



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

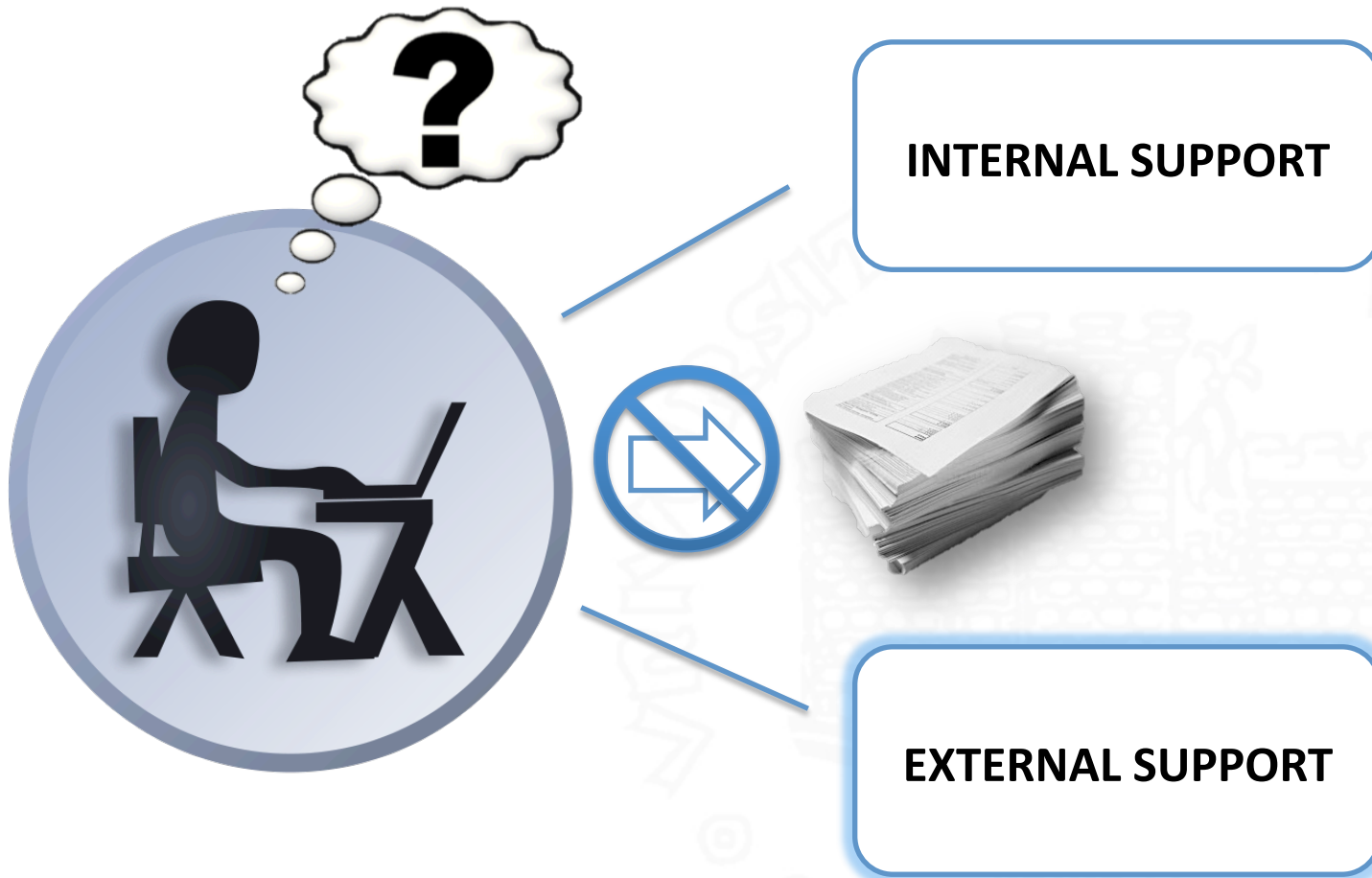


Translation 101



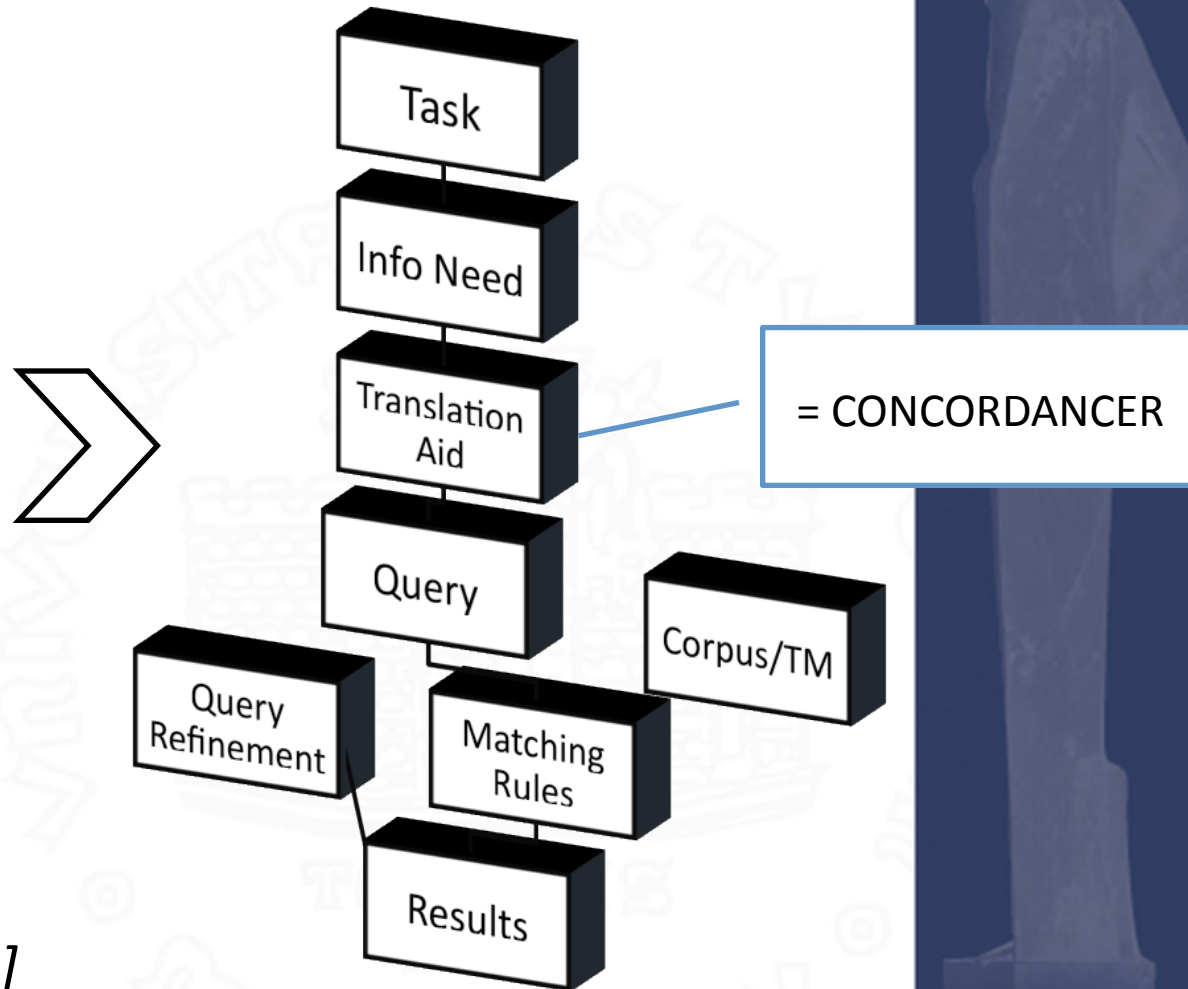


Translation Problems 101





Concordance Search as Information Need

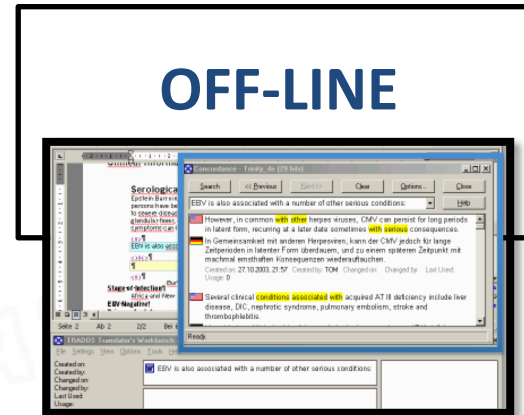


[IR model for Web Search]

(adapted from Broder 2002: 4)

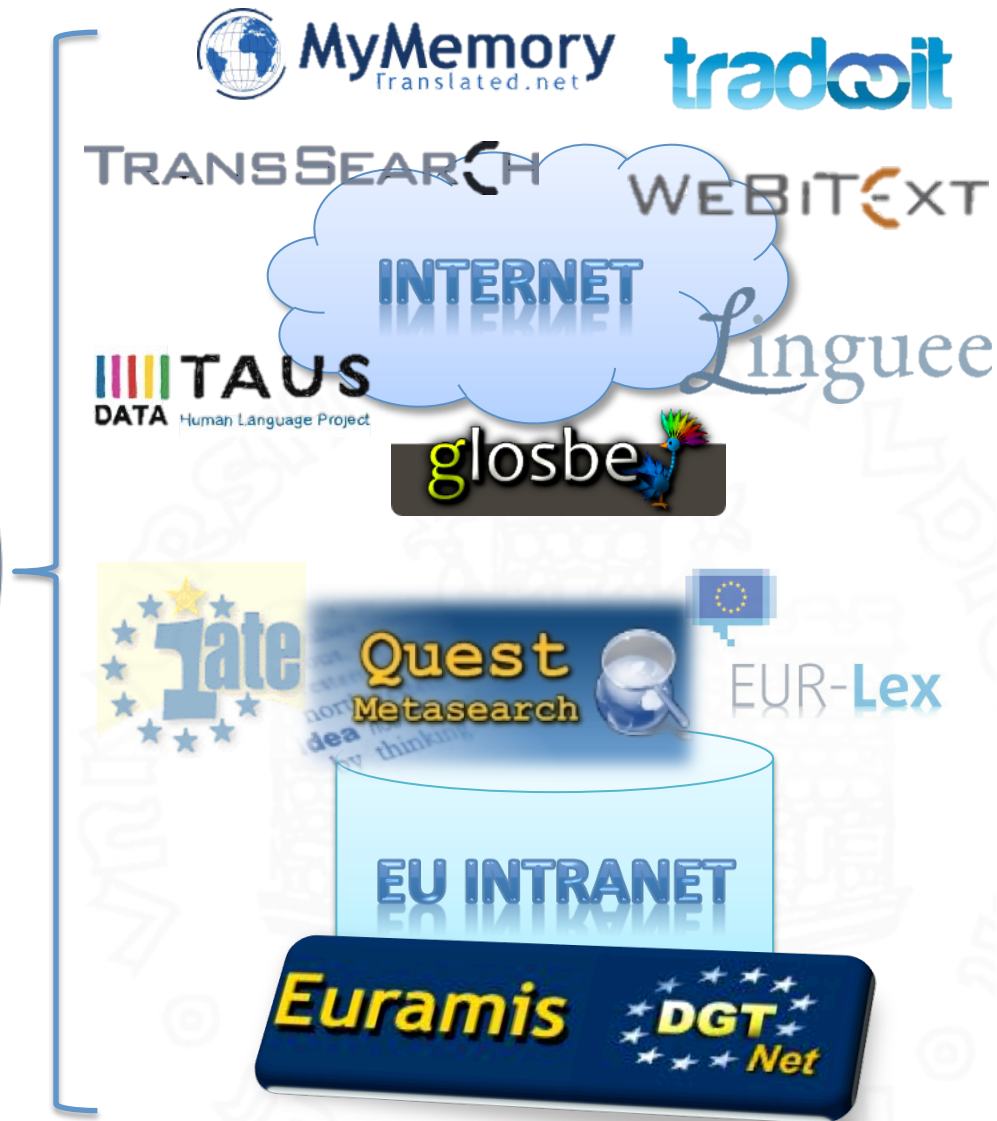


Local vs. Cloud Solutions (Concordancers)





Internet vs. EU intranet





The Euramis Dataset

```
Terminal — less — 179x69
```

Timestamp	Time	Command	Success	Lang	Lang2
2010-09-01	09:06:35,755	ep	false	EN	SK	*	*	*	*	I	A	B	0	0	30	trie hearing
2010-09-01	09:06:40,135	cms	true	EN	CS	*	*	*	*	I	S	B	2	30	30	market economy investor consent
2010-09-01	09:06:50,147	ep	true	EN	PL	*	*	*	*	I	S	B	1	30	30	30 agreed EU position
2010-09-01	09:07:04,375	cms	false	EN	RO	*	*	*	*	I	A	A	6	20	20	party to the complaint proc
2010-09-01	09:07:04,414	ep	true	EN	ET	*	*	*	*	I	S	B	1	0	30	airated autoclaved concrete
2010-09-01	09:07:05,303	cdr	true	EN	NL	*	*	*	*	I	S	B	0	0	30	times of major fiscal const
2010-09-01	09:07:11,204	cms	true	EN	PL	*	*	*	*	I	S	B	0	0	30	gave its consent
2010-09-01	09:07:11,834	ep	true	EN	PL	*	*	*	*	I	S	B	0	12	30	autoclaved concrete
2010-09-01	09:07:13,904	cdr	true	EN	NL	*	*	*	*	I	S	B	0	0	30	deductions to be made from
2010-09-01	09:07:15,534	cms	true	EN	DA	*	*	*	*	I	S	B	0	2	30	major fiscal constraints
2010-09-01	09:07:18,444	cms	true	EN	PL	*	*	*	*	I	S	B	0	0	30	autoclaved
2010-09-01	09:07:21,094	cdr	true	EN	NL	*	*	*	*	I	S	B	3	11	30	complaint procedure
2010-09-01	09:07:21,597	ep	true	EN	ET	*	*	*	*	I	S	B	4	30	30	national co-financing
2010-09-01	09:07:23,835	cms	true	EN	CS	*	*	*	*	I	S	B	3	30	30	fiscal constraints
2010-09-01	09:07:24,739	cms	true	EN	PL	*	*	*	*	I	S	B	2	30	30	Parliament gave its consent
2010-09-01	09:07:26,204	ep	true	EN	PL	*	*	*	*	I	S	B	0	4	30	30 agreed EU position
2010-09-01	09:07:29,004	cms	false	EN	RO	*	*	*	*	I	A	E	0	0	0	linked to an IMF agreement
2010-09-01	09:07:32,742	cms	false	EN	DA	*	*	*	*	I	S	B	0	0	30	sexual abuse
2010-09-01	09:07:34,407	ep	false	EN	PL	*	*	*	*	I	S	B	5	30	30	30 agreed position
2010-09-01	09:07:37,075	cms	false	EN	RO	*	*	*	*	I	A	E	2	17	17	IMF agreement
2010-09-01	09:07:37,771	cms	false	EN	DA	*	*	*	*	I	S	B	3	0	30	

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

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



#2 Query (complex search)

Terminal — less — 179x69

less	bash	
2018-09-01 09:06:35,755	ep false EN SK	* * * * * I A 0 0 0 30 trie hearing
2018-09-01 09:06:40,135	cms true EN CS	* * * * * I S 0 2 30 30 market economy investor
2018-09-01 09:06:50,147	ep true EN PL	* * * * * I S 0 1 30 30 consent
2018-09-01 09:07:04,375	cms false EN RO	* * 2010, 2009 * * * I A A 6 20 30 agreed EU position
2018-09-01 09:07:04,414	ep true EN ET	* * * * * I S 0 1 0 30 party to the complaint procedure
2018-09-01 09:07:05,303	cdr true EN NL	* * * * * I S 0 0 0 30 aired autoclaved concrete
2018-09-01 09:07:11,204	cms true EN PL	* * * * * I S 0 0 0 30 times of major fiscal constraints
2018-09-01 09:07:11,834	ep true EN PL	* * * * * I S 0 0 12 30 gave its consent
2018-09-01 09:07:13,004	cdr true EN NL	* * * * * I S 0 0 0 30 autoclaved concrete
2018-09-01 09:07:15,534	cms true EN DA	* * * * * I S 0 0 2 30 deductions to be made from the quotas
2018-09-01 09:07:18,444	cms true EN PL	* * * * * I S 0 0 0 30 major fiscal constraints
2018-09-01 09:07:21,094	cdr true EN NL	* * * * * I S 0 3 11 30 autoclaved
2018-09-01 09:07:21,597	ep true EN ET	* * * * * I S 0 4 30 30 complaint procedure
2018-09-01 09:07:23,035	cms true EN CS	* * * * * I S 0 3 30 30 national co-financing
2018-09-01 09:07:24,739	cms true EN PL	* * * * * I S 0 2 30 30 fiscal constraints
2018-09-01 09:07:26,284	ep true EN PL	* * * * * I S 0 0 4 30 Parliament gave its consent
2018-09-01 09:07:29,094	cms false EN RO	* * 2010, 2009 * * * I A E 0 0 30 agreed EU position
2018-09-01 09:07:32,742	cms false EN DA	* * * * * I S 0 0 0 30 linked to an IMF agreement
2018-09-01 09:07:34,407	ep false EN PL	* * * * * I S 0 5 30 30 sexual abuse
2018-09-01 09:07:37,075	cms false EN RO	* * 2010, 2009 * * * I A E 2 17 30 agreed position
2018-09-01 09:07:37,771	cms false EN DA	* * * * * I S 0 1 0 30 IMF agreement

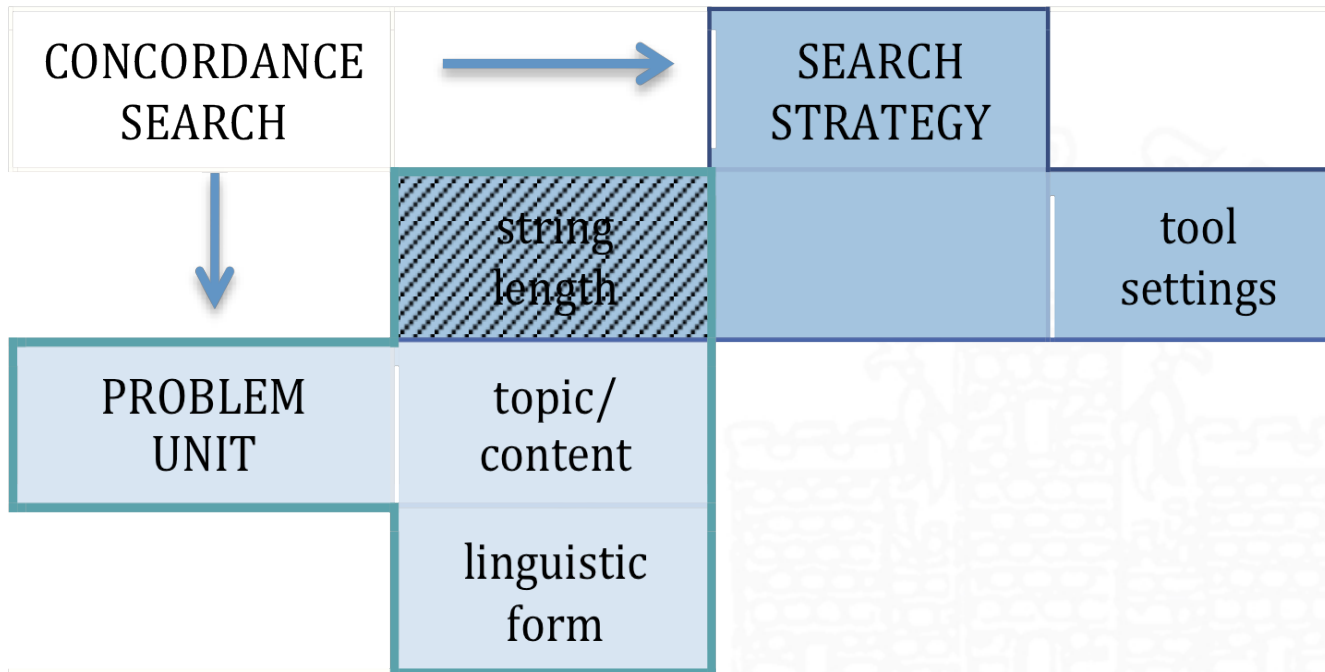
HOW
(Search Strategy)

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%
<i>Total</i>	100%	100%

- Few settings and filters (20.6% adv. mode)
→(WEB) mostly basic searches



#2 Problem Unit (*What*)

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





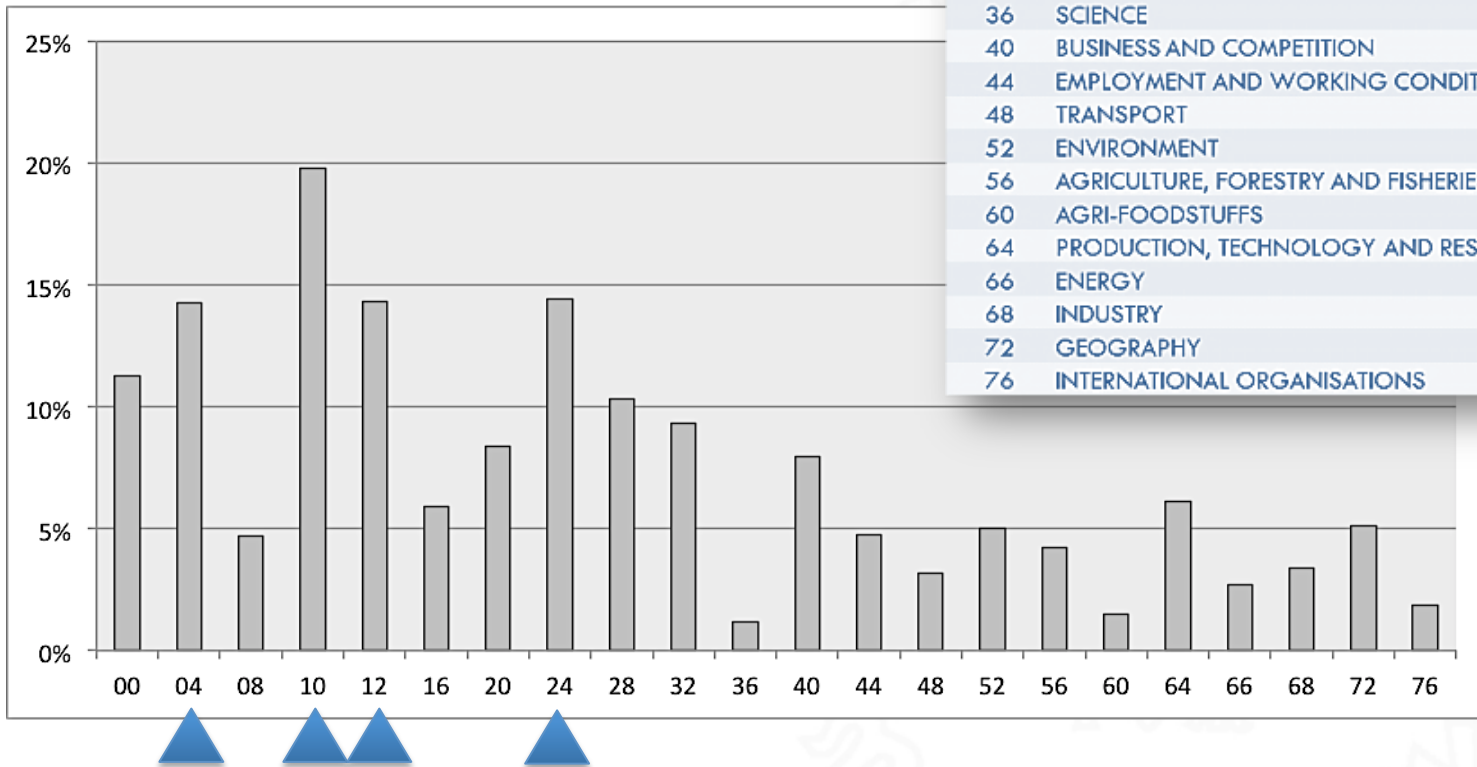
#2 Query – Content Analysis

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

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



[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

