Sapienza
Università di Roma

Dipartimento
di Informatica

EuroSense: Automatic Harvesting of Multilingual Sense Annotations from Parallel Text

http://lcl.uniroma1.it/eurosense

Claudio Delli Bovi
dellibovi@di.uniroma1.it bn:17381128n

Jose Camacho Collados
collados@di.uniroma1.it bn:17381131n

Alessandro Raganato
raganato@di.uniroma1.it bn:17381127n

Roberto Navigli
navigli@di.uniroma1.it bn:09353187n

What is it?

- **EuroSense** is a multilingual **sense-annotated resource**, automatically built via the joint disambiguation of the **Europarl** parallel corpus **[2]** in **21** languages, with almost **123 million** sense annotations for over **155 thousand** distinct concepts and entities, drawn from the multilingual sense inventory of **BabelNet** [4].
- **EuroSense**'s disambiguation pipeline is designed to exploit at best the **cross-language complementarities** of the parallel corpus, without relying on word alignments against a pivot language.

OPUS

BabelNet

Babelfy

The tools

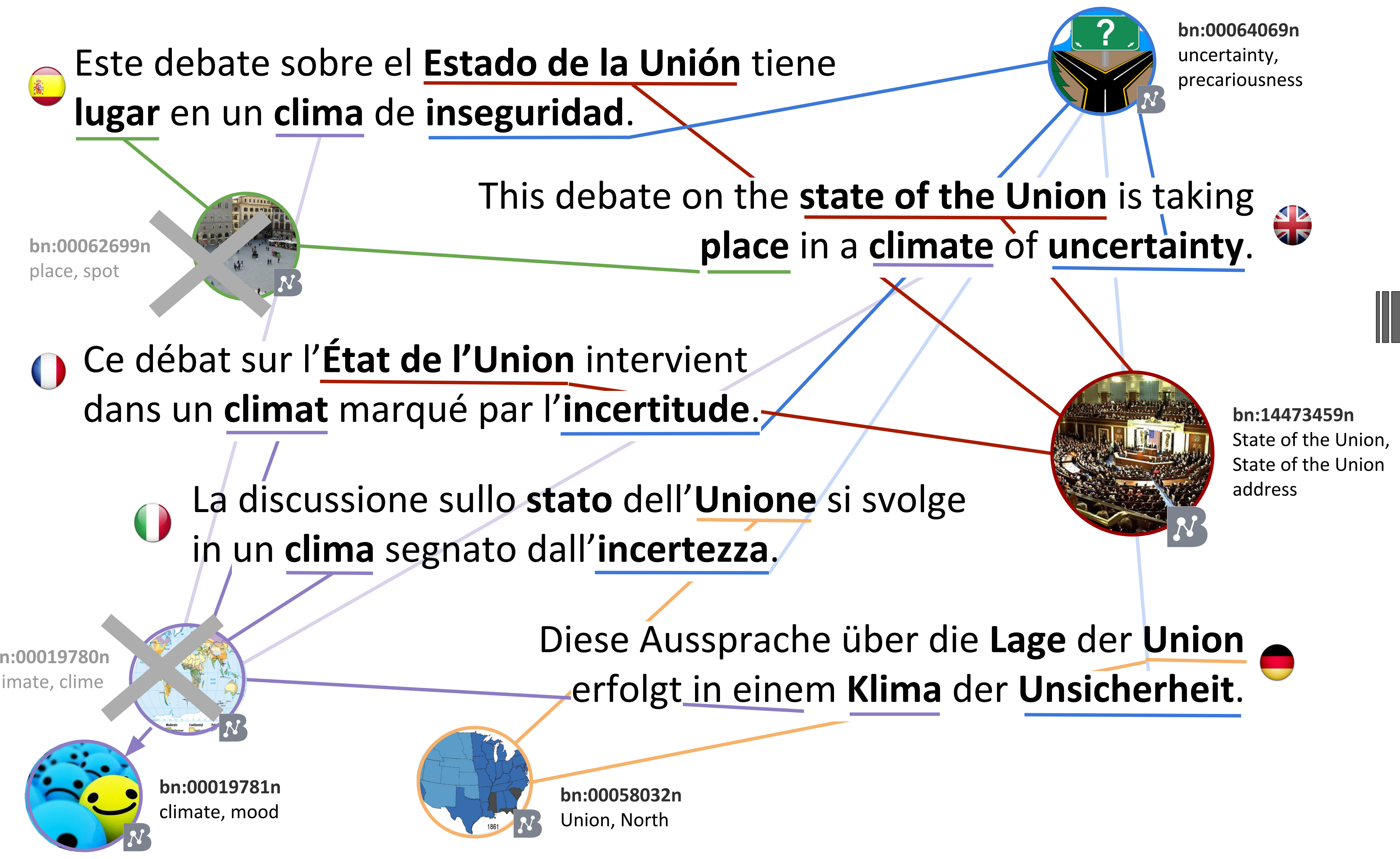
- **Babelfy [3]** is a state-of-the-art graph-based multilingual **disambiguation and entity linking system** powered by BabelNet
  http://babelfy.org
- **Nasari [1]** is a **language-independent vector representation** of concepts and entities from BabelNet and WIkipedia,
  http://lcl.uniroma1.it/nasari

Stage 1: Multilingual Disambiguation

- **Multilingual preprocessing (tokenipart-of-speech tagging, lemmatization)** with TreeTagger + Babelfy's preprocessing pipeline;
- For each sentence, gather all its available translations together in a **multilingual text;**
- **Multilingual disambiguation** using Babelfy's **densest subgraph** algorithm in such a way that it favors sense assignments that are consistent across languages.

$$\text{coherence}(d) = \frac{\#\ \text{connections}_d}{\sum_{i \in D} \#\ \text{connections}_i}$$

Coherence Score


Este debate sobre el **Estado de la Unión** tiene **lugar** en un **clima** de **inseguridad**.

This debate on the **state of the Union** is taking **place** in a **climate** of **uncertainty**.

Ce débat sur l'**État de l'Union** intervient dans un **climat** marqué par l'**incertitude**.

La discussione sullo **stato** dell'**Unione** si svolge in un **clima** segnato dall'**incertezza**.

Diese Aussprache über die **Lage** der **Union** erfolgt in einem **Klima** der **Unsicherheit**.

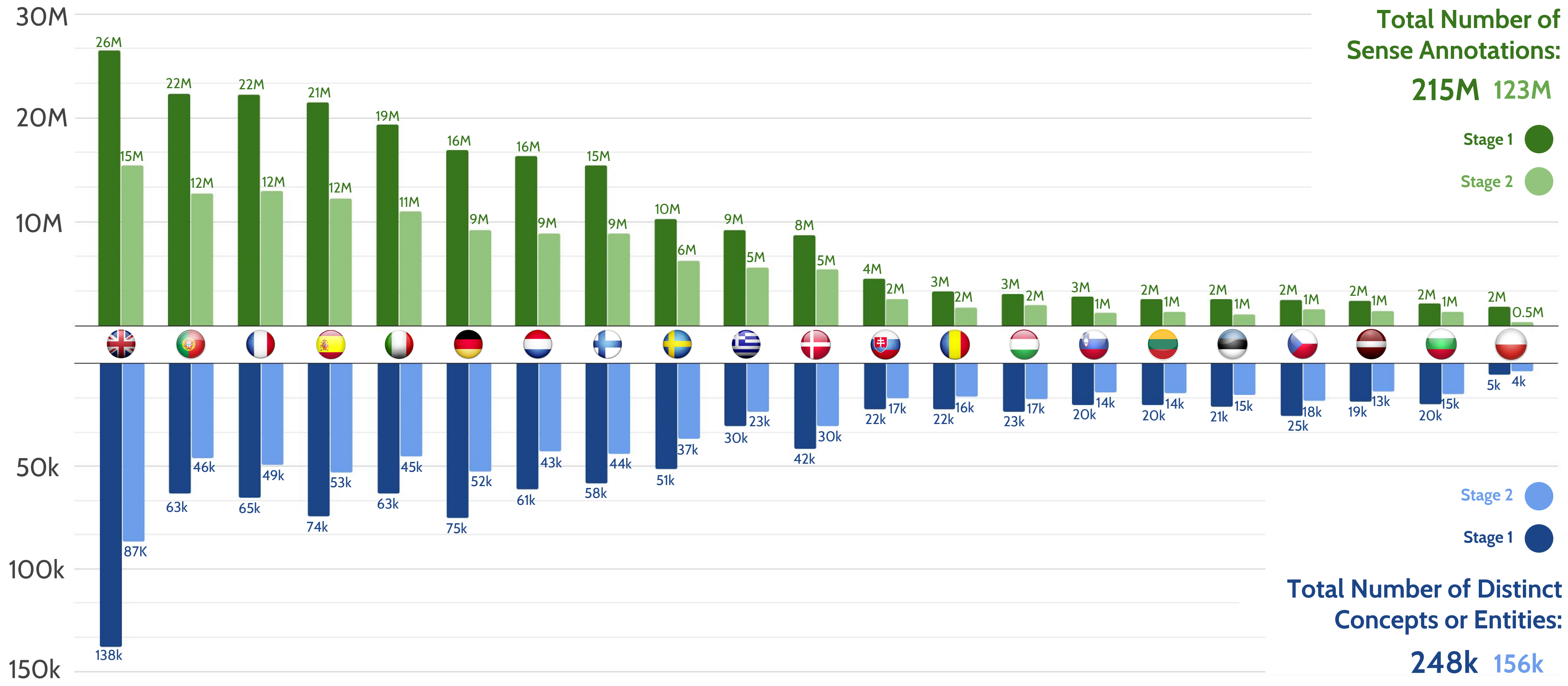
Stage 2: Similarity-based Refinement

- For each sentence, identify a subset **D** of **high-confidence disambiguations** (using the coherence score) from stage 1;
- Take the **Nasari vectors** associated with the disambiguations in **D** and compute the **centroid of D**;
- **Re-disambiguate** the mentions associated with the remaining disambiguations with the sense $\hat{s}$ having the **closest Nasari vector to the centroid**:

$$\hat{s} = \underset{s \in S_w}{\operatorname{argmax}}\ cos\left(\frac{\sum_{d \in D} \vec{d}}{|D|}, \vec{s}\right)$$

Similarity Score

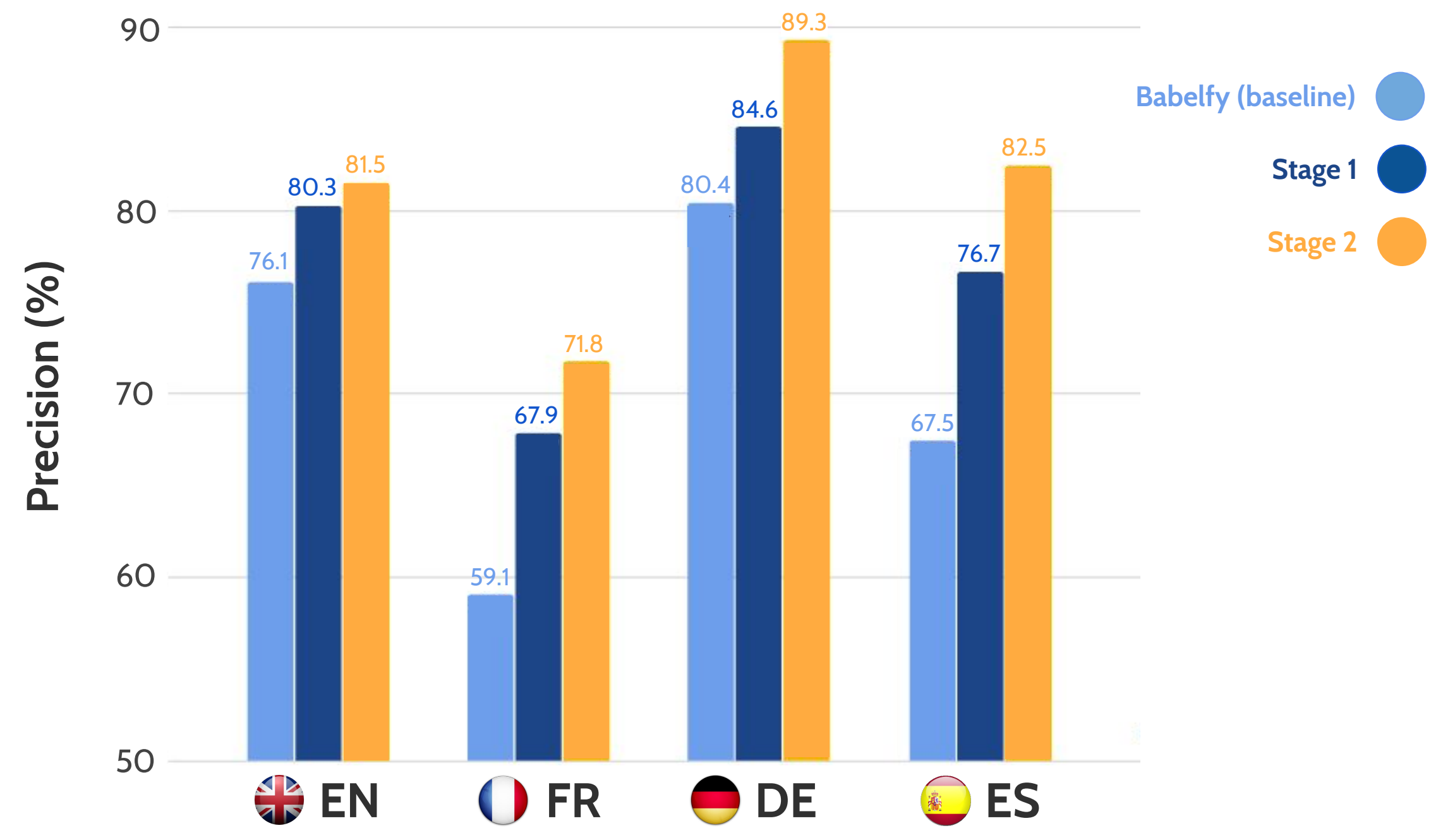
EuroSense: Statistics by Language



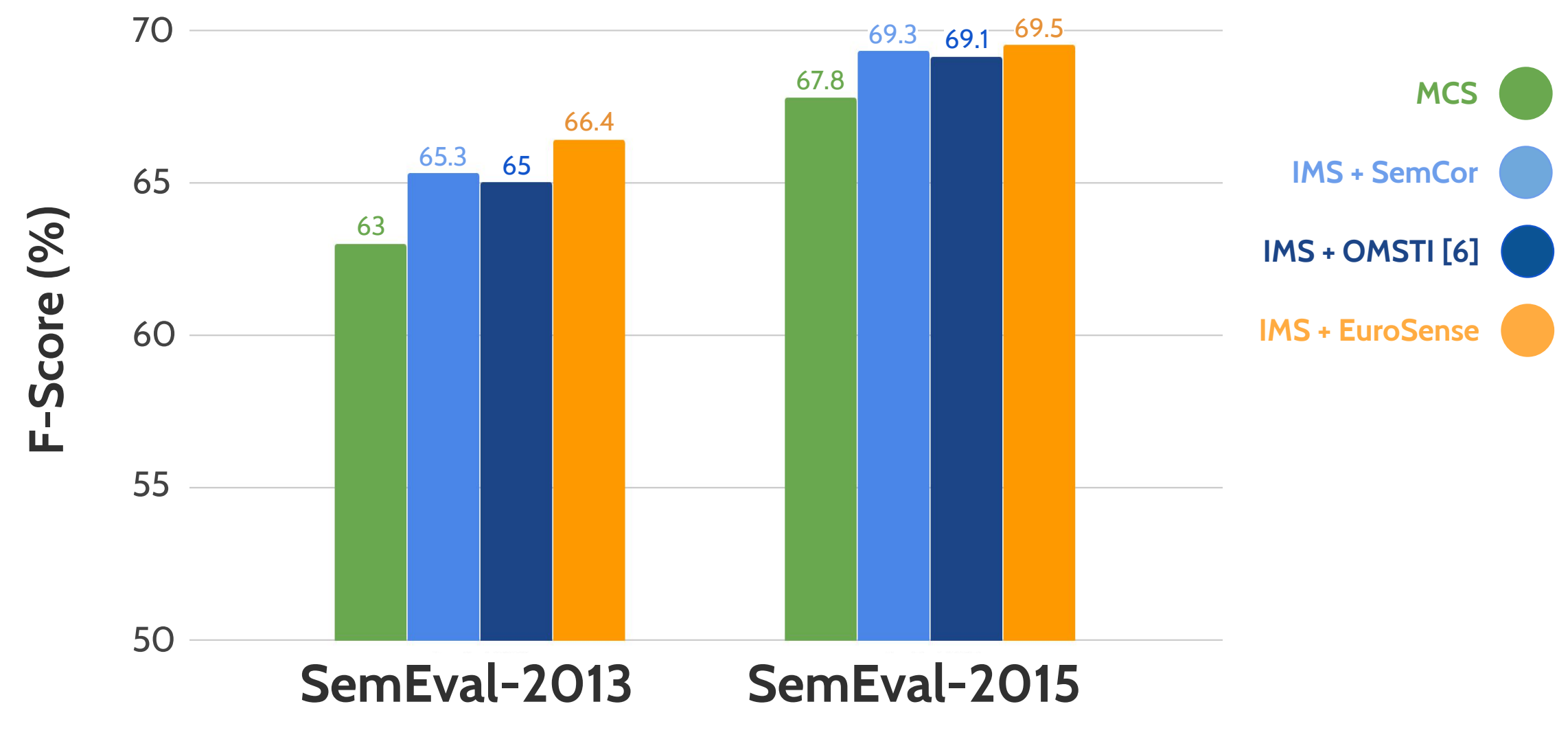
Total Number of Sense Annotations:
215M 123M

Total Number of Distinct Concepts or Entities:
248k 156k

Experimental Evaluation



- **Intrinsic Evaluation: Annotation Quality**
  4 languages, 2 human judges per language, 50 random sentences for each configuration (baseline, stage 1, stage 2)



- **Extrinsic Evaluation: Word Sense Disambiguation**
  EuroSense's English sense annotations as training set for a supervised WSD system: It Makes Sense (IMS) **[5]**

References

[1] J. Camacho-Collados, M. T. Pilehvar, and R. Navigli. 2016. *Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities.* AIJ, 240:36–64.
[2] P. Koehn. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation.* MT summit. vol. 5, pp. 79–86.
[3] A. Moro, A. Raganato, and R. Navigli. 2014. *Entity Linking meets Word Sense Disambiguation: a Unified Approach.* TACL, 2:231–244.
[4] R. Navigli and S. P. Ponzetto. 2012. *BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network.* AIJ, 193:217–250.
[5] Z. Zhong and H. T. Ng. 2010. *It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text.* ACL: System Demonstrations, pp. 78–83.
[6] K. Taghipour and H. T. Ng. 2015. *One Million Sense-tagged instances for word sense disambiguation and induction.* CoNLL, pp. 338–344.

erc The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE Contract no. 726487

Claudio Delli Bovi is supported by a Sapienza Research Grant 'Avvio alla Ricerca 2016'. Jose Camacho-Collados is supported by a Google PhD Fellowship in NLP