

# What kinds of geographical information are there in the Portuguese Web?

Marcirio Silveira Chaves<sup>1</sup>

Diana Santos<sup>2</sup>

Linguateca

<sup>1</sup>Node of XLDB at University of Lisbon

<sup>2</sup>Node of Oslo at SINTEF ICT

## Objective

To present some results on the **geographical information** in the Portuguese web and the **overlap with people's and organization's named entities**, using a **geographic ontology** based on **authoritative sources** and a named entity recognizer.

## Context and Available Resources

- Geographic Knowledge Base (**GKB**) – geographical information about Portugal exported as an ontology [**Geo-Net-PT01**([xldb.fc.ul.pt/geonetpt](http://xldb.fc.ul.pt/geonetpt))]
- **WPT 03** - Portuguese web collection with 12 Gbytes and 3.7 million pages comprising 1.6 billion words. 68.6 % of these pages are in Portuguese. ([www.linguateca.pt/WPT03](http://www.linguateca.pt/WPT03))
- **SIEMÊS** - named entity recognition (NER) system

## Getting to Know Better the Geographic References

NEs detected in a 32,000 documents sample of WPT 03						
MW: multi-word						
DNEs: distinct named entities						
GN: Geo-Net-PT01						
	# of NEs (%)	# of DNEs	# of MW NEs (%)	# of MW DNEs (%)	Overlap	
					WPT 03 and Geo-Net-PT01	
					# of MW DNEs containing a name in GN (%)	# of DNEs occurring in GN (%)
PEO	250,585 (26.48)	77,228	140,155 (55.93)	58,991 (76.39)	<b>24,105 (31.21)</b>	521 (0.67)
ORG	418,915 (44.27)	114,353	214,698 (51.25)	89,790 (78.52)	<b>26,789 (23.43)</b>	462 (0.40)
LOC	276,775 (29.25)	47,972	90,018 (32.52)	36,395 (75.87)	22,959 (47.86)	<b>4,576 (9.53)</b>
Sum	<b>946,275 (100.00)</b>	<b>239,553</b>	444,871 (47.01)	185,176 (77.30)	73,853 (30.83)	5,559 (2.32)

### Has the kind of location occurring in Portuguese web texts different properties?

Geographical information can be found in (almost) all sorts of texts.

Distribution of the types contained in the local (LOC) category

Distribution of NEs per document

Type	# of DNEs (%)	# of MW DNEs (%)
POV (names of pop. places)	<b>33,827 (70.51)</b>	<b>24,037 (71.06)</b>
ENDRALAR (full address)	3,505 (7.31)	3,313 (94.52)
SOCCUL (society/culture)	3,474 (7.24)	3,161 (90.99)
PAIS (country)	1,987 (4.14)	1,419 (71.41)
RLG (religion)	1,197 (2.50)	1,113 (92.98)
Other (Σ11 types)	3,982 (8.30)	3,352 (84.18)
Sum	<b>47,972 (100,00)</b>	<b>36,395 (75,87)</b>

	Total	Distinct		Total	Distinct
Avg. PEOs per doc. with PEOs	11.65	7.82	<b>Median LOCs</b>	4	3
Avg. ORGs per doc. with ORGs	13.81	9.78	Stdev LOCs	149.7	57.54
<b>Avg. LOCs per doc. with LOCs</b>	11.31	<b>7.34</b>	# docs. with 1 LOC	5,443	6,184
Avg. NEs per doc. with NEs	30.04	20.47	# docs. > 3 LOCs	12,913	11,640
Maximum # of LOCs. in 1 doc.	20,594	6,472	# docs. > 30 LOCs	1,483	713

## Concluding Remarks

- First measure of geographical information (as far as named entities are concerned) in the Portuguese web
- Geographic ontology built from web texts can complement administrative sources