UNIVERSITÀ
DEGLI STUDI DI TRIESTE

How long is a piece of string?

Concordance searches &
User behaviour investigated

Paola Valli
University of Trieste

paola.valli@phd.units.it

TC34 – London, 29-30 November 2012

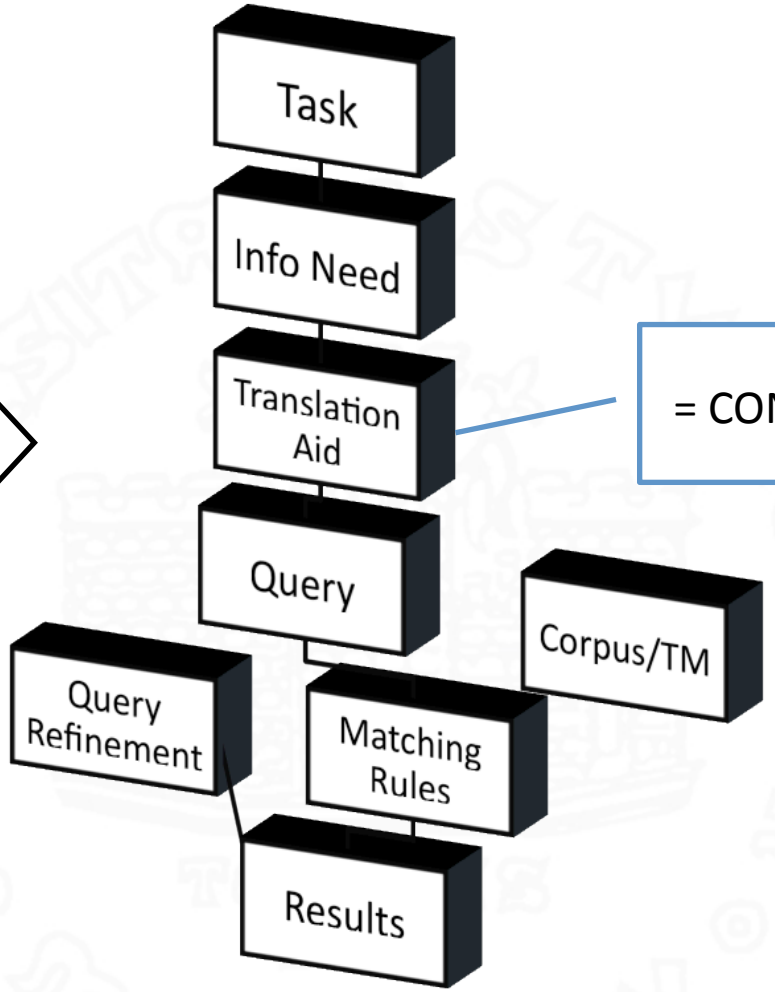
Translation 101

Translation Problems 101

INTERNAL SUPPORT

EXTERNAL SUPPORT

Concordance Search as Information Need



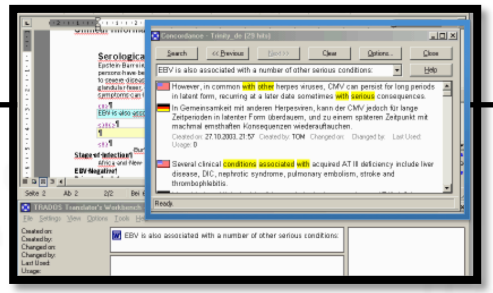
[IR model for Web Search]

(adapted from Broder 2002: 4)

Local vs. Cloud Solutions

(Concordancers)

OFF-LINE



INTERNET

EU INTRANET

Internet vs. EU intranet

MyMemory Translated.net

tradoit

TRANSSEARCH

WEBITEXT

INTERNET

Linguee

TAUS DATA Human Language Project

glosbe

IATE

Quest Metasearch

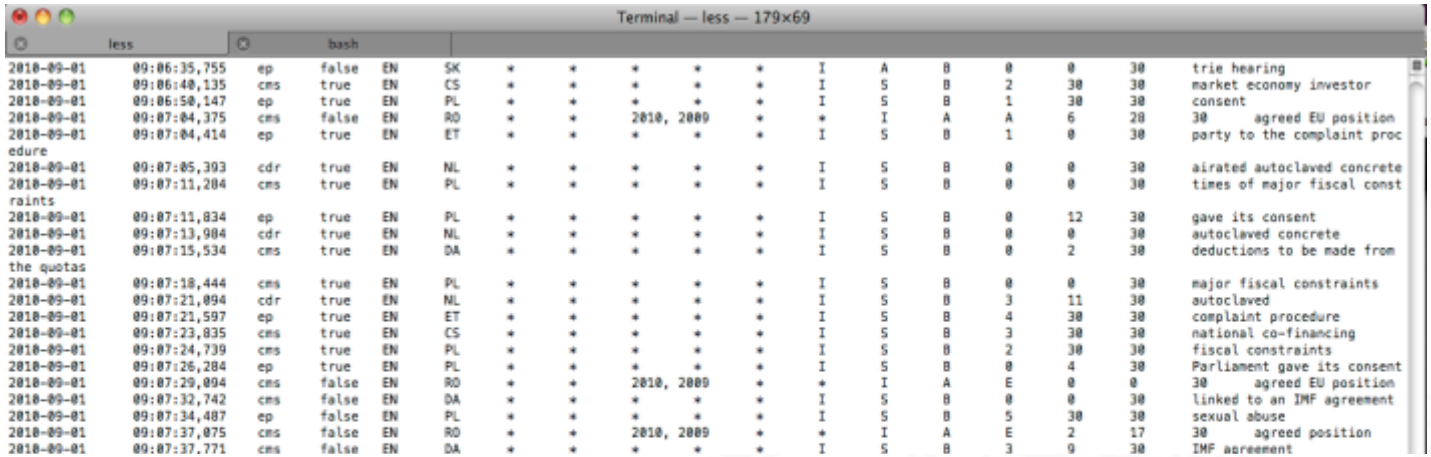
EUR-Lex

EU INTRANET

Euramis

DGT Net

The Euramis Dataset



- Total queries: ~ 724k
- Time span: 1 month (Sept 2010)
- Language pairs: EN > 20 EU languages
- Avg. lang. subset size: ~ 36k strings

Concordance Search logs vs. Web Queries

Three levels of analysis:

1. Session
2. Query
3. Term

#1 Session

- SESSION: = search episode, = info need, multiple interactions
- 4 requisites for session extraction
- Search sessions vs. Spot searches
- 36% search sessions (~ 13k strings/lang)
- Avg. session length: 2.27 (FR) – 2.59 (BG)

→ (WEB) Avg. session length: 2-4 queries

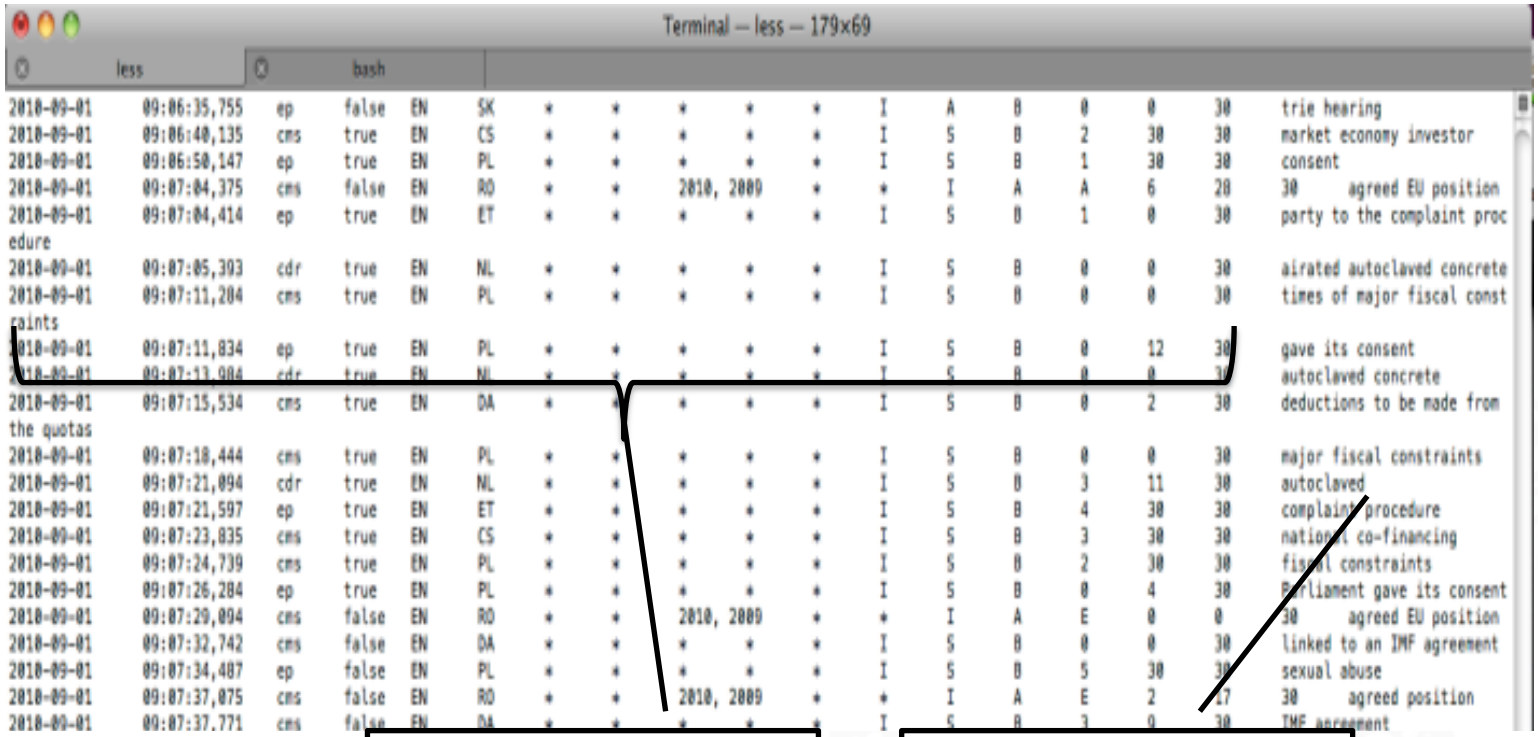
#1 Session (dynamic dimension)

- QUERY REFINEMENT: = need, ≠ strategy
- Taxonomy of reformulations:

A	RESUBMISSION		D	EXPANSION	
	A1. Repeated query	7.74%		*D1. Left expansion*	1.51%
	A2. Wildcards			*D2. Right expansion*	2.05%
B	FORMAL CHANGES			*D3. Middle expansion*	
	B1. Casing			*D4. Cross expansion*	
	B2. Punctuation			*D5. Addition of plural 's'*	
	B3. Locale		E	REPLACEMENT	
C	REDUCTION			*E1. Tense change*	
	C1. Left trim	25.62%		*E2. Paraphrasing*	
	C2. Right trim	23.05%		*E3. Synonym / Antonym*	
	C3. Middle trim	4.23%		*E4. Word substitution*	3.44%
	C4. Cross trim			*E5. Typo Fix*	
	C5. Plural / Genitive 's'	4.04%	F	MIXED STRATEGY	

- Reduction/recall vs. Expansion/precision (WEB)

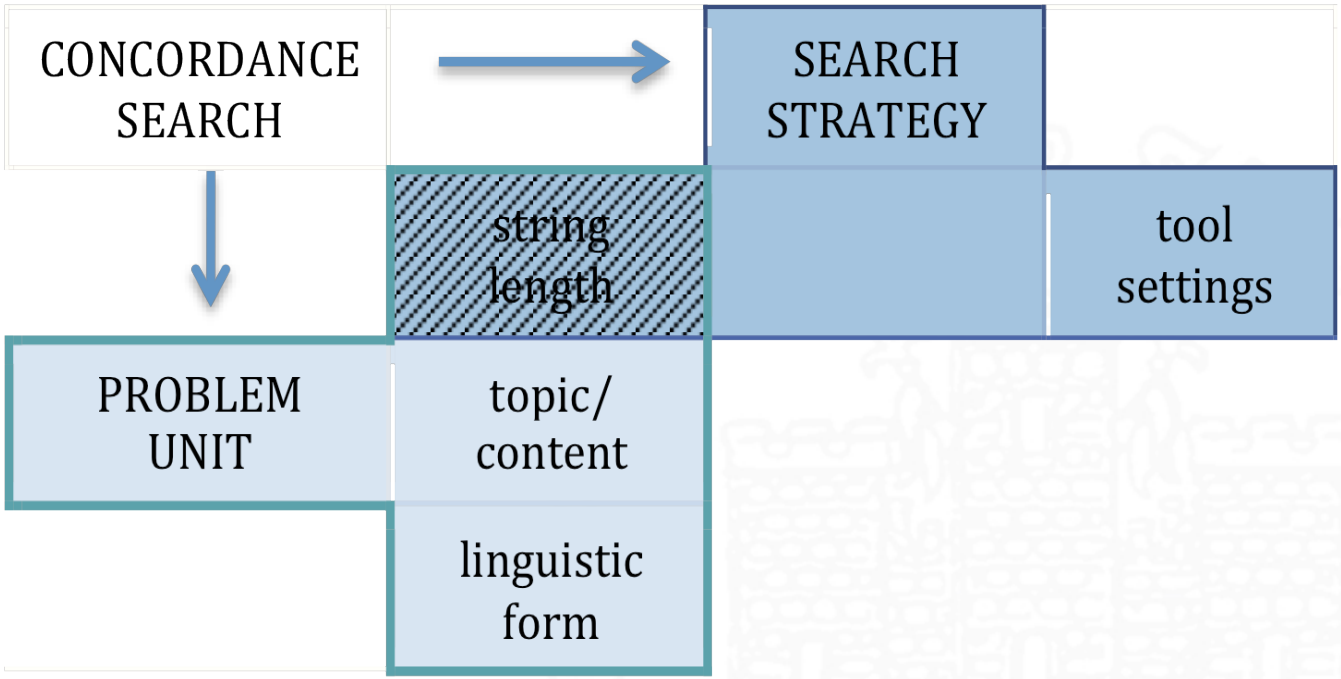
#2 Query (complex search)



HOW
(Search Strategy)

WHAT
(Problem Unit)
+ **HOW**

#2 Query (static dimension)



#2 Search Strategy (*How*)

- Avg. query length: 2-3 words

→(WEB) 2-3 words

	TransSearch ('08) (6 years /7.2m)	Euramis (1 month /724k)
Single-word queries	13.2%	13.83%
Two-word queries	39.6%	34.02%
Three-word queries	27.7%	20.33%
Four-word queries	13.0%	12.27%
Five-word queries	4.3%	6.66%
Six-word queries & above	2.2%	12.90%
Total	**100%**	**100%**

- Few settings and filters (20.6% adv. mode)

→(WEB) mostly basic searches

(Macklovitch et al. 2008)

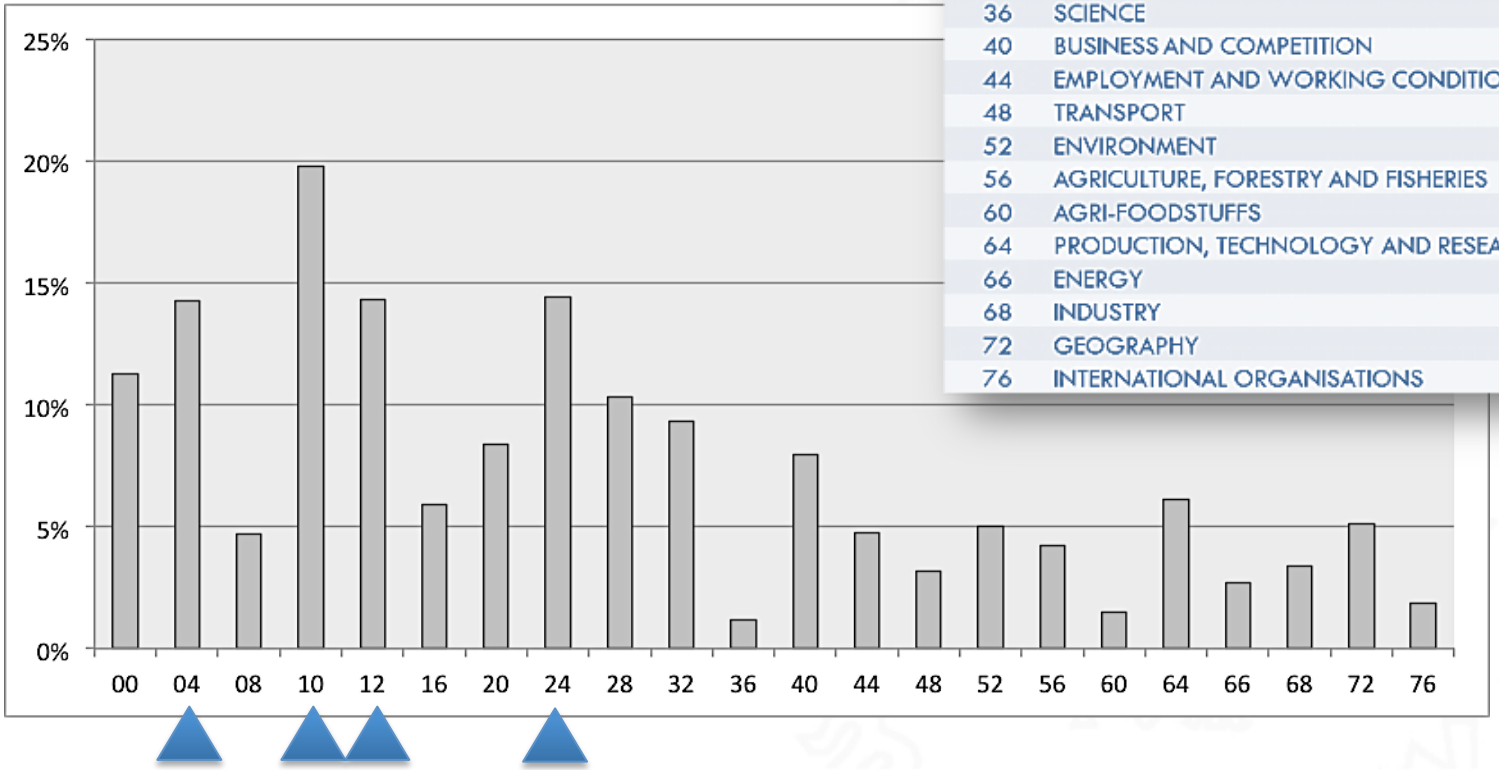
#2 Problem Unit (*What*)

- **Size**: Multi-Word Units/Search approach
- **Content**: Semi-automatic topical categorization (EuroVoc)
- **Linguistic form**: Syntactic perspective

#2 Query – Content Analysis

LGP: 42%
LSP: 58%
multiple matches: 36%
(avg. of 20 lang.)

DOMAIN	LABEL
00	**EUROJARGON**
04	**POLITICS**
08	INTERNATIONAL RELATIONS
10	**EUROPEAN COMMUNITIES**
12	**LAW**
16	ECONOMICS
20	TRADE
24	**FINANCE**
28	SOCIAL QUESTIONS
32	EDUCATION AND COMMUNICATIONS
36	SCIENCE
40	BUSINESS AND COMPETITION
44	EMPLOYMENT AND WORKING CONDITIONS
48	TRANSPORT
52	ENVIRONMENT
56	AGRICULTURE, FORESTRY AND FISHERIES
60	AGRI-FOODSTUFFS
64	PRODUCTION, TECHNOLOGY AND RESEARCH
66	ENERGY
68	INDUSTRY
72	GEOGRAPHY
76	INTERNATIONAL ORGANISATIONS



#3 Term (≠ Term!)

- Term = string of characters (~ token)
- 1-gram analysis:
  - 27 one-grams in top 100 most searched strings
  - 13 of which acronyms
- Acronyms as a type of Named Entity?
- if NE = acronym + spelled-out form:

→ #9 'ERDF' (155) +

#29 'european regional development fund' (126)

#3 Named Entity (Title)

Catherine
Ashton

Vice-President

High Representative of the Union for Foreign Affairs and Security Policy

source: http://ec.europa.eu/commission_2010-2014/index_en.htm

Freq	# w	String
83	2	high representative
5	3	eu high representative
2	3	high representative ashton
1	3	ashton high representative
2	3	high representative vp
1	3	high representative cfsp
21	3	high representative vice-president (vice president/vicepresident)
1	4	high representative catherine asthon
6	4	high representative vice-president ashton
1	4	eu high representative vice-president
1	4	vice-president-high representative ashton
3	4	high representative for the cfsp
1	4	vice-president-high representative
4	4	high representative and vice-president
2	5	eu high representative vice-president ashton
1	5	catherine ashton the high representative
2	5	high representative for foreign affairs
1	5	high representative for foreign relations
1	5	eu high representative catherine ashton
3	5	high representative and vice-president ashton
6	6	high representative of the european union
2	6	high representative vp of the commission
1	6	vice-president of the commission high representative
9	6	high representative vice-president of the commission
8	6	high representative of the union (for)
3	7	high representative for foreign and security policy
27	8	high representative for foreign affairs and security policy
1	8	high representative for foreign relations and security policy
4	8	high representative of the union for foreign affairs
1	9	vice president of the european commission and high representative
3	9	high representative for the common foreign and security policy
1	9	high representative of the european union for foreign affairs
6	9	eu high representative for foreign affairs and security policy
1	10	high representative for foreign affairs and security policy vice president
1	11	catherine ashton eu high representative for foreign affairs and security policy
6	11	european union high representative for the common foreign and security policy
53	11	high representative of the union for foreign affairs and security policy
3	11	high representative of the eu for foreign affairs and security policy
2	11	eu high representative for foreign affairs and security policy commission vice-president
1	12	eu high representative of the union for foreign affairs and security policy
3	12	the high representative of the union for foreign affairs and security policy
17	12	high representative of the european union for foreign affairs and security policy
1	12	commission high representative of the union for foreign affairs and security policy
7	12	vice-president of the commission high representative of the union for foreign affairs
3	13	eu high representative of the european union for foreign affairs and security policy
1	13	the high representative for foreign affairs and security policy vice-president of the commission
1	14	the high representative of the union for foreign affairs and security policy and vice-president
1	15	high representative of the union for foreign affairs and security policy vice-president of the commission
3	15	vice president of the european commission and high representative for foreign affairs and security policy
13	15	vice-president of the commission high representative of the union for foreign affairs and security policy
2	16	eu vice president of the european commission and high representative for foreign affairs and security policy
3	16	vice-president of the commission and high representative of the union for foreign affairs and security policy
3	16	high representative of the union for foreign affairs and security policy and vice-president of the commission
3	17	eu high representative of the union for foreign affairs and security policy and vice-president of the commission

56!!

[tot_fr = 353, avg_fr = 5.98, range_fr 1-83] - [avg_len = 8.1 w, range_len 2-17 w]

Lessons Learned

- Broadly similar to Web queries
- Some specificities due to translation task (e.g. search aim)
- Frequency lists to identify problems for which alternative forms of support are lacking (e.g. acronyms & NE)
- Need to broaden our understanding of NEs?
- Old & new challenges for effective NE identification for translation purposes

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Thank you!

paola.valli@phd.units.it