Fragmentary Texts and Digital Collections of Fragmentary Authors

Monica Berti
"I Frammenti degli Storici Greci"
Dipartimento di Antichità e Tradizione Classica
Università di Roma Tor Vergata, Italia

Marco Büchler
Natural Language Processing Group
Institute of Mathematics and Computer Science
University of Leipzig, Germany

Digital Classicist 2010 Summer Seminar Programme
Institute of Classical Studies, London, 30th July 2010

What is a fragment?
(*Oxford English Dictionary*, s.v. fragment)

- a part broken off or otherwise detached from a whole
- a part remaining or still preserved when the whole is lost or destroyed
- an extant portion of a writing or composition which as a whole is lost
- a portion of a work left uncompleted by its author

Different kinds of fragments

- material fragments
- textual fragments

material fragments

material fragments
= physical remains of ancient evidence

reconstruction of the monument

textual fragments (1)

textual fragments
= material fragments bearing textual evidence

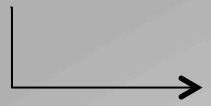
→ surviving broken off pieces of ancient writings

textual fragments (2)

Athenaeus, *Deipnosophistai* 10.67 (447c)

Ἑλλάνικος δ᾽ ἐν Κτίσεσι καὶ ἐκ ῥιζῶν, φησι κατασκευάζεται τὸ βρῦτον γράφων ὧδε· ‘πίνουσι δὲ βρῦτον ἔκ τινων ῥιζῶν, καθάπερ οἱ Θρᾷκες ἐκ τῶν κριθῶν’. Ἑκαταῖος δ᾽ ἐν δευτέρῳ Περιηγήσεως εἰπὼν περὶ Αἰγυπτίων ὡς ἀρτοφάγοι εἰσὶν ἐπιφέρει· ‘τὰς κριθὰς ἐς τὸ πῶμα καταλέουσιν’. ἐν δέ τῇ τῆς Εὐρώπης περιόδῳ Παίονάς φησι πίνειν βρῦτον ἀπὸ τῶν κριθῶν καὶ παραβίην ἀπὸ κέγχρου καὶ κόνυζαν. ‘ἀλείφονται δέ’, φησίν, ‘ἐλαίῳ ἀπὸ γάλακτος’. καὶ ταῦτα μέν ταύτῃ.

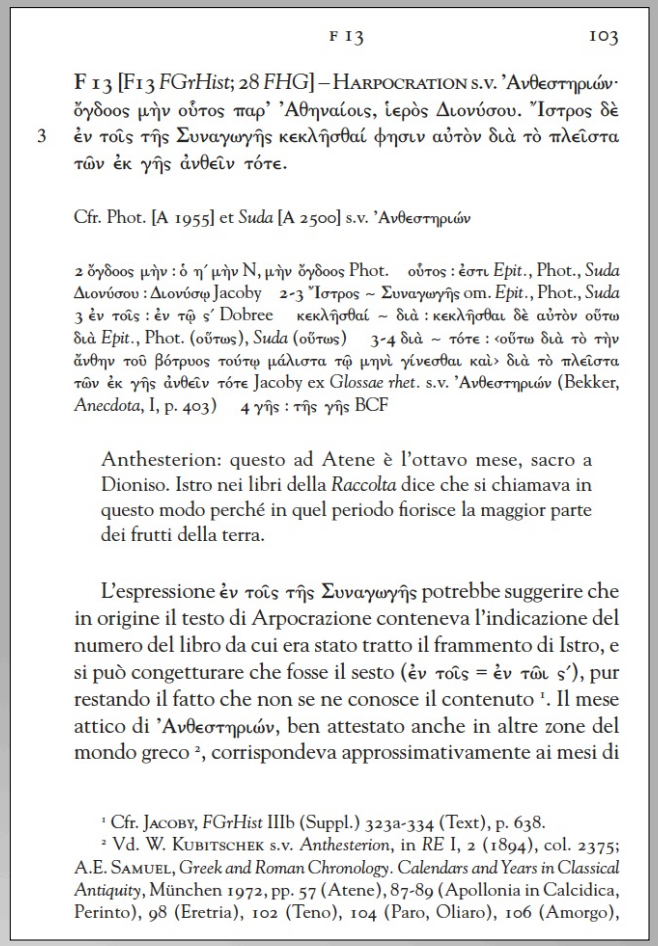
Hellanicus in *The Foundings* says that beer is made also of rye; he writes as follows: ‘They drink beer made of rye, as the Thracians drink it made of barley’. Hecataeus, in the second book of his *Description*, after saying of the Egyptians that they were bread-eaters, continues: ‘They grind up the barley to make the drink’. And in *The Description of Europe* he says that the Paeonians drink a beer made from barley, also *parabias*, made from millet, and even fleabane. ‘They also anoint themselves’, he says, ‘with an oil made from milk’. So much for that. (trans. Gulick)



textual fragments
= quotations of lost works embedded into other texts

print collections of fragmentary texts

- textual excerpts drawn from many different sources
- excerpts arranged according to various criteria
- length of the excerpts different from one edition to another
- when printed the excerpt gives a false illusion of materiality
- duplication of the same text in multiple editions



F 13 103

F 13 [F13 *FGrHist*; 28 *FHG*] – Harpocration s.v. Ἀνθεστηριών· ὄγδοος μὲν οὗτος παρ' Ἀθηναίοις, ἱερὸς Διονύσου. Ἴστρος δὲ
3 ἐν τοῖς τῆς Συναγωγῆς κεκλῆσθαί φησιν αὐτὸν διὰ τὸ πλεῖστα τῶν ἐκ γῆς ἀνθεῖν τότε.

Cfr. Phot. [A 1955] et *Suda* [A 2500] s.v. Ἀνθεστηριών

2 ὄγδοος μὲν : ὁ η΄ μὲν N, μὲν ὄγδοος Phot. οὗτος : ἐστι *Epit.*, Phot., *Suda* Διονύσου : Διονύσῳ Jacoby 2-3 Ἴστρος ~ Συναγωγῆς om. *Epit.*, Phot., *Suda* 3 ἐν τοῖς : ἐν τῷ ς΄ Dobree κεκλῆσθαί ~ διὰ : κεκλῆσθαι δὲ αὐτὸν οὕτω διὰ *Epit.*, Phot. (οὕτως), *Suda* (οὕτως) 3-4 διὰ ~ τότε : ‹οὕτω διὰ τὸ τὴν ἄνθην τοῦ βότρυος τούτῳ μάλιστα τῷ μηνὶ γίνεσθαι καὶ› διὰ τὸ πλεῖστα τῶν ἐκ γῆς ἀνθεῖν τότε Jacoby ex *Glossae rhet.* s.v. Ἀνθεστηριών (Bekker, *Anecdota*, I, p. 403) 4 γῆς : τῆς γῆς BCF

Anthesterion: questo ad Atene è l'ottavo mese, sacro a Dioniso. Istro nei libri della *Raccolta* dice che si chiamava in questo modo perché in quel periodo fiorisce la maggior parte dei frutti della terra.

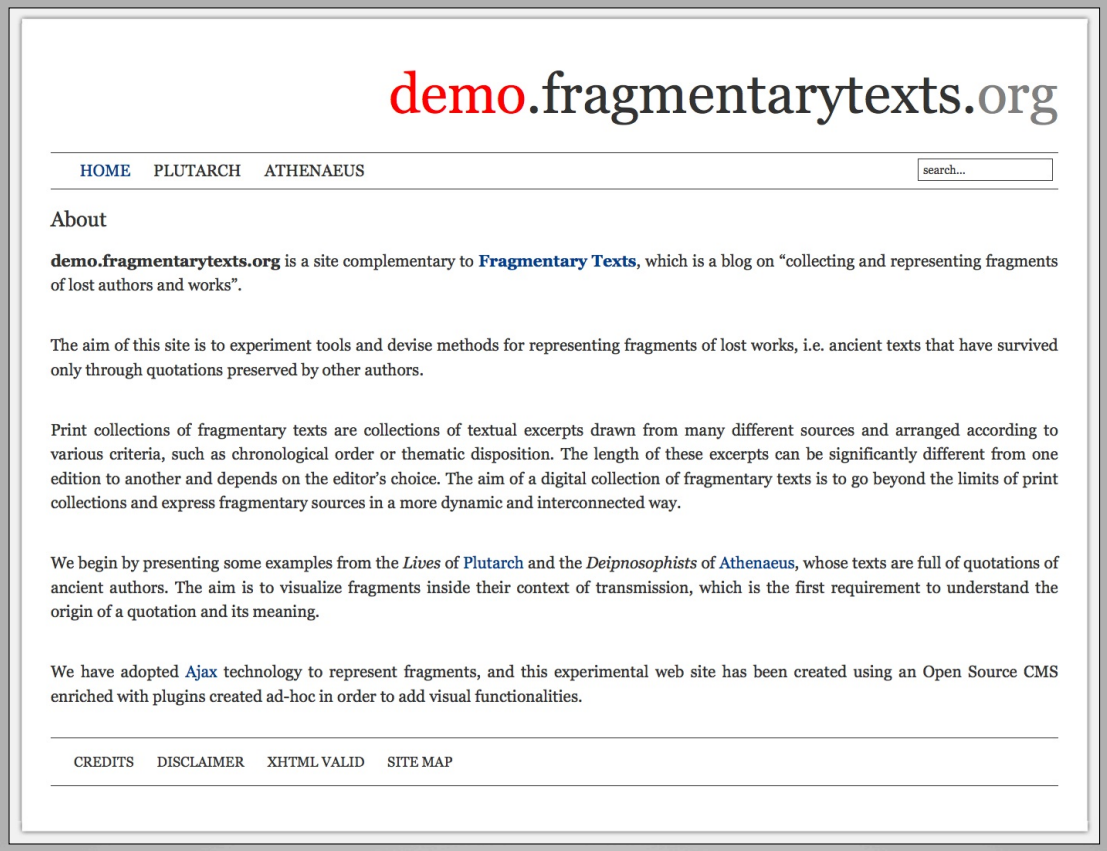
L'espressione ἐν τοῖς τῆς Συναγωγῆς potrebbe suggerire che in origine il testo di Arpocrazione conteneva l'indicazione del numero del libro da cui era stato tratto il frammento di Istro, e si può congetturare che fosse il sesto (ἐν τοῖς = ἐν τῶι ς΄), pur restando il fatto che non se ne conosce il contenuto [1]. Il mese attico di Ἀνθεστηριών, ben attestato anche in altre zone del mondo greco [2], corrispondeva approssimativamente ai mesi di

[1] Cfr. Jacoby, *FGrHist* IIIb (Suppl.) 323a-334 (Text), p. 638.
[2] Vd. W. Kubitschek s.v. *Anthesterion*, in *RE* I, 2 (1894), col. 2375; A.E. Samuel, *Greek and Roman Chronology. Calendars and Years in Classical Antiquity*, München 1972, pp. 57 (Atene), 87-89 (Apollonia in Calcidica, Perinto), 98 (Eretria), 102 (Teno), 104 (Paro, Oliaro), 106 (Amorgo),

representing textual fragments

- construct truly hypertextual editions, including not only excerpts but links to the scholarly sources from which those excerpts are drawn
- create meta-information through an accurate and elaborate semantic markup
- produce meta-editions consisting not only of isolated quotations, but also of pointers to the original contexts from which the fragments have been extracted
- provide scholars with an interconnected corpus of primary and secondary sources of fragments that also includes critical apparatuses, commentaries, translations, and modern bibliography on ancient texts

- textual fragment as hypertext
  - a text derived from another text and interconnected to many other different typologies of texts
- textual fragment as multitext
  - the result of a work of stratification of manuscripts and scholarly conjectures



http://demo.fragmentarytexts.org

Bibliography

Berti, M. et. al. "Collecting Fragmentary Authors in a Digital Library." In Proceedings of the 2009 Joint International Conference on Digital Libraries (JCDL '09). Austin, TX, 259-62. New York, NY: ACM Digital Library

Romanello, M. et al. "When Printed Hypertexts Go Digital: Information Extraction from the Parsing of Indices." In Hypertext 2009: Proceedings of the 20th ACM Conference on Hypertext and Hypermedia, Turin, Italy, 357-58. New York, NY: ACM Digital Library

Romanello, M. et al. Rethinking Critical Editions of Fragmentary Texts by Ontologies." In ELPUB 2009: 13th International Conference on Electronic Publishing: Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies, Milan, Italy, 155-74

Berti, M. "Fragmentary Texts and Digital Libraries". in *Philology in the age of Corpus and Computational Linguistics*. CHS Publication. Ed. G. Crane, A. Lüdeling, and M. Berti (forthcoming)

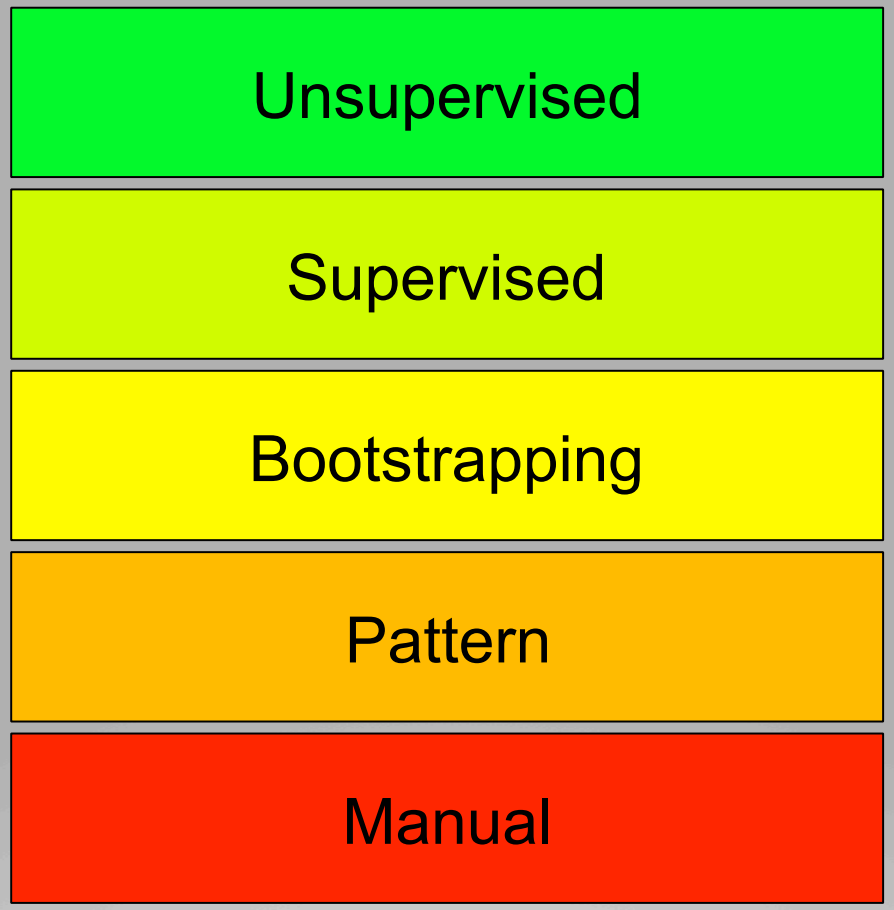
What is text mining?

„Process of deriving high-quality information from text“
(Feldman & Sanger 2006)

What is text mining?

„Text Mining is 'big reading'."
(Craig Bellamy on Twitter, Jul. 5th 2010)

Classes of text minings tools



Tasks of extracting and collecting fragmentary authors

- **Task 1**: Associations between person and work names
- **Task 2**: Extraction of fragments of an author
- **Task 3**: Finding new quotations and parallel texts
- **Task 4**: Expansion of the fragments' set

Task 1: Workflow person name extraction

- **Step 1**: Extraction of candidates by pattern such as
  - VN VN
  - VN ETH
  - VN LOC

Pattern

- **Step 2**: Resolving morphological dependencies using Morpheus

Pattern

- **Step 3**: Statistical evidence criterion

Unsupervised

- **Step 4**: Generating a similarity graph of those candidates and building valid concept classes

Unsupervised

- **Step 5**: Applying validated patterns on text in order to extract less frequent occurrences

Supervised

- **Step 6**: Iterating step 2 - 5

Botstrapping

Task 1: Some results of the PN extractor

- **Step 1**: Extraction of candidates by pattern such as
  - Ἑλλάνικος Λέσβιος (VN ETH)
- **Step 2**: Resolving morphological dependencies
  - Removing candidates like Ἑλλάνικος Ἀκουσιλάῳ VN VN
- **Step 3**: Statistical evidence criterion like min freq is 4.
- **Step 4**: Generating a similarity graph of those candidates and building valid concept classes e.g.

  Ἑλλάνικος Λέσβιος (VN ETH)

  Ἑλλάνικος ὁ Λέσβιος (VN ZN ETH)

Step 5: Applying validated patterns on text in order to extract less frequent occurrences

- Ἑλλάνικός τε ὁ Λέσβιος
- Ἑλλάνικος δὲ ὁ Λέσβιός
- Λέσβιος Ἑλλάνικος
- ...
- Overall after 1 iteration 16 different versions of Hellanicus of Lesbos

Task 3: Finding new quotations and parallel texts: pseudo algorithm

1 V = segment_corpus(C) with $v_1, v_2, ..., v_n \in V$, $\cup v_i = C$ and $v_i \neq v_j$
2 **for each** $v_i \in V$
3 F_i=train_features(v_i);

Training

4 **for each** $v_i \in V$
5 **for each** $f_k \in F_i$
6 $e_i=(v_i,v_j) \in E$=select all v_j containing feature f_k
7 **for each** $e_i \in E$

Linking

8 s_i=scoring($e_i=(v_i,v_j) \in E$; F_i; F_j);
9 if(s_i<*threshold*){$E=E\backslash\{e_i\}$}

Scoring

Task 3: Finding new quotations and parallel texts: Types of Completeness

Extraction of fragmentary authors

- *String approaches:*
  - *GST*
  - Letter n-grams
- Syntactic approaches (*(literal) quotations*):
  - N-gram expansion
  - Word n-grams
  - Distance based co-occurrences
- Semantic approaches (*parallel texts*):
  - *Semantic clustering*
  - *Semantic graph based approach(es)*
  - Relations of contrastive semantics
  - *Radius retrieval*
- *More complex approaches:*
  - *DCT*
  - *Winnowing*

Task 2: Extraction of fragments: Role of named entities

		Complete graph	w_id>=100 && freq(word)>1	w_id>=300 && freq(word)>1	w_id>=500 && freq(word)>1	Named Entities	Normalised Named Entities	Normalised Text and Named Entities
Graph properties	Number of nodes	538,572	388,929	363,359	353,618	1,149	4,487	2,178
	Number of co-occurrences	57,762,474	34,818,138	25,615,956	21,004,538	15,436	126,188	152,856
	Number of significant co-occurrences	30,382,422	21,739,476	17,687,582	15,462,940	14,876	69,858	84,124
	Percentage	0.53	0.62	0.69	0.74	0.96	0.55	0.55
	Average degree	56.41	55.90	48.68	43.73	12.95	15.57	38.62
Argumentation trail properties	Number of trails	$> 10^8$	$> 10^8$	$> 10^8$	$> 10^8$	361.094	7.958.240	3.087.581
	Average degree	15.34	9.93	7.70	6.79	7.03	7.77	9.93
	Average degree of internal node (trail length 2)	31.34	21.08	14.33	11.45	7.02	10.15	12.31
	Average degree of internal node (trail length 3)	301.38	362.56	285.86	231.39	55.66	76.06	81.86

Task 2: Extraction of fragments: Possible ways?

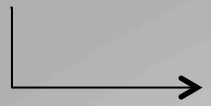
- **Option 1**: Statistical based — Supervised
- **Option 2**: Pattern based — Pattern
- **Option 3**: Completely different? — Unsupervised

Again: textual fragments

Athenaeus, *Deipnosophistai* 10.67 (447c)

Ἑλλάνικος δ᾽ ἐν Κτίσεσι καὶ ἐκ ῥιζῶν, φησι κατασκευάζεται τὸ βρῦτον γράφων ὧδε· ‘πίνουσι δὲ βρῦτον ἔκ τινων ῥιζῶν, καθάπερ οἱ Θρᾷκες ἐκ τῶν κριθῶν’. Ἑκαταῖος δ᾽ ἐν δευτέρῳ Περιηγήσεως εἰπὼν περὶ Αἰγυπτίων ὡς ἀρτοφάγοι εἰσὶν ἐπιφέρει· ‘τὰς κριθὰς ἐς τὸ πῶμα καταλέουσιν’. ἐν δέ τῇ τῆς Εὐρώπης περιόδῳ Παίονάς φησι πίνειν βρῦτον ἀπὸ τῶν κριθῶν καὶ παραβίην ἀπὸ κέγχρου καὶ κόνυζαν. ‘ἀλείφονται δέ’, φησίν, ‘ἐλαίῳ ἀπὸ γάλακτος’. καὶ ταῦτα μέν ταύτῃ.

Hellanicus in *The Foundings* says that beer is made also of rye; he writes as follows: ‘They drink beer made of rye, as the Thracians drink it made of barley’. Hecataeus, in the second book of his *Description*, after saying of the Egyptians that they were bread-eaters, continues: ‘They grind up the barley to make the drink’. And in *The Description of Europe* he says that the Paeonians drink a beer made from barley, also *parabias*, made from millet, and even fleabane. ‘They also anoint themselves’, he says, ‘with an oil made from milk’. So much for that. (trans. Gulick)



textual fragments
= quotations of lost works embedded into other texts

Why was it reused?

'They drink beer made of rye, as the Thracians drink it made of barley'.
the Paeonians drink a beer made from barley, also *parabias*, made from millet, and even fleabane.
'They also anoint themselves', he says, 'with an oil made from milk'.

Some „significance" related properties:
• **tf.idf**: Except „Thracian" and „Paeonians" all other words have a term weight of 0 (function words) or are weak content words.
• **Difference analysis**: no discriminating words
• **Log-likelihood ratio:** no discriminating words

Dale Chall Readability Index: [6.59;9.36] AVG: 7.85 (level of 9th – 10th grade of a secondary school)

Is there any measurable content in this fragments?

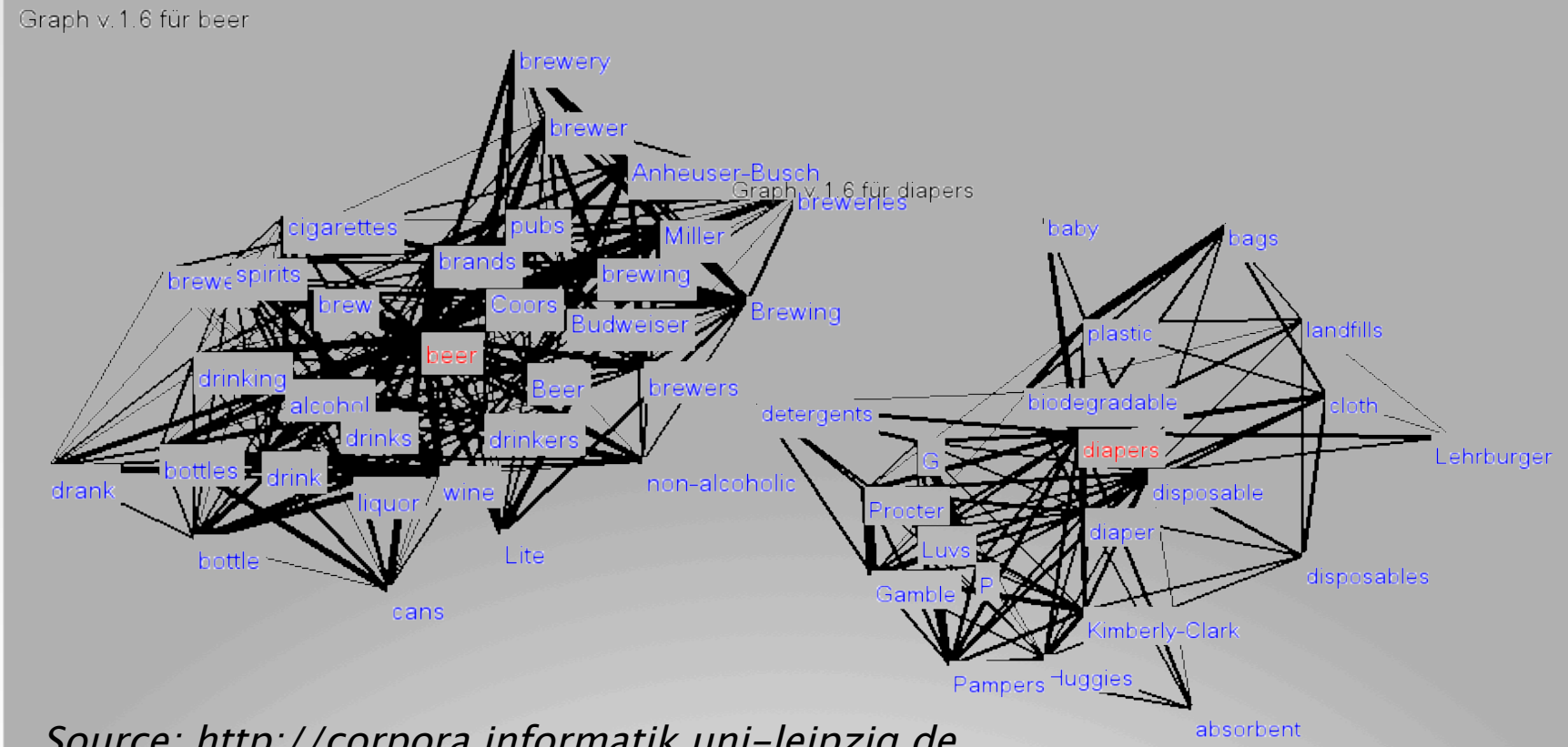
Definition/Motivation

- Definition Co-occurrences:
  - Common occurrence of at least two objects/events within a dedicated window
  - » Possible windows in Classical Studies: line, sentence, paragraph, document, author, century
- Motivation:
  - Psycholinguistic experiments: Given a word: What is the first word test persons answer?

Stimulus	Response Prob.	# of Prob.'s	Co-occurrence	Significance
butter	Bread	60	Bread	51
	soft	40	Cheese	49
	Milk	32	Sugar	29
	Margarine	27	Milk	23
	Cheese	20	Margarine	22
	Fat	16	Farina	18
	yellow	14	Eggs	16
	Bread and butter	8	Pound	14
	Box / can	6	Meat	13
	eat	6		

Overview of the approach (Humanists friendly)

- **Observation:**
  - Every word has a contextual profile (based on co-occurrences) in which it is typically used



Source: http://corpora.informatik.uni-leipzig.de

An example of Data Mining: Relation between beer and diapers I

There is a story that a large supermarket chain, usually Wal-Mart, did an analysis of customers' buying habits and found a statistically significant correlation between purchases of **beer** and purchases of nappies (**diapers** in the US). It was theorized that the reason for this was that fathers were stopping off at Wal-Mart to buy nappies for their babies, and since they could no longer go down to the pub as often, would buy beer as well. As a result of this finding, the supermarket chain is alleged to have the nappies next to the beer, resulting in increased sales of both.

Contrastive relation: (**beer**, **diapers**)

Source: http://web.onetel.net.uk/~hibou/Beer%20and%20Nappies.html

An example of Data Mining: Relation between beer and diapers II

There is a story that a large supermarket chain, usually Wal-Mart, did an analysis of customers' buying habits and found a statistically significant correlation between purchases of **beer** and purchases of **nappies** (**diapers** in the US). It was theorized that the reason for this was that fathers were stopping off at Wal-Mart to buy nappies for their babies, and since they could no longer go down to the pub as often, would buy beer as well. As a result of this finding, the supermarket chain is alleged to have the nappies next to the beer, resulting in increased sales of both.

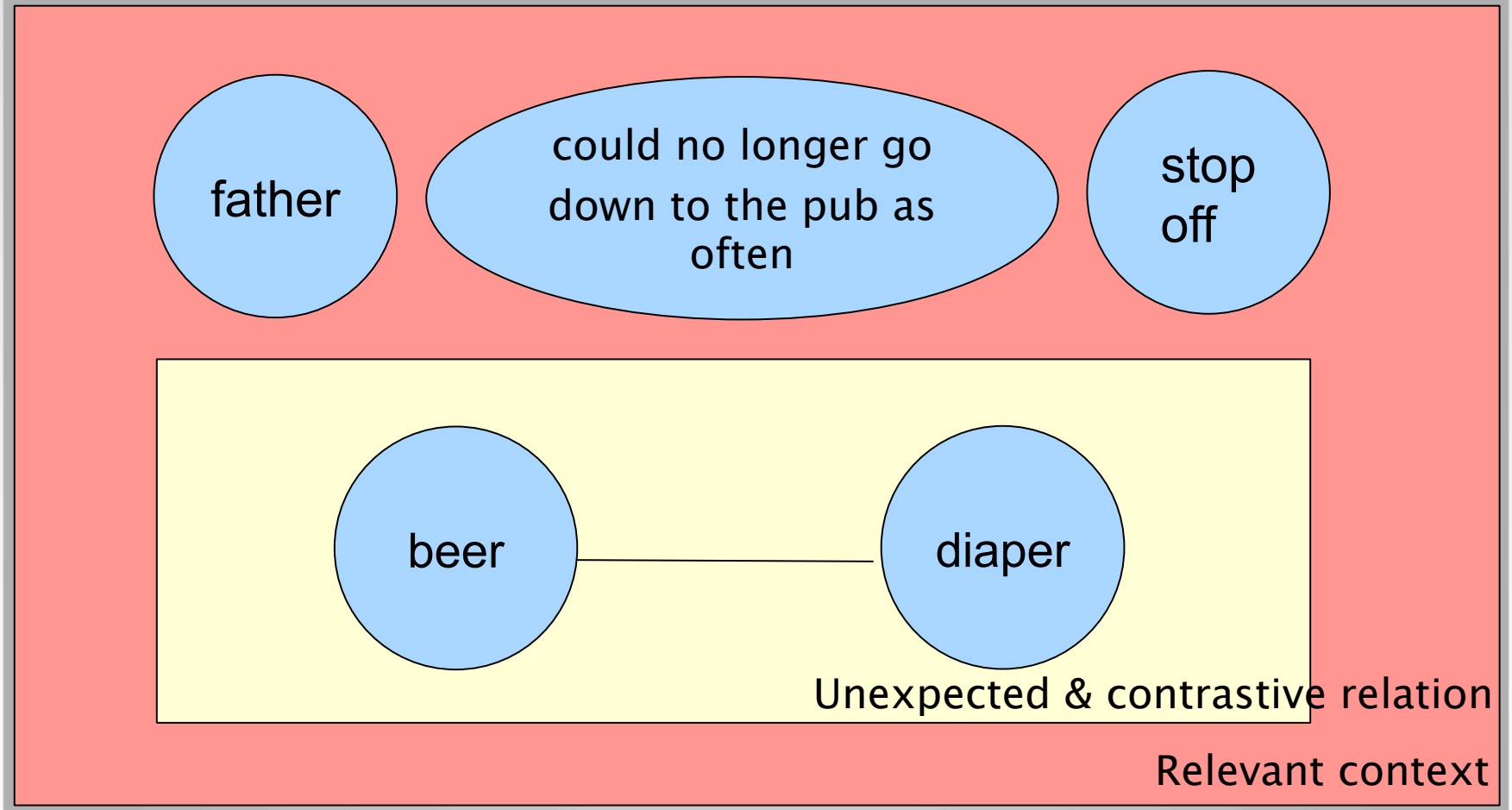
Latent relation: (**beer**, **diapers**)

Context: fathers, stopping off, Wal-Mart, could no longer go down to the pub as often

Result of this relation: nappies next to the beer

Source: http://web.onetel.net.uk/~hibou/Beer%20and%20Nappies.html

What should be the result?



Some examples (if the original text is still existent)

- Relation of (**Ὀδόμαντοι, πέος**)
  - engl.: (*Odomastai (a folk in Thrace), peni*s)
  - Context: Found in an Ancient comedy (Aristophanes, 5th c. BC)

- Relation of (**κοπρολόγος, ψάλτρια**) – engl.: (shit collector, *dancing girl*)
  - Context: ἀστυνόμοι – engl.: (*protecting the city, public festivals*)
  - Found in Aristotle (4th c. BC)

Results

- Lots of *contrastive semantic relations* can be found (manual evaluation is still in progress)

- But depending on text sort:
  - Other clusters can be found additionally
    - As shown in examples **comedy**
    - **Sarcasm**
    - **Cynicism**
    - **Artificial ambiguity** like „*Michael Schumacher* the red *king*“ (translated from a German corpus)
    - **Scope to gnomology & philosophical texts**

Contrastive semantic relations from a bird's eye view

- What did I do with the example of *beer* and *diaper*?
  - If I would write it down: A **semantically textual reference**.
- Is there a relation between **contrastive relations** and **textual reuse**?
  - Clearly, yes.
  - First evaluation results: More than 90% of the latent relations (Settings: minimum frequency: 2, Except the contrastive relation itself not more than 2 additional associations)

Focus:

Here: Why is knowledge reused?

Nobody would reuse something like: „Milk is white and good for you“.

Why: It's well-known.

Why was it reused?

'They drink *beer* made of *rye*, as the *Thracians* drink it made of *barley*'.

the Paeonians drink a beer made from barley, also *parabias*, made from millet, and even fleabane.

'They also anoint themselves', he says, 'with an oil made from milk'.

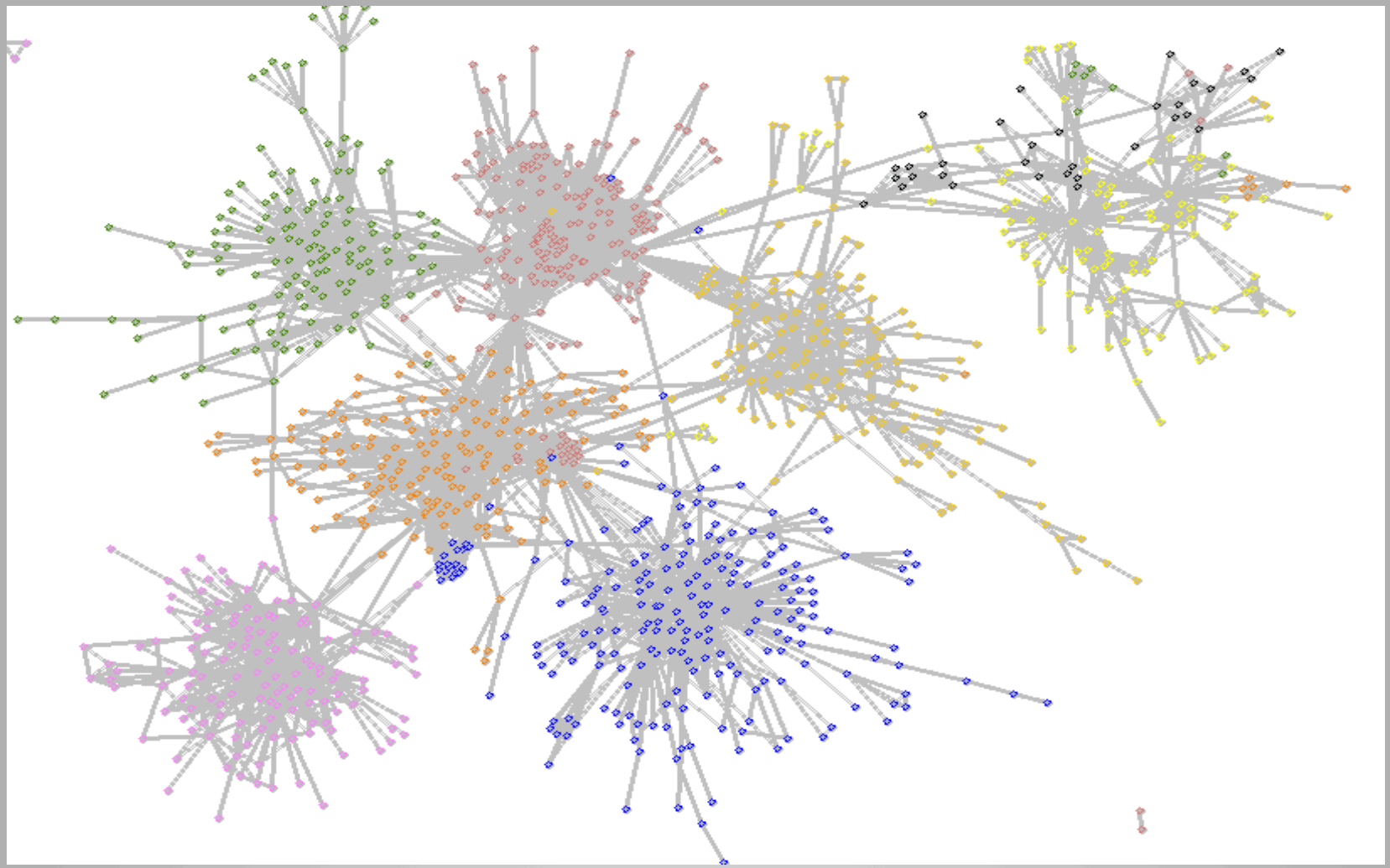
Dissimilarities in the contextual usage (TLG):

- **(milk,oil)**: 72%

- **(fleabane, millet)**: 92%, **(parabias, millet)**: 97%, **(fleabane, parabias)**: 94%, **(barley, fleabane)**: 94%, ...

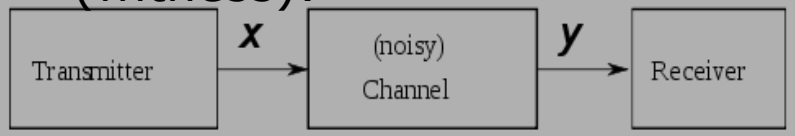
- **(rye, barley)**: 80%

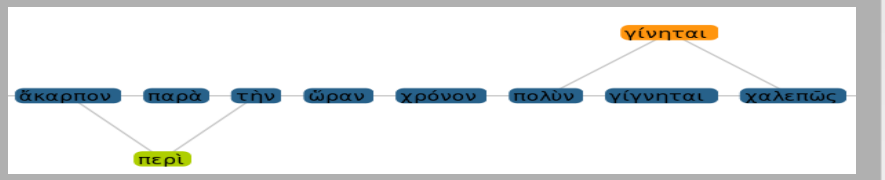
Further work: Semantic spaces



How can Marco benefit from work with Monica?

- **NEW QUESTION:** Shannon's Noisy Channel Theorem (witness):





- **NEW QUESTION**: Not HOW but why is something quoted?
  - Contrastive semantics

- **EVALUATION:** How to evaluate text reuse & knowledge transfer?
  - Collection of fragmentary authors as highly reviewed *Gold Standard*

Summary

To be, or not to be, that is the question
Hamlet, Shakespeare