid Rodda | they/them/theirs | University of Oxford

University College | martinaastrid.rodda@classics.ox.ac.uk

Flexible formulae?

The issue of formulaic variation

Oral vs. written poetry:

(1) Traditional style
vs. individual expression

(2) Repetition vs. variation

Lord, Albert Bates. 2000. *The Singer of Tales*. 2nd ed. Cambridge, MA; London: Harvard University Press.

Approaching formulaic variation

- Two axes of variation:

  (1) Syntax → repeated structures or flexible units (in performance)?

  (2) Semantics → what holds a formula together?

- Quantitative approach:

  Jenset, Gard B., and Barbara McGillivray. 2017. *Quantitative Historical Linguistics: A Corpus Framework*. Oxford: Oxford University Press.

Some thoughts on method

Comparison with a baseline

Formulaic material only

Syntactic variation: yay!

- Formulae vs non-formulaic phrases: very different behaviour
- Formulae are used flexibly, not memorised by rote
- Patterns in formulaic flexibility ~ patterns in idiom flexibility

Semantic variation?

Expectations

- Constructional/idiom pre-emption
- Formular economy

  → formulaic material shows **more** variety in meaning?
- NB both of these are **tendencies**, not laws

***insert citation

Quantifying meaning

Distributional Semantics:

- Operationalising meaning as a function of lexical context
- Vector space model(s)

…he had the wreath in his hands…
…a laurel wreath with ribbons…
…make elegant wreaths for festivals…

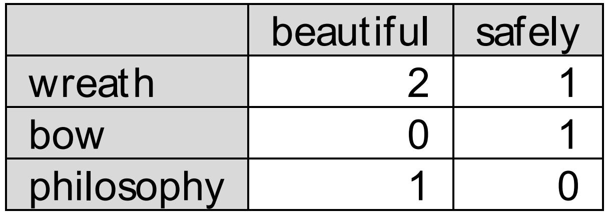
…he had the bow in his hands…
…Syrian curved bows on horseback…
…arrows from the bow at the suitors…

Harris, Zellig S. 1954. "Distributional structure", *Word* 10: 146–162.
Fabre, Cécile, and Alessandro Lenci. 2015. "Distributional Semantics today", *TAL – Traitement Automatique des Langues* 56: 7–20.

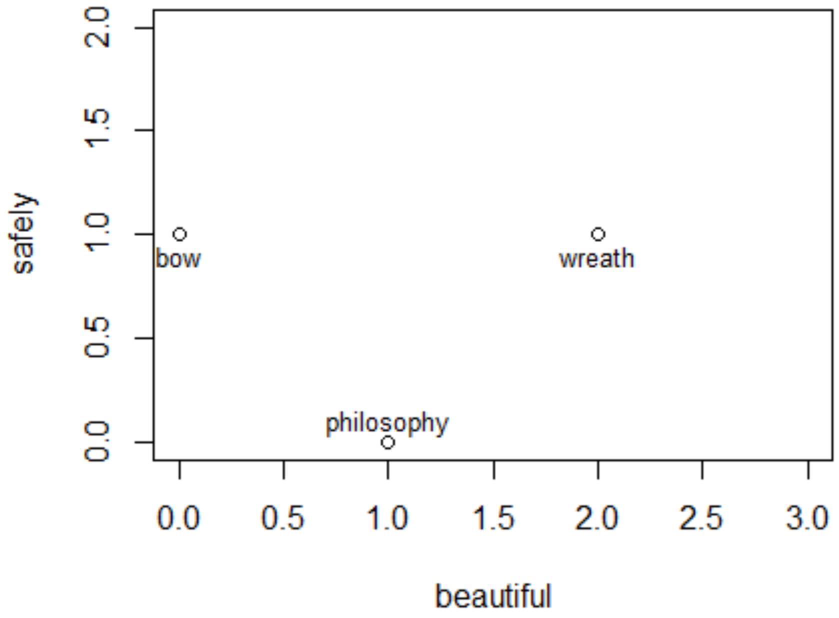
From co-occurrences to models

		Context words							
		beautiful	safely	sacred	laurel	curved	ride	belief	teach
Target words	wreath	2	1	1	1	0	0	0	0
	bow	0	1	0	0	2	1	0	0
	philosophy	1	0	0	0	0	0	1	1

From co-occurrences to models



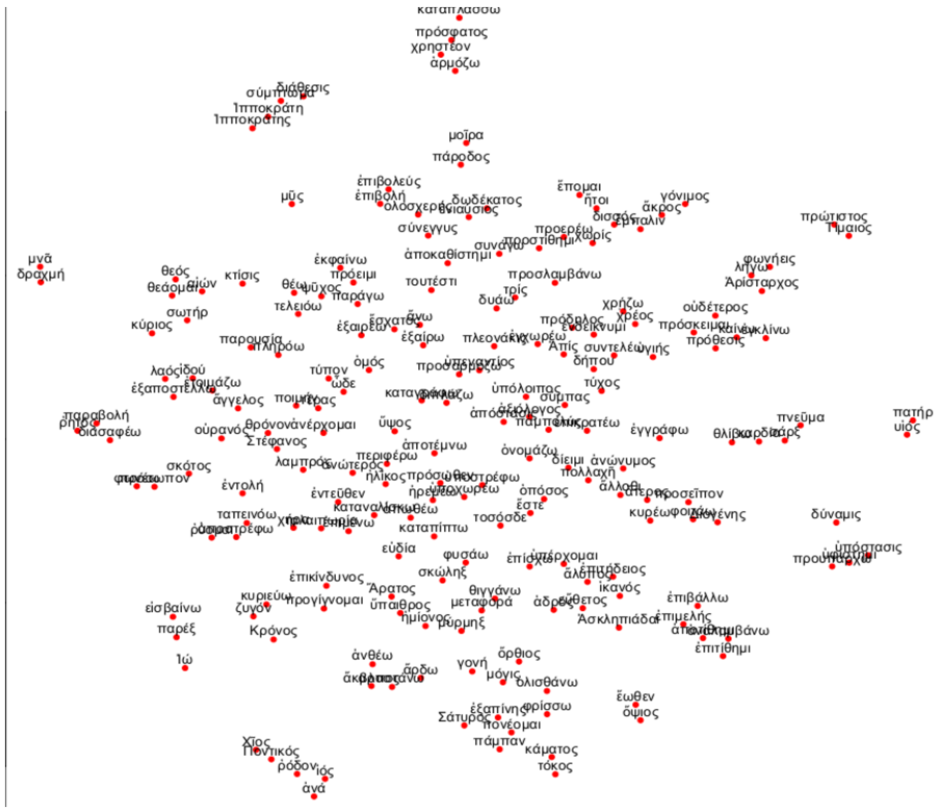
	beautiful	safely
wreath	2	1
bow	0	1
philosophy	1	0



Quantifying meaning

Distributional Semantics:

- Operationalising meaning as a function of lexical context
- Vector space model(s)



Rodda, Martina A., Marco S.G. Senaldi, and Alessandro Lenci. 2017. "Panta Rei: Tracking semantic change with Distributional Semantics in ancient Greek." *Italian Journal of Comp. Linguistics* 3 (1): 11–24.

Materials & methods

- Vector space model from the Diorisis corpus
- Evaluated for accuracy against ancient scholarship, modern lexicography, and an NLP resource (Ancient Greek WordNet)
- Parameter optimisation: frequency filtering, size of context window

Rodda, Martina Astrid, Philomen Probert, and Barbara McGillivray. forthc. 2020. "Vector Space Models of Ancient Greek Word Meaning, and a Case Study on Homer." *TAL – Traitement Automatique Des Langues* 60 (3).

A potential problem

Is our corpus **too small** for Distributional Semantics?

Preliminary results

- 5 transitive VP formulae of holding/thinking:
  - 3x with χείρ ‘hand’ (+ ἔχω, αἱρέω, τίθημι)
  - 1x intellectual perception (εὖ εἰδέναι)
  - 1x sensory perception (ἰδεῖν ὀφθαλμοῖσ/ι)
- Measure: variance of the distribution of object distances

→ best predictors: formula type (physical ~ metaphorical); syntactic flexibility (accusative vs. genitive objects)

The corpus is not too small!

Next step: complete analysis of TrV+Obj formulae

What I looked at …

- TrV + AccObj phrases (μῆνιν ἄειδε, ἄλγε᾽ ἔθηκε)
- Target texts: *Iliad*, *Odyssey*, *Theogony*, *Works and Days*
- Ancient Greek and Latin Dependency Treebank (AGLDT)

Bamman, David, and Gregory Crane. 2011. "The Ancient Greek and Latin Dependency Treebanks." In *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, 79–98. Berlin; Heidelberg: Springer.

What I looked at ...

- 15 formulaic verbs with >10 object types (91x to 11x)
- Formulaic material **only!**

Pavese, Carlo Odo, and Federico Boschetti. 2003. *A Complete Formular Analysis of the Homeric Poems*. 3 vols. Lexis Research Tools 2. Amsterdam: Hakkert.

Pavese, Carlo Odo, and Paolo Venti. 2000. *A Complete Formular Analysis of the Hesiodic Poems: Introduction and Formular Edition*. Lexis Research Tools 4. Amsterdam: Hakkert.

What I looked at ...

- Existing verb valency database
- Authors: Aeschylus, Aesop, Pseudo-Apollodorus, Athenaeus, Diodorus Siculus, Herodotus, Lysias, Plato, Plutarch, Polybius, Sophocles, Thucydides

McGillivray, Barbara, and Alessandro Vatri. 2015. “Computational Valency Lexica for Latin and Greek in Use: A Case Study of Syntactic Ambiguity.” *Journal of Latin Linguistics* 14: 101–26.

What I looked at …

- Cosine similarity
- Centroid of objects
- 30 distributions of distances
- Measures: median similarity & variance

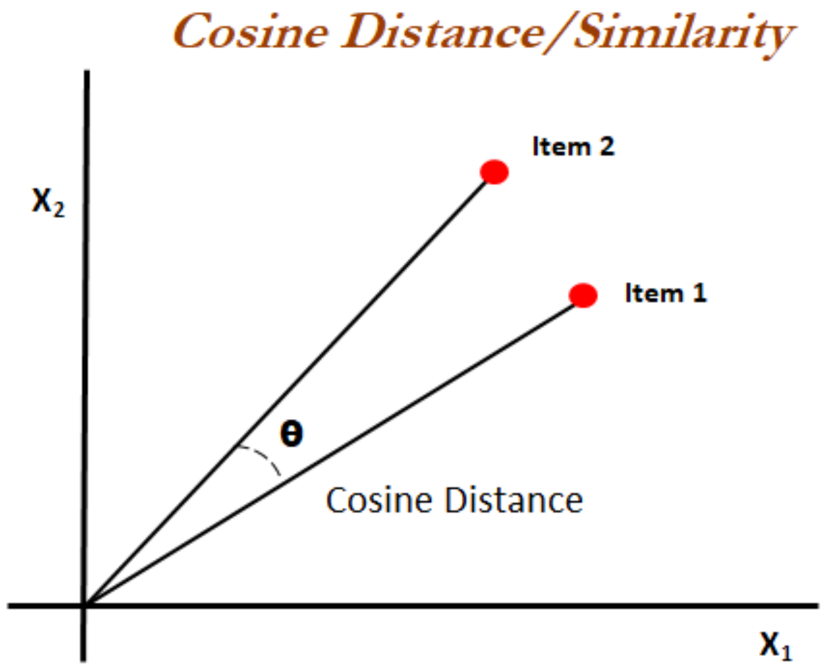
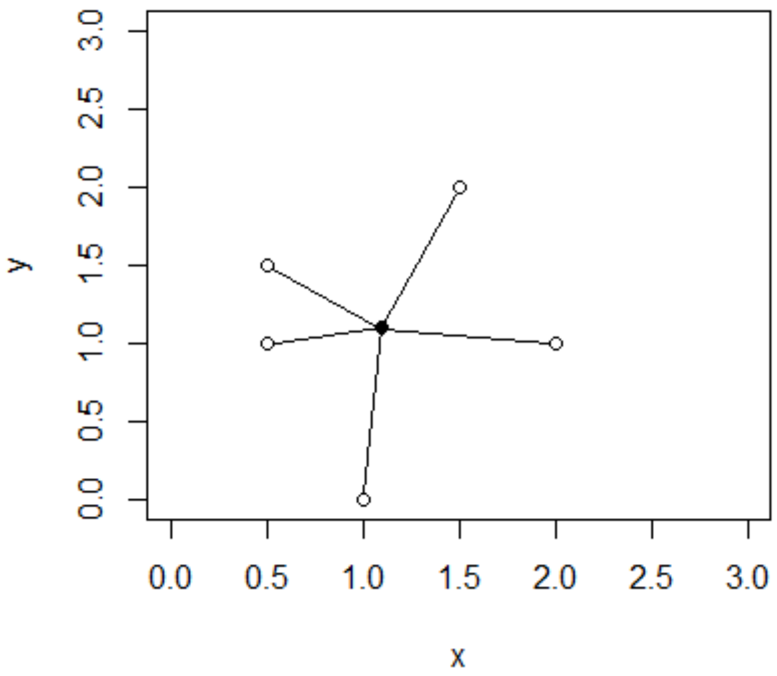


Image source: Dangeti, Pratap. 2017. *Statistics for Machine Learning*. Packt Pub.

What I looked at …

- Cosine similarity
- Centroid of objects
- 30 distributions of distances
- Measures: median similarity & variance



What I looked at …

- Cosine similarity
- Centroid of objects
- 30 distributions of distances
- Measures: median similarity & variance

ἔχω
αἱρέω
δίδωμι
εἶπον
φέρω
βάλλω
εἶδον
τίθημι
οἶδα
χέω
ἄγω
λύω
τίκτω
ἵημι
ἀκούω

… And what I found

- **No significant differences** between formulaic and non-formulaic pairs
- (No significant effects of frequency or productivity)
- … What happened to formular economy?

More things to look into

- New look at the data

→ are centroids the best similarity measure?

→ cluster analysis?

- Control for confounding effects

→ later hexameter poetry

- Maybe there is just no difference!

→ then why the concept of formular economy?

More things to look into

- New look at the data

→ are centroids the best similarity measure?

→ cluster analysis?

- Control for confounding effects

→ later hexameter poetry

- Maybe there is just no difference!

→ then why the concept of formular economy?

More things to look into

- New look at the data

→ are centroids the best similarity measure?

→ cluster analysis?

- Control for confounding effects

→ later hexameter poetry

- Maybe there is just no difference!

→ then why the concept of formular economy?

Can Distributional Semantics teach us something about formulae? Maybe!

Dr Martina Astrid Rodda | they/them/theirs | University of Oxford

University College | martinaastrid.rodda@classics.ox.ac.uk

Improving morphological analysis of Greek with Transformer-based approaches: first results with ELECTRA

(First presented at the Day of Computational Approaches to Ancient Greek and Latin, Groningen)

Alek Keersmaekers

Wouter Mercelis

University of Leuven
RU Quantitative Lexicology and Variational Linguistics

Introduction: morphological tagging

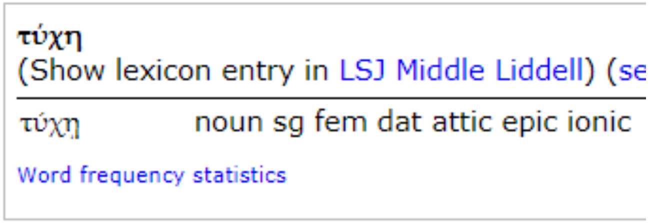
- Part-of-speech tagging: one of the oldest tasks in NLP
- Most approaches use the **context** of a word to decide on the correct morphological analysis
- E.g. “ἀλλ’ ἀγαθῇ **τύχῃ**” > τύχῃ is likely a dative feminine noun (and not a verb) because it is preceded by a dative feminine adjective
- Such patterns are automatically learned from large corpora (supervised machine learning)
- Typically, the context is defined as a fixed number of preceding words
- Through specific techniques (e.g. decision trees), the tagger can learn which words from the context are particularly relevant for the morphological analysis (e.g. *ἀλλ’* would not be relevant here)

Morphological tagging: optimizations for Ancient Greek (1)

- Ancient Greek morphological tags can be extremely complex (e.g. *participle,sg,aor,act,masc,dat*), which increases the number of possibilities to be taken into account
- Therefore morphological tags are typically split into individual attributes, which are predicted semi-independently and combined in some way (see later)
  - This is also done for context features: e.g. τῇ νῦν αὐτῷ **παρούσῃ** τύχῃ > the tense or voice of παρούσῃ is less relevant than the case/number/gender

Morphological tagging: optimizations for Ancient Greek (2)

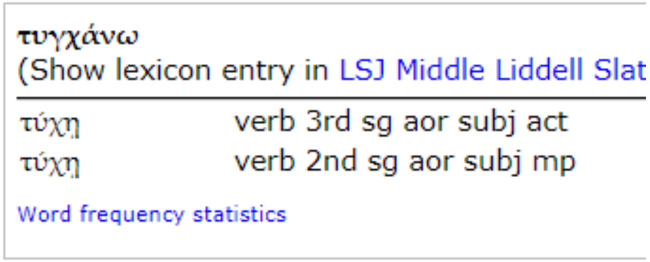
- – Due to the morphological complexity of Greek, corpora typically do not contain all inflected forms of each word (e.g. the treebanks do not contain a single attestation of λελυκώς)
- – Therefore a ‘morphological dictionary’ is typically used (e.g. *Morpheus*), so that the tagger already knows which possibilities to select from (e.g. τύχῃ: only 3 possibilities are considered)



τύχη
(Show lexicon entry in LSJ Middle Liddell) (se

τύχῃ	noun sg fem dat attic epic ionic

Word frequency statistics



τυγχάνω
(Show lexicon entry in LSJ Middle Liddell Slat

τύχῃ	verb 3rd sg aor subj act
τύχῃ	verb 2nd sg aor subj mp

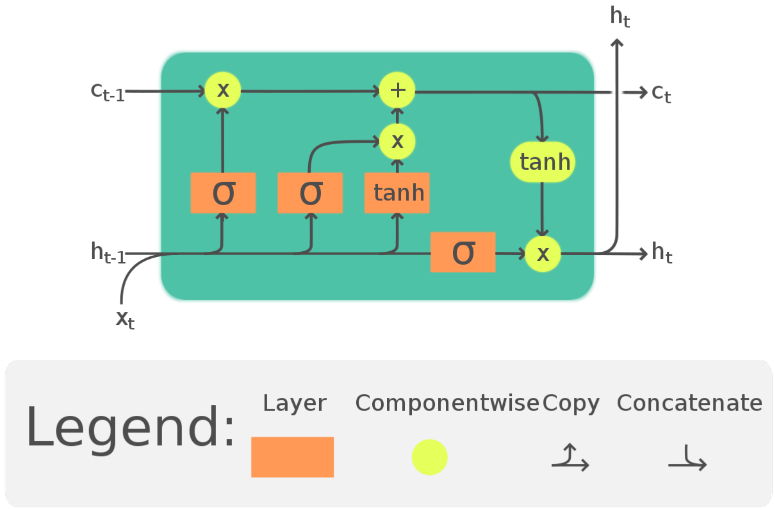
Word frequency statistics

Morphological tagging: problems

- Some tagging tools (e.g. *RFTagger*, *MarMoT*) that implement these optimizations therefore are able to reach a relatively high tagging accuracy for Ancient Greek (about 95%)
- However, one problem is the limited syntactic context these taggers use (*RFTagger*: context of about 4 preceding words)
  - E.g. **ἔστι** γὰρ τὸ πλῆθος τοῦ ἀργυρίου οὐκ **ὀλίγον**
- Syntactic context is often highly crucial for Greek morphological prediction (e.g. case, agreement)
- Also, the ‘semantic’ context of a word is not used (just the part-of-speech tags of the preceding words)

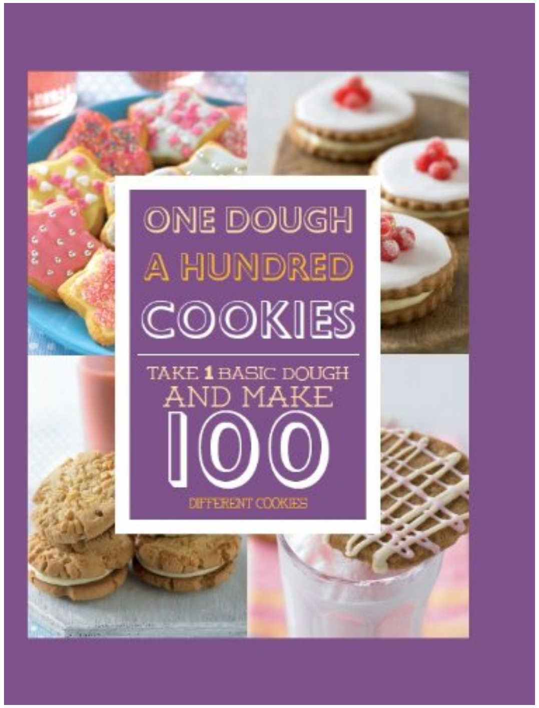
Neural language models

- Type of distributional language model
- Information about the usage of a word (syntax, semantics, …) is automatically learned from a large, unannotated corpus
- No handcrafted features necessary



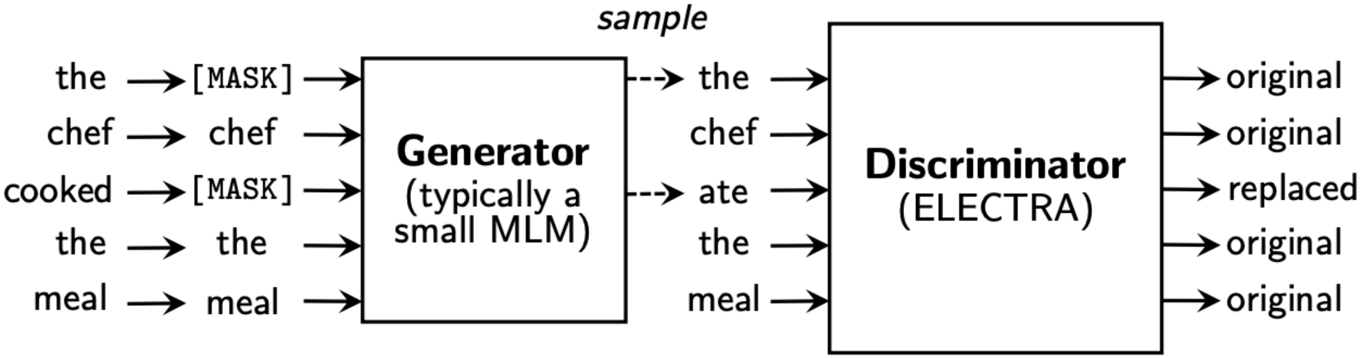
Transformer models (1)

- E.g. *BERT, ELECTRA*, *XLNET*
- Very good at automated processing of natural language sentences (‘self-attention’)
- Transfer learning:
  - First, the model learns general information about the usage of Greek words from a large corpus (GLAUx: about 29M tokens)
    = **pre-training**
  - Second, extra information from an annotated corpus (e.g. morphology, syntax) is added to the network = **fine-tuning**
  - Therefore, morphological and syntactic analysis are not dependent on each other anymore (avoids error propagation)



Transformer models (2)

- So how is information about the usage of a word learned?
- Predicting words on the basis of their context
- E.g. ELECTRA:
  - Some words in the training sentences are randomly replaced by a plausible alternative
  - The network has to learn which words are replaced and which are original



Fine-tuning ELECTRA for morphological tagging

- Token-classification task
- Training a classifier for each morphological attribute:
  - Part-of-speech tag
  - Number, gender, case
  - Person, tense, mood, voice
  - Degree
- Example: person (4 possible labels):
  - 1, 2, 3 for first, second and third person
  - _ for all words that do not have a person (e.g. prepositions)

Input data

- Language model: trained on literary part *GLAUx* (8BC-3AD) + papyri (3BC-8AD) (about 29M tokens: small for transformer models)

- Tagging model finetuned on treebank data (Perseus, PROIEL, Gorman, Leuven, PapyGreek, Harrington, Yordanova) (about 1.1M tokens)

Results: general

- Comparison with RFTagger: accuracy **0.946**
- In a first step, we combined all individual predictions
  - Can lead to ‘strange’ predictions e.g. adjective 1 sg pres ind act neut pos
- Next, we constrained the output to valid Greek morphological tags
  - Tag probability is calculated by multiplying individual prediction probabilities: e.g. P(noun sg masc acc) = P(noun) * P(sg) * P(masc) * P(acc)
- Finally, we further constrained the output to tags that were either in the training data or considered valid analyses by *Morpheus* (if possible)

	Accuracy
Combining predictions	0.958 (+1.2%)
Constraining to valid tags	0.963 (+1.7%)
Lexicon	0.969 (+2.3%)

Results by genre

	Correct	Total	Accuracy
Religion	33799	34308	98.5%
Rhetorical	24503	25209	97.2%
History & Biography	93926	96683	97.1%
Dialogue	1902	1961	97.0%
Narrative, Mythography & Paradoxography	9371	9663	97.0%
Epic Poetry	49117	50653	97.0%
Example Sentence	4152	4289	96.8%
Epistolography	1026	1060	96.8%
Polyhistory	9512	9874	96.3%
Documentary	8434	8780	96.1%
Science & Philosophy	16277	17020	95.6%
Drama	25019	26396	94.8%
Lyric Poetry	1000	1062	94.2%

Result by attribute

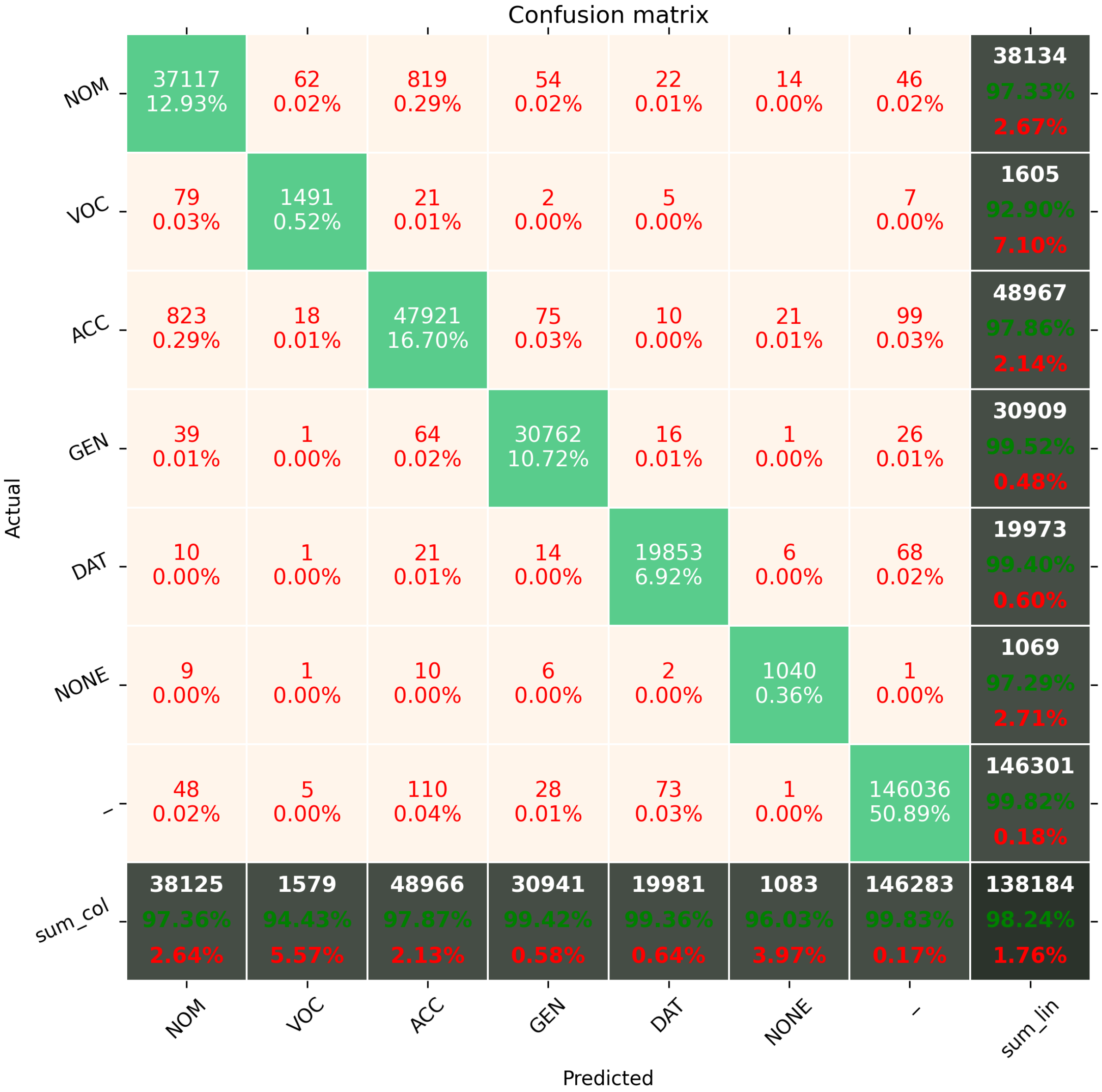
– Like with RFTagger, case, gender and mood are among the most difficult categories, but big improvements!

	RFTagger	ELECTRA
Number	0.991	0.994 (+0.3%)
Voice	0.990	0.994 (+0.4%)
Part-of-speech	0.986	0.989 (+0.3%)
Person	0.978	0.988 (+1.0%)
Mood	0.972	**0.984 (+1.2%)**
Tense	0.976	0.983 (+0.7%)
Case	0.960	**0.982 (+2.2%)**
Gender	0.963	**0.977 (+1.4%)**
Degree	0.961	0.974 (+1.3%)

Results: case

- Genitive and dative are excellent
- Points of confusion:
  - nominative and accusative (neuter!)
  - nominative and vocative (mostly identical)
  - accusative and _ (adverbs/adverbial accusatives)
- Improvement vs. RFTagger especially strong with nominative/accusative
  - Nominative
    - 0.945 precision (RFTagger) vs. 0.974 (ELECTRA)
    - 0.926 recall (RFTagger) vs. 0.973 (ELECTRA)
  - Accusative
    - 0.943 precision (RFTagger) vs. 0.979 (ELECTRA)
    - 0.958 recall (RFTagger) vs. 0.979 (ELECTRA)
- Total accuracy: 98.2%

Confusion matrix: case



Confusion matrix

Actual
Predicted

	NOM	VOC	ACC	GEN	DAT	NONE	_	sum_lin
NOM	37117 12.93%	62 0.02%	819 0.29%	54 0.02%	22 0.01%	14 0.00%	46 0.02%	38134 97.33% 2.67%
VOC	79 0.03%	1491 0.52%	21 0.01%	2 0.00%	5 0.00%		7 0.00%	1605 92.90% 7.10%
ACC	823 0.29%	18 0.01%	47921 16.70%	75 0.03%	10 0.00%	21 0.01%	99 0.03%	48967 97.86% 2.14%
GEN	39 0.01%	1 0.00%	64 0.02%	30762 10.72%	16 0.01%	1 0.00%	26 0.01%	30909 99.52% 0.48%
DAT	10 0.00%	1 0.00%	21 0.01%	14 0.00%	19853 6.92%	6 0.00%	68 0.02%	19973 99.40% 0.60%
NONE	9 0.00%	1 0.00%	10 0.00%	6 0.00%	2 0.00%	1040 0.36%	1 0.00%	1069 97.29% 2.71%
_	48 0.02%	5 0.00%	110 0.04%	28 0.01%	73 0.03%	1 0.00%	146036 50.89%	146301 99.82% 0.18%
sum_col	38125 97.36% 2.64%	1579 94.43% 5.57%	48966 97.87% 2.13%	30941 99.42% 0.58%	19981 99.36% 0.64%	1083 96.03% 3.97%	146283 99.83% 0.17%	138184 98.24% 1.76%

Error analysis (1)

- Analysis of 200 random errors
- In 37% (74) of all cases, ELECTRA actually made the right prediction (gold data was wrong)!
- Another 16% (32) are caused by inconsistencies in the data
  - E.g. “λίαν ἐν **Ἀρμενίοις** ἀφανῆ ἄνδρα > is this a noun or an adjective?
- 12% (24): the ‘meaning’ of a word is wrongly understood by ELECTRA > could theoretically be resolved with better models/more data
  - E.g. “πολλά δέ τ’ **ἀσθμαίνοντα** λέων ἐδάμασσε βίηφιν” > analyzed as neuter plural instead of masculine singular
- 8% (15) of the problems occur in elliptic contexts
  - E.g. “ἀεί διελέγετο σκοπῶν **τί** εὐσεβές”

Error analysis (2)

- 7% (13) are forms that are not recognized by *Morpheus* (e.g. dialectal, post-classical) > expanding lexicon/inflections
- 6% (12) of all problems are related to co-reference resolution (often word in other sentence)
  - E.g. “τοιήνδε δέ ἐξ **αὐτῶν** παρενθήκην ἐποιήσατο.”
- 4% (8) of problems are related to genre/dialect/diachrony
  - E.g. “τοῦ δέ ὄρεος **τό** περικληίει”: analyzed as article
- Finally, 6% (11) are difficult to explain (should be possible) > neural networks often a black box
  - E.g. “ἐλπίζω δέ ἐν κυρίῳ Ἰησοῦ Τιμόθεον ταχέως πέμψαι ὑμῖν , ἵνα καί ἐγώ **εὐψυχῶ** γνούς τά περί ὑμῶν”: analyzed as indicative rather than subjunctive

Conclusion

- Transformer models (ELECTRA) are able to make significant improvements for Greek morphological tagging, even with limited training data
- These improvements are particularly pronounced with difficult morphosyntactic categories (e.g. case, gender, mood)
- Some remaining problems can be tackled by improving the quality of the data + expanding Morpheus lexicon
- Nevertheless, several remaining problems are difficult to resolve, even for humans