Bringing Modern Spell Checking Approaches to Ancient Texts

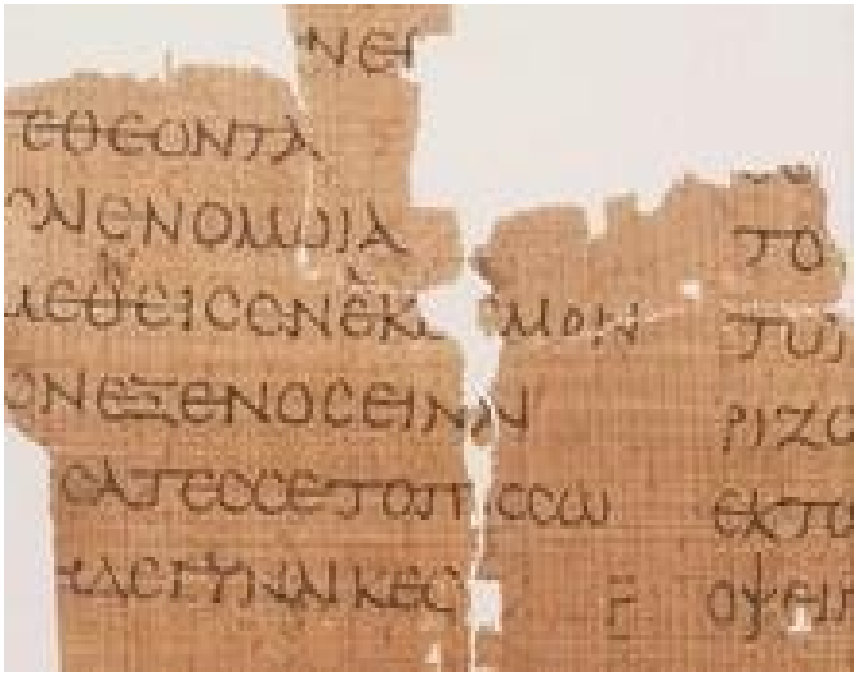
Automatized Suggestions for Incomplete/Damaged Words

2011 Digital Classicists Seminar
Leipzig, 2011/07/29

Marco Büchler
Natural Language Processing Group
Department of Computer Science
University of Leipzig

Agenda

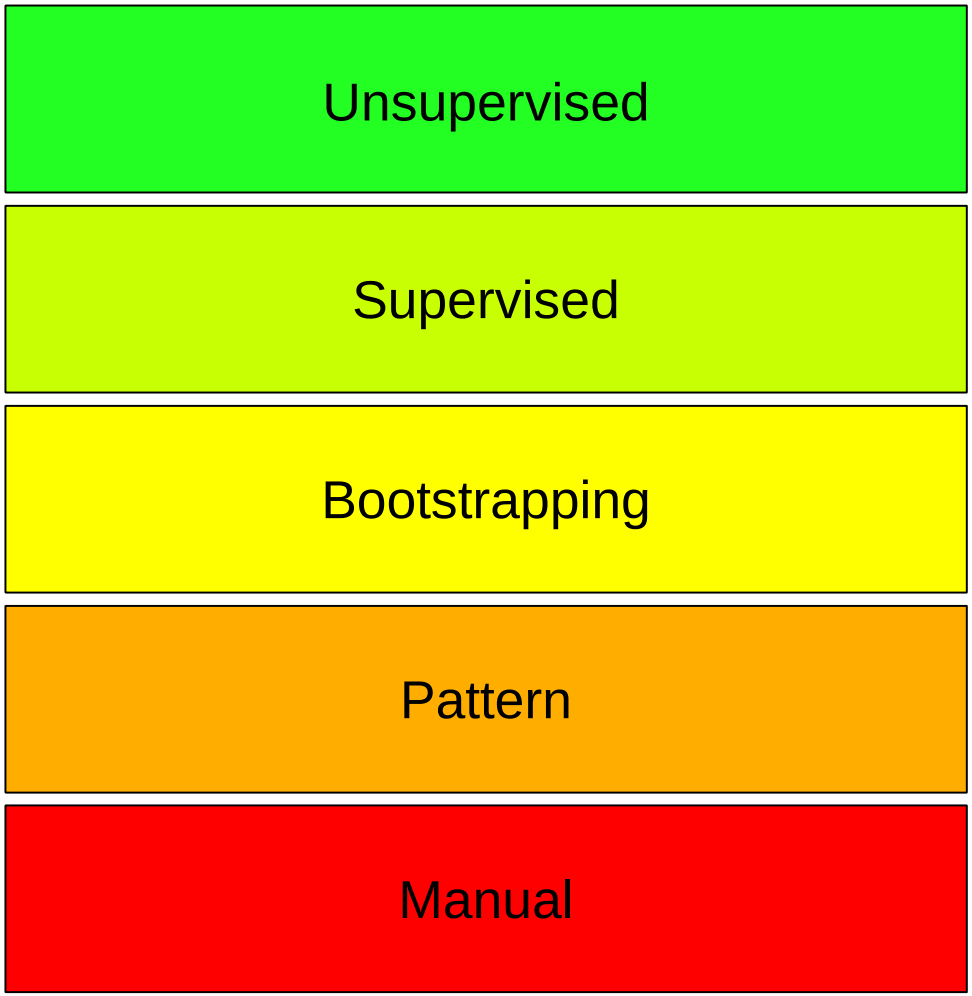
- What can you do with automatic suggestions of words for filling gaps?
- Comparison of different strategies for gap reconstruction
- 2 tasks of gap reconstruction
  - Identification of damaged words
  - Generation of suggestions
- Explanation by way of example
- Some problems / further steps



Textcorrection

Possible passage in the text:
Platon Timaios, 38c7 bis 38d4 (from TLG-Online):
σώματα δὲ αὐτῶν ἑκάστων ποιήσας ὁ θεὸς ἔθηκεν εἰς τὰς
@1 περιφορὰς ἃς ἡ θατέρου περίοδος ᾖειν, ἑπτὰ οὔσας ὄντα
(d.) ἑπτά, σελήνην μὲν εἰς τὸν περὶ γῆν πρῶτον, ἥλιον δὲ εἰς
τὸν δεύτερον ὑπὲρ γῆς, ἑωσφόρον δὲ καὶ τὸν ἱερὸν Ἑρμοῦ
λεγόμενον εἰς [τὸν] τάχει μὲν ἰσόδρομον ἡλίῳ κύκλον ἰόντας,τὴν δὲ ἐναντίαν εἰληχότας αὐτῷ δύναμιν·

Classes of mining tools



Text Completion
By (information) retrieval

Text completion today I

New Search | DDBDP Search | Numbers Search | News & Updates

Search Duke Databank of Documentary Papyri

Text Search

Search for:word or phrase: doul [Search] [Clear Form]

THEN word or phrase: oikogen within: 10 words

THEN word or phrase: within: 10 words

Use ^ [SHIFT-6] to anchor a substring to a word boundary. Example: ^kai to search words beginning with kai

- (•) input is in beta code
- ( ) input is **not** in beta code
- [ ] lemmatized search

Limit by

Place name

Publication series: [Select Publication Series] volume:

APIS collection [Select APIS Collection]

Date On or after: CE On or before: CE

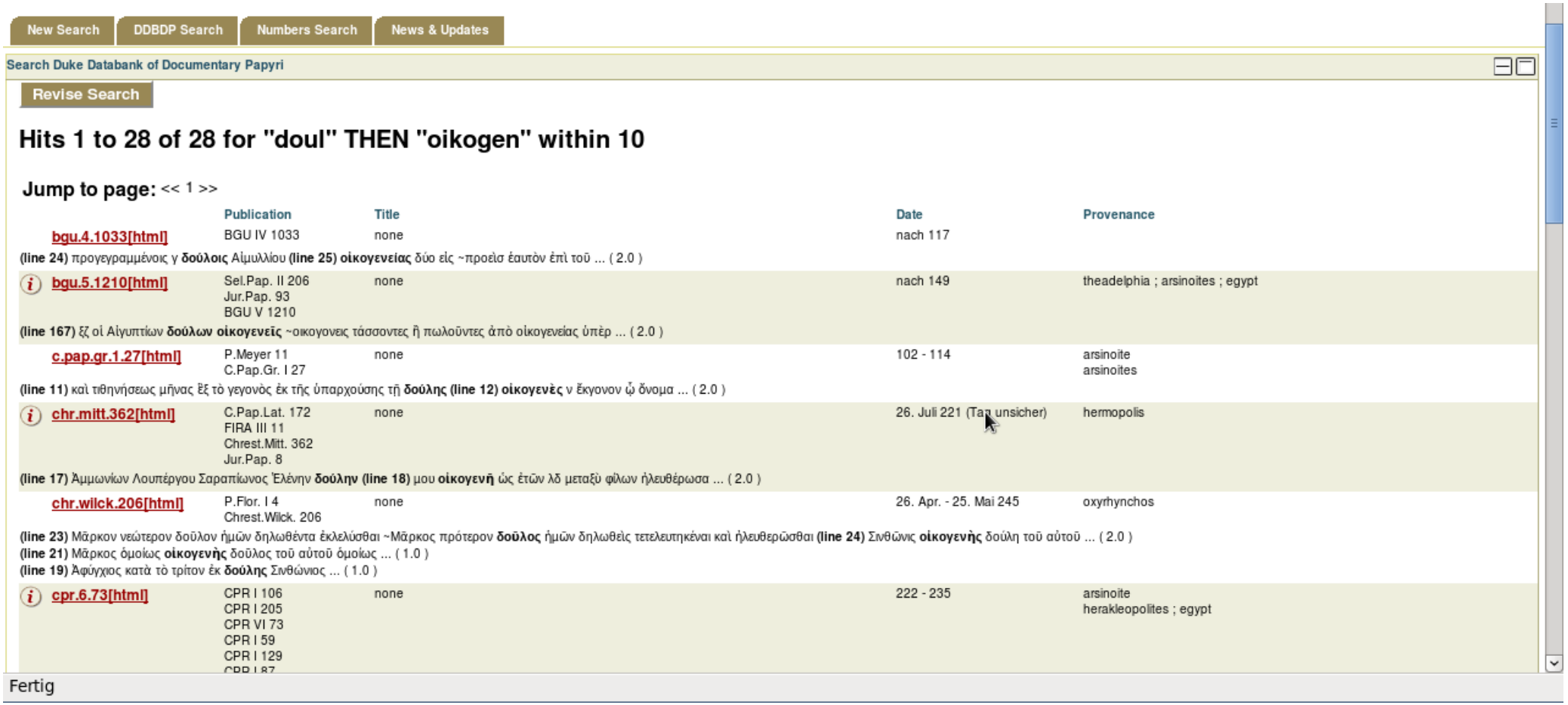
Options

- [ ] respect capitalization
- [ ] respect diacritics/accents
- [ ] hide highlighted fragments (faster)

Show records with:

- [ ] **images** first
- [ ] **translations** first

Source: papyri.info



New Search | DDBDP Search | Numbers Search | News & Updates

Search Duke Databank of Documentary Papyri

Revise Search

Hits 1 to 28 of 28 for "doul" THEN "oikogen" within 10

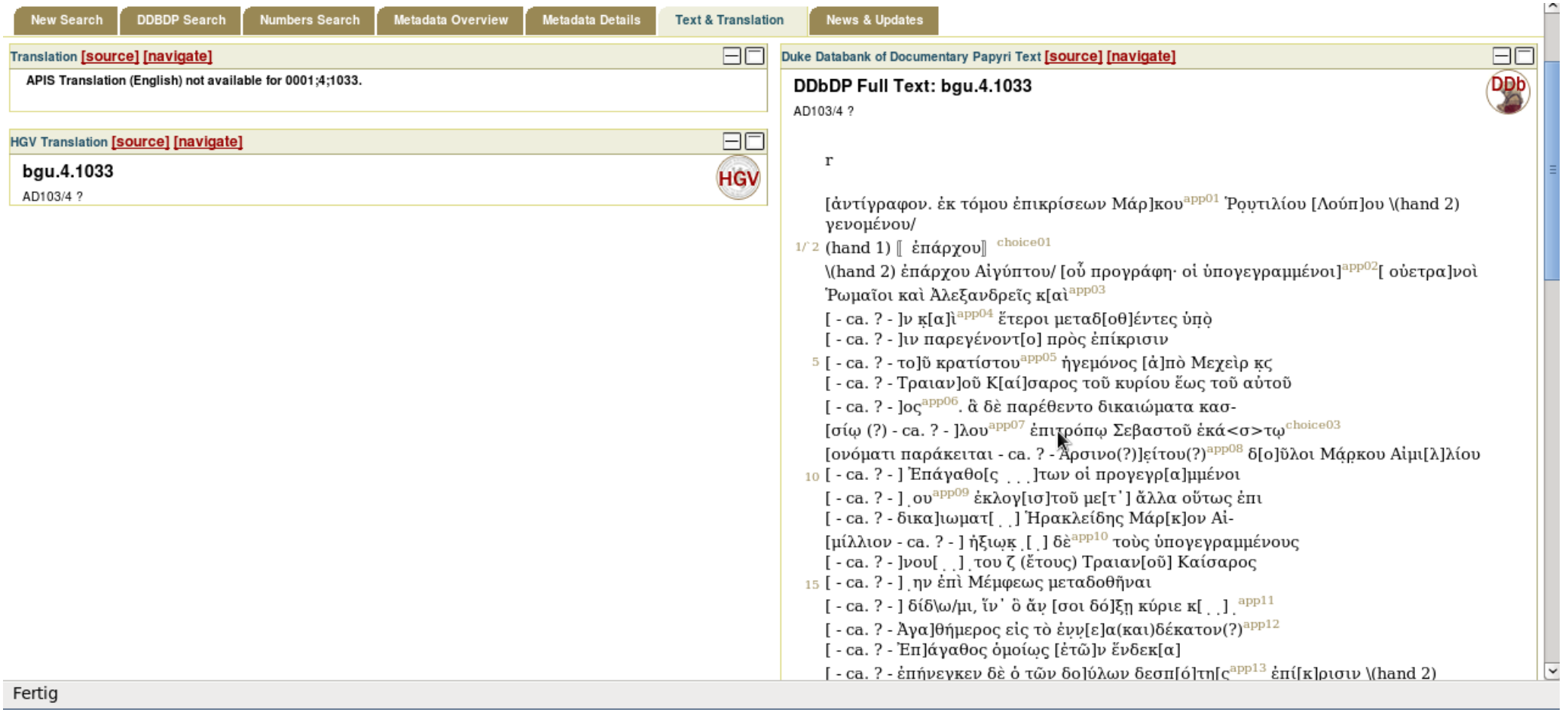
Jump to page: << 1 >>

	Publication	Title	Date	Provenance
bgu.4.1033[html]	BGU IV 1033	none	nach 117	
(line 24) προγεγραμμένοις γ **δούλοις** Αἰμιλλίου **(line 25) οἰκογενείας** δύο εἰς ~προεὶσ ἑαυτὸν ἐπὶ τοῦ ... (2.0)				
bgu.5.1210[html]	Sel.Pap. II 206 Jur.Pap. 93 BGU V 1210	none	nach 149	theadelphia ; arsinoites ; egypt
(line 167) ξζ οἱ Αἰγυπτίων **δούλων οἰκογενεῖς** ~οικογονεις τάσσοντες ἢ πωλοῦντες ἀπὸ οἰκογενείας ὑπὲρ ... (2.0)				
c.pap.gr.1.27[html]	P.Meyer 11 C.Pap.Gr. I 27	none	102 - 114	arsinoite arsinoites
(line 11) καὶ τιθηνήσεως μῆνας ἓξ τὸ γεγονὸς ἐκ τῆς ὑπαρχούσης τῇ **δούλης (line 12) οἰκογενὲς** ν ἔκγονον ᾧ ὄνομα ... (2.0)				
chr.mitt.362[html]	C.Pap.Lat. 172 FIRA III 11 Chrest.Mitt. 362 Jur.Pap. 8	none	26. Juli 221 (Tag unsicher)	hermopolis
(line 17) Ἀμμωνίων Λουπέργου Σαραπίωνος Ἑλένην **δούλην (line 18)** μου **οἰκογενῆ** ὡς ἐτῶν λδ μεταξὺ φίλων ἠλευθέρωσα ... (2.0)				
chr.wilck.206[html]	P.Flor. I 4 Chrest.Wilck. 206	none	26. Apr. - 25. Mai 245	oxyrhynchos
(line 23) Μᾶρκον νεώτερον δοῦλον ἡμῶν δηλωθέντα ἐκλελύσθαι ~Μᾶρκος πρότερον **δοῦλος** ἡμῶν δηλωθεὶς τετελευτηκέναι καὶ ἠλευθερῶσθαι **(line 24)** Σινθῶνις **οἰκογενὴς** δούλη τοῦ αὐτοῦ ... (2.0) **(line 21)** Μᾶρκος ὁμοίως **οἰκογενὴς** δοῦλος τοῦ αὐτοῦ ὁμοίως ... (1.0) **(line 19)** Ἀφύγχις κατὰ τὸ τρίτον ἐκ **δούλης** Σινθώνιος ... (1.0)				
cpr.6.73[html]	CPR I 106 CPR I 205 CPR VI 73 CPR I 59 CPR I 129 CPR I 87	none	222 - 235	arsinoite herakleopolites ; egypt

Fertig

<u>Problem 1:</u> Lots of texts needs to be read. Do you read the first hit in the same unbiased way as the last one?

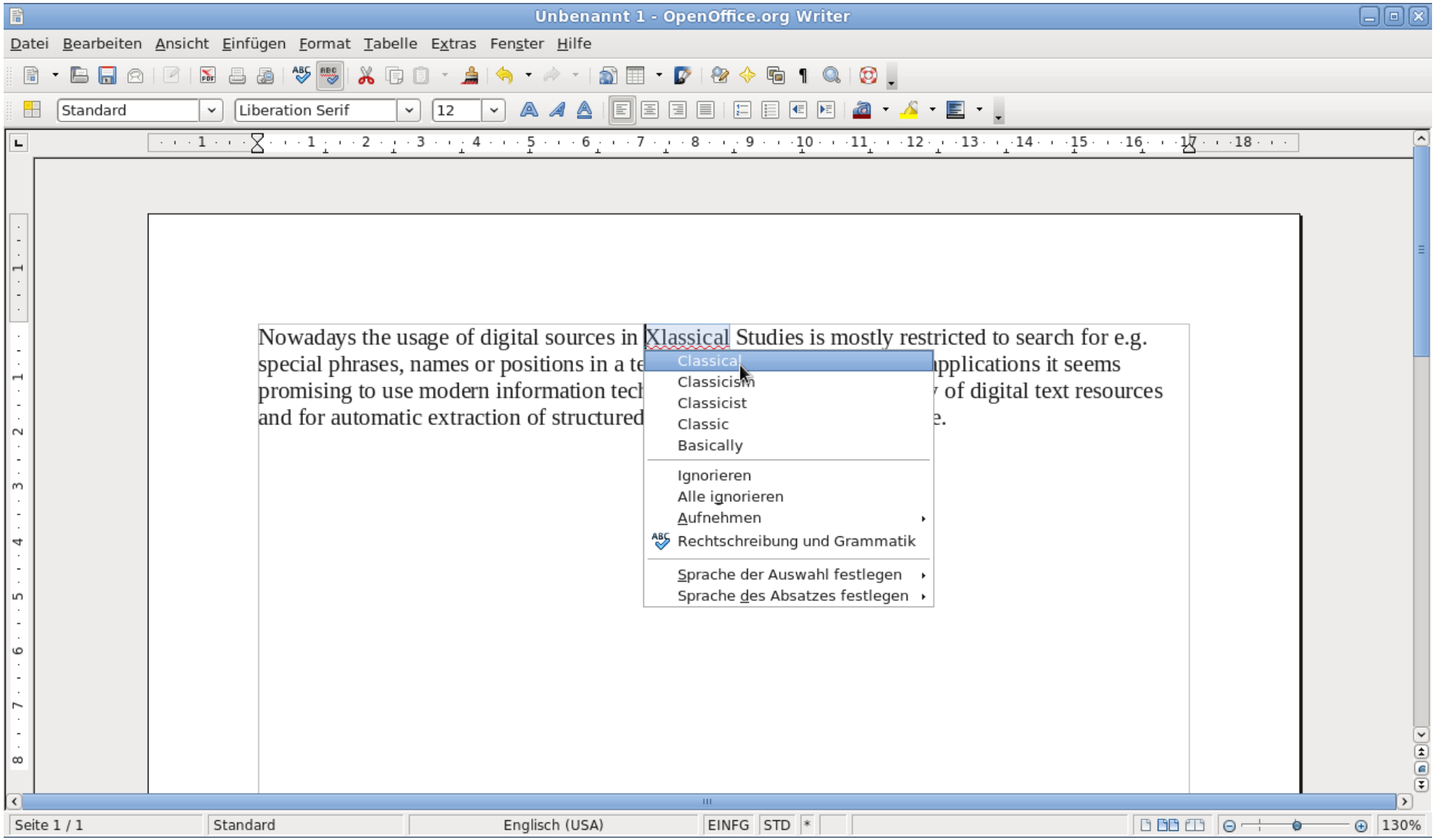
Source: papyri.info



<u>Problem 2:</u> Finding the right "signal words". Which words are good indicators for text completion?

Source: papyri.info

Spell checking today



Can we bring ancient texts and modern spell checking approaches together?

OR

Language models meet Classical Studies

Text Completion
By (text) mining

- **Finding words** that contain an error (Detection of candidates)
  - Ancient texts: Leiden conventions
  - Modern texts:
    - Trust of correctness by likelihood ration
    - Redundancy: „ negative information“ for misspelled or fragmentary words

- **Correcting candidate words (features are pre-computed on different corpora)**
  - By semantic approaches
  - By syntactical approaches
  - Word properties
  - String similarity
  - Named Entities
  - *Morphological knowledge*

- Currently processed corpora: **TLG**, PHI7, PHI7_INS, **PHI_DDP**, **epiDuke**

- *Pre-processing:*
  - All texts are segmented into sentences.
  - Meta information such as dating or classification are extracted.
  - Tokenisation

- *Training:*
  - Features (e. g. “signal words”) for every word are pre-computed in the background (up to 100s of millions datasets)
  - Features are classified by different approaches

- *Scoring the overall list:*
  - Main idea/assumption: Every known word in a corpus is a potential candidate for text completion.
  - That means: TLG about 1.7M words, epiDuke avout 550T words
  - Every approach delivers an independent list of candidates having a score between 0 and 1.
  - Overall candidate list is scored by the sum of a word's individual score by a selected algorithm

Textcorrection

Possible passage in the text:
Platon Timaios, 38c7 bis 38d4 (from TLG-Online):
σώματα δὲ αὐτῶν ἑκάστων ποιήσας ὁ θεὸς ἔθηκεν εἰς τὰς
@1 περιφορὰς ἃς ἡ θατέρου περίοδος ᾖειν, ἑπτὰ οὔσας ὄντα
(d.) ἑπτά, σελήνην μὲν εἰς τὸν περὶ γῆν πρῶτον, ἥλιον δὲ εἰς
τὸν δεύτερον ὑπὲρ γῆς, ἑωσφόρον δὲ καὶ τὸν ἱερὸν Ἑρμοῦ
λεγόμενον εἰς [τὸν] τάχει μὲν ἰσόδρομον ἡλίῳ κύκλον ἰόντας,τὴν δὲ ἐναντίαν εἰληχότας αὐτῷ δύναμιν·

- Detection of words by Leiden Conventions (Source: Wikipedia):
  - [abc]: letters missing from the original text due to lacuna, but restored by the editor
  - <ab>: characters erroneously omitted by the ancient scribe, restored by the editor
  - [[abc]]: deleted letters
  - ...

Finding candidates

V.3	Erased and lost	⟦ [...] ⟧ ⟦ ... ⟧ ⟦ [...] ⟧	<del rend="erasure"><gap reason="lost" quantity="3"/></del>
V.3	Erased and lost	⟦ [c. 5] ⟧	<del rend="erasure"><gap reason="lost" quantity="5"/></del>
V.3	Erased and lost	⟦ [---] ⟧	<del rend="erasure"><gap reason="lost" extent="unknown"/></del>
VI.1	Text struck over erasure	⟪ abc ⟫	<add place="overstrike">αβγ</add>
VI.1	Overstruck text, incomprehensible	⟪ ABC ⟫	<add place="overstrike"><orig>αβγ</orig></add>
VI.1	Overstruck text ambiguous	⟪ ạḅc̣ ⟫	<add place="overstrike"><unclear>αβγ</unclear></add>
VI.2	Overstruck text, lost but restored	⟪ [abc] ⟫	<add place="overstrike"><supplied reason="lost">αβγ</supplied></add>
VI.3	Overstruck text, completely lost	⟪ [...] ⟫	<add place="overstrike"><gap reason="lost" quantity="3" unit="character"/></add>
VI.3	Overstruck text, lost, extent approximate	⟪ [c. 5] ⟫	<add place="overstrike"><gap reason="lost" quantity="5" unit="character" precision="low"/></add>

Gabriel Bodard (et al.), (2006-2009), _EpiDoc Cheat Sheet: Krummrey-Panciera sigla & EpiDoc tags_, version 1085, accessed: 2010-07-04. available <http://epidoc.svn.sourceforge.net/viewvc/epidoc/trunk/guidelines/msword/cheatsheet.doc>

- **Semantically** best word (co-occurrences)
- **Syntactically** best word (N-gram)
- **String similar** best word (Levenshtein, FastSS)
- **Word length** (Stoichedon texts)
- Best word by **domain classification** by
  - Mathematics and mechanics
  - Centuries
  - Cities
  - Jurisdiction
  - Slave trading

- ***Toy sample: A b C d <E> G h.***

- *Semantic approach:*
  - Features: sentences based co-occurrences (function words filtered)
  - Toy sample: A, C, G are selected as semantic profile
  - Looking for words that have the best overlap with the semantic profile (all permutations are possible)
  - Real world example: <E>=Τροία: Τροίας, Τρωΐα, Τροίᾳ, Τροίαν, Τροίη,῞**Ιλιος, ῞Ιλιος, ῞Ιλου**

- *Syntactical approach:*
  - Method: Looking for immediately neighboured words (bi-gram level)
  - Toy sample: d, G are selected as features

- *Word similarity:*
  - Method: letter bi-gram overlapping (word)
  - *Real word examples: γίγνηται and γίνηται or συναγαγόντες and ξυναγαγόντες*

- *Named Entity list and word length*

Which approach for what?

- *Semantic approach:*
  - If a word occurs typically in static semantic context
  - If a word occurs in a context that has a significant amount of content words

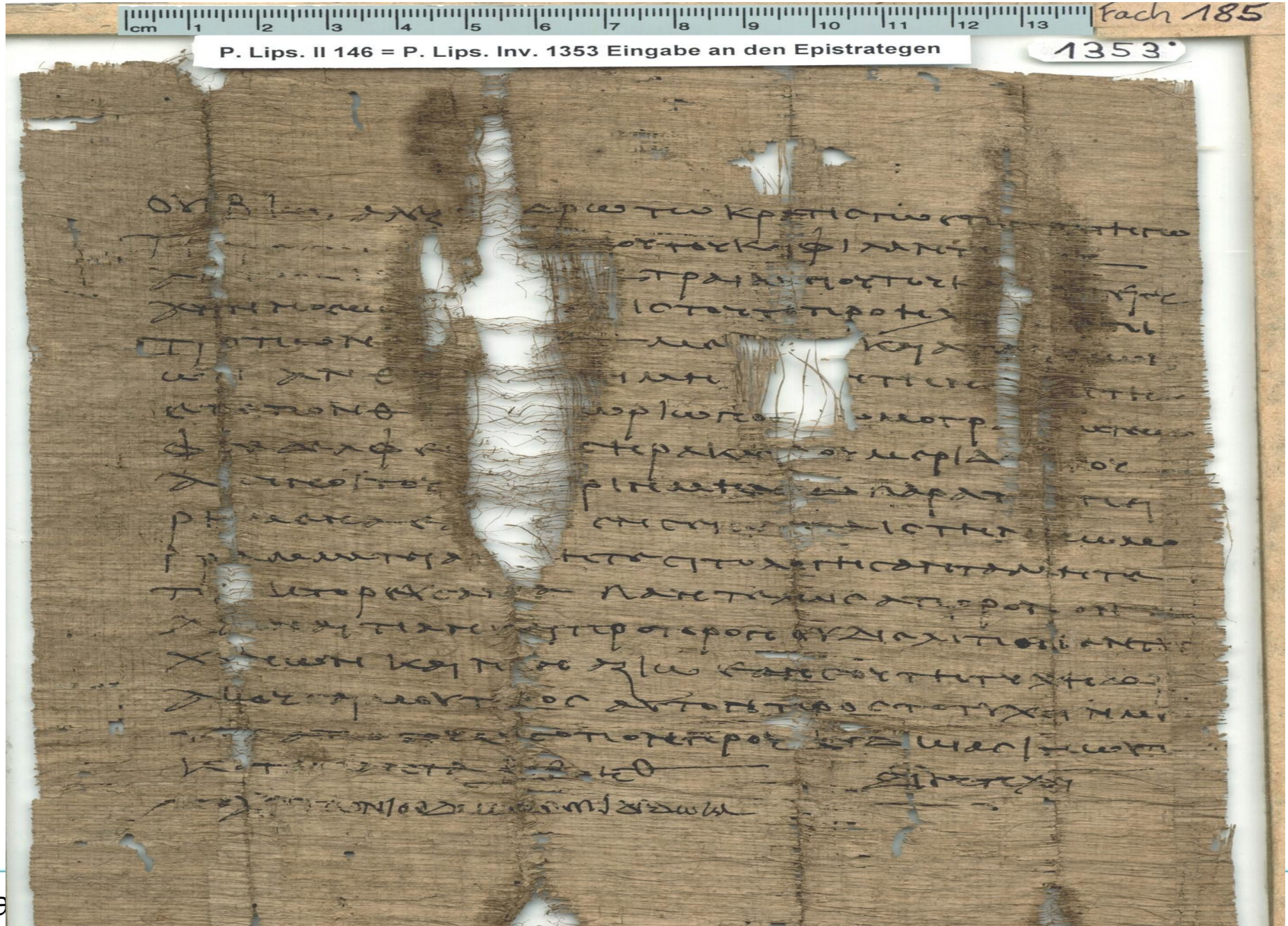
- *Syntactical approach:*
  - If a word occurs in a quite static syntactic pattern
  - If a word is part of a multi word expressions like *King Alexander the Great*

- *Word similarity:*
  - If a misspelled or fragmentary word still has enough recovered letters
  - Free of any sense: words with e.g. 2 letters. NO RESTRICTION

- *Word length*
  - If word length is known

- *Named Entity list*
  - If it makes sense to restrict the candidate list to person names (e.g. deletion).

- ***What is the ORIGINAL missing word?***



Οὐιβίῳ **Ἀλεξά[ν]δρῳ** τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου
Ἀντωνίο[υ Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ Στρα]τείου
Ἀντινοέως. [οὐκ ἂν] εἰς τοῦτο προήχθ[η]ν, ἐπ-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὢν ἄνθρ[ωπος,] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην
ἐπεπόνθ[ειν ὑπὸ] Ὡρίωνο[ς κ]ωμογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινοίτου. [οὐ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἐα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο-
γραμματείαν [μ]ήτε σιτολογήσαντα μήτε
πρ[α]κτορεύσαντα παντελῶς ἄπορον ὄν[τ]α.
δι᾽ ἣν αἰτίαν καὶ πρότερον οὐ διέλιπον ἐντυγ-
χάνων καὶ νῦν ἀξιῶ, ἐάν σου τῇ τύχῃ δόξ[ῃ],
ἀκοῦσαί μου π[ρ]ὸς αὐτὸν πρὸς τὸ τυχεῖν με
τῆς ἀπὸ σοῦ [μι]σοπονήρου ἐγδ[ι]κίας, ἵν᾽ ὧ ὑπὸ [σ]οῦ
κατὰ πάντα βεβοηθ(ημένος). διευτύχει
Ἀντώνιος Δόμνος ἐπιδέδωκα.

Step 1: Take text and copy it to the web page

Text

Οὐιβίῳ [_]λεξά[____] τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου.
Ἀντωνίο[υ Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ Στρα]τείου
Ἀντινοέως. [οὐκ ἂν] εἰς τοῦτο προήχθ[η]ν, ἐπι-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὢν ἄνθρ[ωπος,] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην
ἐπεπόνθ[ειν ὑπὸ] Ὡρίωνο[ς κ]ωμογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινοίτου. [οὗ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἑαυτὸν ἐνσείσαντα εἰς τὴν κωμη-

TLG PHI7 Epiduke Send

#	Sentence

Step 2: Choosing damaged word

Text

Οὐιβίῳ [_]λεξά[____] τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου.
Ἀντωνίο[υ Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ Στρα]τείου
Ἀντινοέως. [οὐκ ἂν] εἰς τοῦτο προήχθ[η]ν, ἐπι-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὢν ἄνθρ[ωπος,] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην
ἐπεπόνθ[ειν ὑπὸ] Ὡρίωνο[ς κ]ωμογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινοίτου. [οὗ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἑα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο-

TLG
PHI7
Epiduke
Send

#	Sentence
0	Οὐιβίῳ [_]λεξά[____] τῷ κρατίστῳ ἐπιστρατήγῳ παρὰ Ἀντ[ωνίου] [Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου Ἀντωνίο[υ] [Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ] [Στρα]τείου Ἀντινοέως
1	[οὐκ] [ἂν] εἰς τοῦτο προήχθ[η]ν ἐπι- τρόπων [μέγιστ]ε μέ[τριος] καὶ ἀπράγμων ὢν ἄνθρ[ωπος] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην ἐπεπόνθ[εν] [ὑπὸ] Ὡρίωνο[ς] [κ]ωμογρα[μ]ματέως Φ[ι]λαδελφεί[ας] [τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ Ἀρσινοίτου
2	[οὗ] [χά]ριν μην[ύ]ω παρὰ τ[ὰ] [ἀ]πει- ρημένα ἑα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο- γραμματείαν [μ]ήτε σπολογήσαντα μήτε πρ[α]κτορεύσαντα παντελῶς ἄπορον ὄν[τ]α
3	δι ἣν αἰτίαν καὶ πρότερον οὐ διέλιπον ἐντυγ- χάνων καὶ νῦν ἀξιῶ ἐάν σου τῇ τύχῃ δόξ[ῃ] ἀκοῦσαί μου π[ρ]ὸς αὐτὸν πρὸς τὸ τυχεῖν με τῆς ἀπὸ σοῦ [μ]σοπονήρου ἐγδ[ι]κίας ἵν ὦ ὑπὸ [σ]οῦ κατὰ πάντα βεβοηθ(ημένος)
4	διευτύχει Ἀντώνιος Δόμνος ἐπιδέδωκα

Strategy 1

Only use of information about the damaged word

Step 3: Choosing set of algorithms

[_]λεξά[___]

Interpreted word : _λεξά____

Length : 9

Candidate	Score	☐ Word length	☐ Neighboured letter bigrams	☐ Word similarity (letters)	☐ Named Entity	☐ Word bigram	☐ Semantic context	☐ Classification	Show
Ἀλεξάνδρα	2	1.0		1.0					
Ἀλεξάνρου	2	1.0		1.0					
Ἀλεξάνδρα	2	1.0		1.0					
Ἀλεξάνδρῳ	2	1.0		1.0					
Ἀλεξάρχου	2	1.0		1.0					
Ἀλεξάνδου	2	1.0		1.0					
ἐνεχάραξα	1	1.0							
ὁμολογει῀	1	1.0							
ὁλοκλήρον	1	1.0							

Strategy 2

Only use of any kind of context information

Step 3: Choosing set of algorithms

[_]λεξά[___]

Interpreted word : _λεξά____

Length : 9

Candidate	Score	☐ Word length	☐ Neighboured letter bigrams	☐ Word similarity (letters)	☐ Named Entity	☐ Word bigram	☐ Semantic context	☐ Classification	Show
νομοῦ	3					0.5	0.4	0.000	
Ἀλεξάνδρῳ	3					1.0	0.8	0.003	
ἀπόδος	2					0.5		0.001	
Δόμνου	2						0.8	0.040	
ἀνέτενα	2						0.2	0.025	
Αὐρηλίου	2						0.4	0.000	
Ἀχιλλεῖ	2					0.5	0.2		
Ἑπτὰ	2						0.2	0.010	
διαδεχομένῳ	2						0.2	0.250	
Σεουηριανῷ	2					0.5	0.2		
Αἰγύπτου	2						0.4	0.001	
ἡγεμόνι	2						0.2	0.001	
Λικιννιανῷ	2						0.2	0.200	

The 'Real' Strategy 2

(remove everthing that you know about the damaged word)

Only use of any kind of context information

Step 1&2: Changing text and choosing damaged word

Text

```
Οὐιβίῳ [_] τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου
Ἀντωνίο[υ Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ Στρα]τείου
Ἀντινοέως. [οὐκ ἂν] εἰς τοῦτο προήχθ[η]ν, ἐπι-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὢν ἄνθρ[ωπος,] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην
ἐπεπόνθ[ειν ὑπὸ] Ὡρίωνο[ς κ]ωμογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινοίτου. [οὔ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἑα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο-
```

TLG
PHI7
Epiduke
Send

#	Sentence
0	Οὐιβίῳ [_] τῷ κρατίστῳ ἐπιστρατήγῳ παρὰ Ἀντ[ωνίου] [Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου Ἀντωνίο[υ] [Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ] [Στρα]τείου Ἀντινοέως
1	[οὐκ] [ἂν] εἰς τοῦτο προήχθ[η]ν ἐπι- τρόπων [μέγιστ]ε μέ[τριος] καὶ ἀπράγμων ὢν ἄνθρ[ωπος] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην ἐπεπόνθ[ειν] [ὑπὸ] Ὡρίωνο[ς] [κ]ωμογρα[μ]ματέως Φ[ι]λαδελφεί[ας] [τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ Ἀρσινοίτου
2	[οὔ] [χά]ριν μην[ύ]ω παρὰ τ[ὰ] [ἀ]πει- ρημένα ἑα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο- γραμματείαν [μ]ήτε σπολογήσαντα μήτε πρ[α]κτορεύσαντα παντελῶς ἄπορον ὄν[τ]α
3	δι ἣν αἰτίαν καὶ πρότερον οὐ διέλιπον ἐντυγ- χάνων καὶ νῦν ἀξιῶ ἐάν σου τῇ τύχῃ δόξ[ῃ] ἀκοῦσαί μου π[ρ]ὸς αὐτὸν πρὸς τὸ τυχεῖν με τῆς ἀπὸ σοῦ [μ]σοπονήρου ἐγδ[ι]κίας ἵν ὦ ὑπὸ [σ]οῦ κατὰ πάντα βεβοηθ(ημένος)
4	διευτύχει Ἀντώνιος Δόμνος ἐπιδέδωκα

[_]

Interpreted word : _

Length : 1

Candidate	Score	☐ Word length	☐ Neighboured letter bigrams	☐ Word similarity (letters)	☐ Named Entity	☐ Word bigram	☐ Semantic context	☐ Classification	Show
νομοῦ	3					0.5	0.4	0.000	
Ἀλεξάνδρῳ	3					1.0	0.8	0.003	
ἀπόδος	2					0.5		0.001	
Δόμνου	2						0.8	0.040	
ἀνέτεινα	2						0.2	0.025	
Αὐρηλίου	2						0.4	0.000	
Ἀχιλλεῖ	2					0.5	0.2		
Ἑπτὰ	2						0.2	0.010	
διαδεχομένῳ	2						0.2	0.250	
Σεουηριανῷ	2					0.5	0.2		
Αἰγύπτου	2						0.4	0.001	
ἡγεμόνι	2						0.2	0.001	
Λικιννιανῷ	2						0.2	0.200	

Strategy 3

The Full Strategy – choosing whatever makes sense

[_]λεξά[___]

Interpreted word : _λεξά____

Length : 9

Candidate	Score	☐ Word length	☐ Neighboured letter bigrams	☐ Word similarity (letters)	☐ Named Entity	☐ Word bigram	☐ Semantic context	☐ Classification	Show
Ἀλεξάνδρῳ	5	1.0		1.0		1.0	0.8	0.003	
γενομένην	3	1.0				0.5	0.2		
διοίκησιν	3	1.0					0.2	0.008	
νομοῦ	3					0.5	0.4	0.000	
Ἀντινοέως	3	1.0					0.8	0.011	
βιβλιδίου	3	1.0					0.4	0.003	
ἐπιπρόπων	3	1.0					0.2	0.005	
Ἀντινοέων	3	1.0					0.4	0.002	
Δημητρίωι	3	1.0					0.2	0.002	
βιβλιδίων	3	1.0					0.2	0.002	
Στρατείου	3	1.0					0.6	0.214	
στρατηγὸς	3	1.0					0.2	0.001	
Ἀλεξάρχου	2	1.0		1.0					

- Apparatus criticus

- Post selection of word list

- More approaches (up to 30 (automatic) approaches are planned/possible)
  - Separation of left hand side and right hand side signal words
  - Semantics:
    - Semantics by Wittgenstein: Co-occurrences
    - Semantics by Firth (here HGV classification): You shall know a word by the company it keeps.
    - Different semantic spaces

- Does any cluster of algorithms fit to different annotations of the Leiden conventions?

- What are good training data for suggesting words?

- What is of interest for correction? (If no Leiden Convention exists)
  - Technically it is difficult to make distinction between spelling errors, OCR errors, or just variants, loan words, person names

Questions

?