



Evaluation of HLT-Tools for Less Spoken Languages

Cristina Vertan

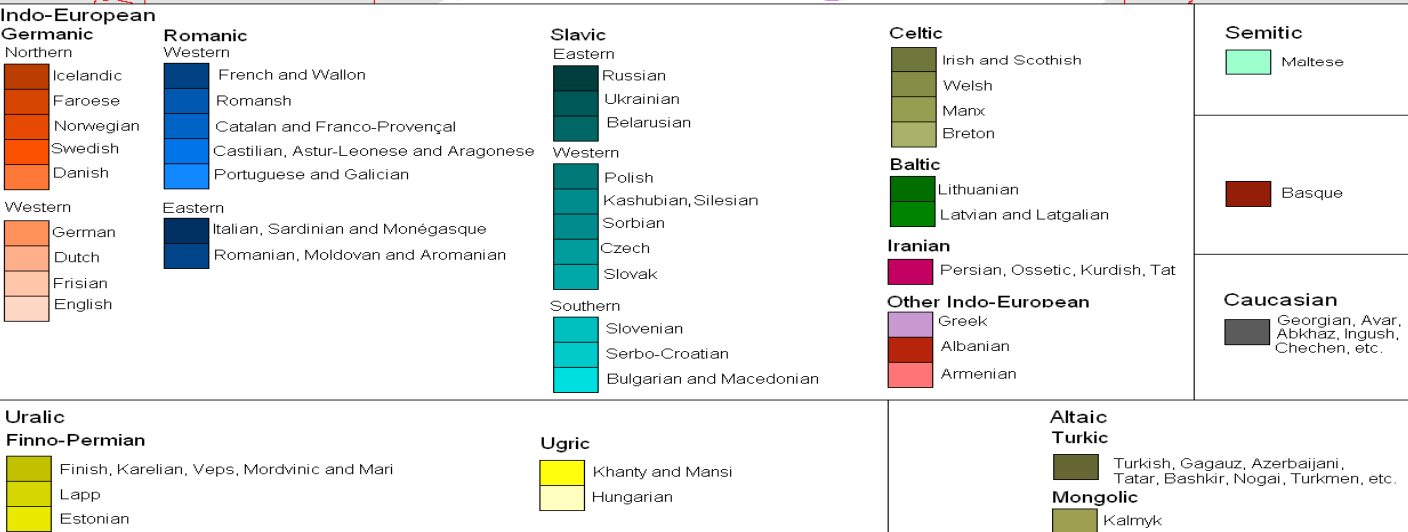
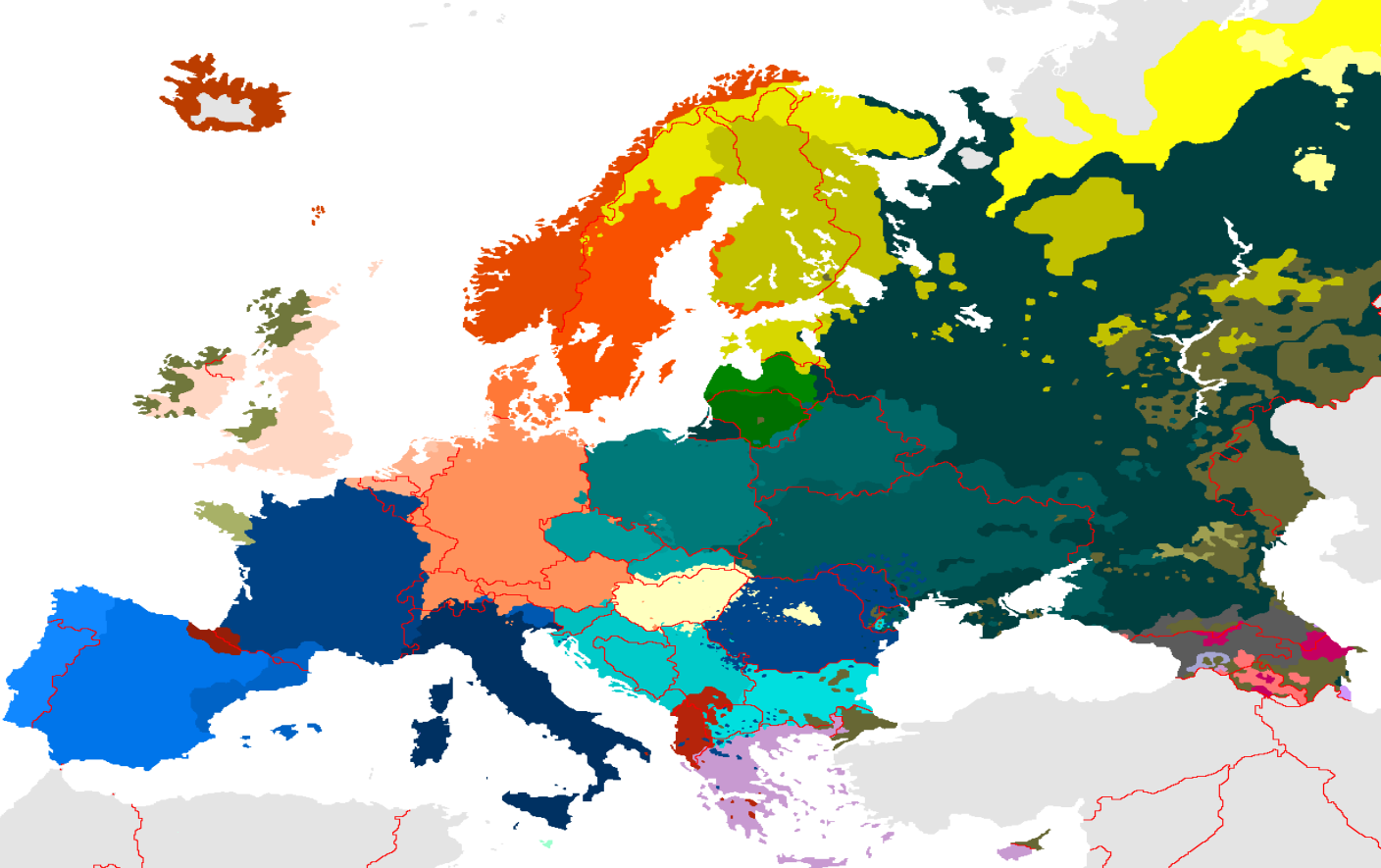
cristina.vertan@uni-hamburg.de

Rationale

- Europe offers an unique context
 - not only due to the variety of languages
 - but also due to the necessity to offer a huge amount of information in each native language (EU)
- The number of mono- and multilingual systems dealing with other languages than the most frequent used ones (EN, FR, DE) is increasing.
- BUT: Most part of these systems are:
 - Developed for a particular scenario
 - Tuned on specific (available) corpora
- AS a RESULT: It is very difficult to evaluate such systems, as there are no:
 - Reference test data
 - Reference evaluation scores

...AND

How to build reference systems and comparable test data for all these languages and their combinations ?



© Wikipedia

Possible Solution

- One cannot build reference systems for all types of NLP tasks and all language combinations

BUT

- We can ensure a reference parallel corpora
- Reference test suites
- Define reference measures for classical NLP Applications (MT, IR, CLIR, QA, etc.) – underway.
- Develop when possible language independent system models (like Moses in MT)

Existent Corpora -1-

- ***JRC-Acquis Multilingual Parallel Corpus***
(<http://langtech.jrc.it/JRC-Acquis.html>)
 - 23rd languages and 231 language pairs

BUT:

- „Community law + acts adopted under the second and third pillars of the EU and the common objectives laid down in the treaties“

Systems trained on JRC-Acquis will produce qualitative lower results when tested on another domain, not because the method is wrong but because the vocabulary is different

Language ISO code	N° of texts	Text body			Signatures	Annexes	Total No words (text + signatures + annexes):
		Total No words	Total No characters	Average No words	Total No words	Total No words	
bg	11384	16140819	104522671	1417.85	2170075	14114612	32425506
cs	21438	22843279	148972981	1065.55	7225300	16763733	46832312
da	23624	31459627	213468135	1331.68	2629786	16855213	50944626
de	23541	32059892	232748675	1361.87	2542149	16327611	50929652
el	23184	36453749	239583543	1572.37	2973574	16459680	55887003
en	23545	34588383	210692059	1469.03	3198766	17750761	55537910
es	23573	38926161	238016756	1651.3	3490204	19716243	62132608
et	23541	24621625	192700704	1045.9	1336051	14995748	40953424
fi	23284	24883012	212178964	1068.67	2677798	12547171	40107981
fr	23627	39100499	234758290	1654.91	3021013	19978920	62100432
hu	22801	28602380	213804614	1254.44	2529488	15056496	46188364
it	23472	35764670	230677013	1523.72	3120797	18331535	57217002
lt	23379	26937773	199438258	1152.22	2436585	15018484	44392842
lv	22906	27592514	196452051	1204.6	1673124	15437969	44703607
mt	10545	20926909	128906748	1984.53	1336042	15620611	37883562
nl	23564	35265161	231963539	1496.57	3039580	18467115	56771856
pl	23478	29713003	214464026	1265.57	2513141	17027393	49253537
pt	23505	37221668	227499418	1583.56	3034308	19350227	59606203
ro	6573	9186947	60537301	1397.68	514296	11185842	20887085
(ro-19211)	19211
sk	21943	26792637	179920434	1221.01	3227852	16190546	46211035
sl	20642	27702305	178651767	1342.04	3103193	16837717	47643215
sv	20243	29433037	199004401	1453.99	2575771	14965384	46974192
Total	463792	636216050	4288962348	1387.23	50068893	358999011	1055583954

Existent Corpora -2-

- OPUS Corpus (<http://urd.let.rug.nl/tiedeman/OPUS/>)
 - Medical texts
 - Operating systems manuals
 - Good language coverage although less than Acquis.

(src)="s11.3"> Die Wirksamkeit von ABILIFY in Dosen von 10 bis 30 mg täglich wurde mit der eines Placebos (einer Scheinbehandlung) verglichen , indem die Veränderung der Symptome der Patienten anhand einer Standardskala für die Schizophrenie gemessen wurde .

(trg)="s10.3"> Eficacitatea ABILIFY în doză de 10 până la 30 mg pe zi a fost comparată cu cea a unui placebo (un preparat inactiv) prin măsurarea schimbării simptomatologiei pacienților , utilizând o scală standard pentru schizofrenie .

(src)="s1.1"> EU Number

(trg)="s1.1"> Numărul UE EU/ 1/ 04/ 276/ 001 EU/ 1/ 04/ 276/ 002 EU/ 1/ 04/ 276/ 003 EU/ 1/ 04/ 276/ 004 EU/ 1/ 04/ 276/ 005 EU/ 1/ 04/ 276/ 006 EU/ 1/ 04/ 276/ 007 EU/ 1/ 04/ 276/ 008 EU/ 1/ 04/ 276/ 009 EU/ 1/ 04/ 276/ 010 EU/ 1/ 04/ 276/ 011 EU/ 1/ 04/ 276/ 012 EU/ 1/ 04/ 276/ 013 EU/ 1/ 04/ 276/ 014 EU/ 1/ 04/ 276/ 015 EU/ 1/ 04/ 276/ 016 EU/ 1/ 04/ 276/ 017 EU/ 1/ 04/ 276/ 018

Other frequently used corpora

- Europarl (11 Languages)
- MULTEXT-EAST (<http://nl.ijs.si/ME/CD/docs/mte-corp.html>)
- Small corpora developed in local projects projects (e.g. ROGER (Romanian - German - English - Russian corpus, aligned at sentence level, 2333 sentences ©NATS, University of Hamburg)

The problem of datas-et profiling -monolingual

Length of text	AP	DOE	FR	PAT	SJM	WSJ	ZF	TIPSTER OVERALL
100	1.333	1.515	1.492	1.315	1.428	1.282	1.47	1.405
200	1.626	1.562	1.666	1.538	1.612	1.55	1.68	1.605
400	1.877	1.762	2.051	2.259	1.869	1.886	1.941	1.949
800	2.144	2.067	2.572	3.065	2.035	2.072	2.305	2.323
1600	2.797	2.315	3.047	4.266	2.476	2.584	2.758	2.892
3200	3.062	2.824	3.841	5.169	3.013	3.225	3.285	3.488
6400	3.561	3.575	5.437	6.009	3.557	3.83	4.238	4.315
16000	4.563	4.737	8.583	9.744	4.153	4.566	5.289	5.948
20000	4.972	5.196	9.199	11.031	4.463	4.988	5.383	6.462
100000	9.14	10.451	15.453	20.764	8.463	9.413	12.017	12.243
1000000	30.573	30.157	50.571	62.637	26.377	30.909	38.105	38.476
10000000	106.845	94.778	144.866	134.017	102.149	116.183	121.798	117.234

Type to token ratio (©A.deRoeck &Al. 2007)

The problem of data-set profiling -multilingual

TIPSTER OVERALL	OU	Bengali	Arabic	Brown Corpus	Length of text
1.405	1.47	1.204819	1.19	1.449	100
1.605	1.694	1.388889	1.342	1.613	200
1.949	2.247	1.67364	1.423	2.424	400
2.323	2.622	1.864802	1.578	2.439	800
2.892	3.053	2.288984	1.774	2.576	1600
3.488	3.673	2.775369	2.082	3.674	3200
4.315	4.312	3.309204	2.357	4.702	6400
5.948	6.24	4.663363	2.771	5.928	16000
6.462	6.944	5.20969	2.875	6.341	20000
12.243	12.41	6.074628			100000
38.476	36.127	10.81093	8.252	20.408	1000000
117.234	82.064				10000000

Type to token ratio (©A.deRoeck &Al. 2007)

Recent Evaluations campaigns

- CLEF (including all subsections) - http://www.clef-campaign.org/2009/2009_agenda.html
- EUROMATRIX <http://www.euromatrix.net/>
- Very often languages / language pairs cannot be involved simply because the training and test corpora are not there

Steps for building a parallel corpus

- Unify annotation for existent parallel corpora: JRC-Acquis, OPUS, Europarl, MULTEX., MULTEX-EAST
- Investigate all EU projects with participation from various language communities Networks of Excellence. Especially in non-technical domains project reports are translated in all participant languages (e.g. CALIMERA <http://www.calimera.org/default.aspx>, I*Teach <http://i-teach.fmi.uni-sofia.bg/>)
- EU Web site at least part of the information is made available in all languages
- Collect parallel corpora produced in different projects

Roadmap -proposal

- Define a standard for minimal parallel corpora annotation (to be used as input data)
- development of a reference parallel corpus in all (most part) of the accepted EU languages.
- Releasing of test suites (parts of this corpus) to be used in test scenarios.
- Define standard test scenarios for evaluating (LT-systems) on these test-suites.
- A dataset profiling study, considering influence of language distribution on the statistical mechanisms for all official EU-languages and major HLT applications
- Implementation of baseline MT-systems using Moses and testing on the reference corpus