# CSMs Today & Tomorrow

## Workshop Summary and Final Discussion


Marco Baroni

*CIMEC, University of Trento*

Stefan Evert

*CogSci, University of Osnabrück*

Alessandro Lenci

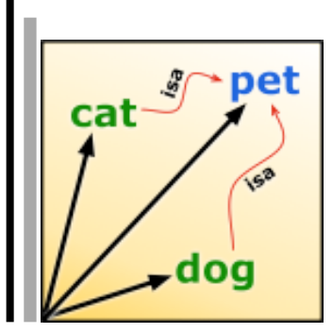
*University of Pisa*


# Workshop summary

## Distributional Lexical Semantics Workshop @ ESSLLI 2008

**We would like to thank the speakers & all participants for an exciting, fruitful and enjoyable workshop!**

# Task 1: Semantic categorization

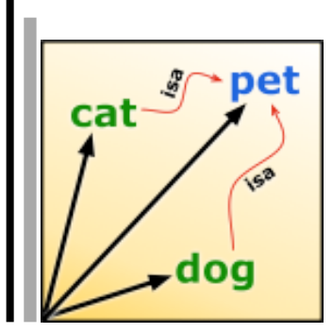



Distributional Lexical Semantics Workshop @ ESSLLI 2008

| | nouns/6 | nouns/3 | nouns/2 | abstract (hi/lo) | abstract (3-way) | verbs/9 | verbs/5 | |
|---|---|---|---|---|---|---|---|---|
| Bullinaria | .886/.120 | *worse?* | | | | .644/.527 | | |
| Shaoul (HAL) | .386/.719 | .523/.869 | .545/.981 | .725/.647 | | .333/.709 | .511/.715 | |
| Van de Cruys (bag of words) | .682/.334 | .705/.539 | .545/.983 | **1.0/0.0** | .700/.605 | .556/.442 | .600/.463 | *NL* |
| Van de Cruys (syntactic) | .841/.173 | **1.0/0.0** | **1.0/0.0** | **1.0/0.0** | **.750/.367** | .556/.408 | **.667/.464** | *NL* |
| Katrenko/Ad. (formal role) | .89x/.13x | **1.0/0.0** | .80/.59 | | | | | *Qualia* |
| Katrenko/Ad. (formal+agent) | .91x/.09x | **1.0/0.0** | .80/.59 | | | | | *Qualia* |
| Versley (decorrelation) | .795/.196 | *Web 1T5* | | | | **.711/.280** | *Web 1T5* | |
| Versley (best feature) | .841/.172 | *ukWaC* | | | | .733/.253 | *Web 1T5* | |
| Versley (comb. feat.) | **.977/.034** | | | | | **.778/.218** | | |
| Peirsman et al. (bag of words) | .82x/.23x | **.84x/.34x** | **.86x/.55x** | **1.0/0.0** | | .56x/.41x | .69x/.39x | |
| | *w = 2* | *w = 2* | *w = 2* | | | *w = 5…7* | | |

**These results probably reflect serious (implicit) overtraining!**

# Task 2: Free association norms



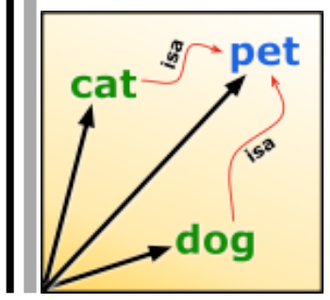

Distributional Lexical Semantics Workshop @ ESSLLI 2008

| | FIRST/other | FIRST/HAPAX/ RANDOM | correlation | prediction (mean rank) | |
|---|---|---|---|---|---|
| **Peirsman et al. (bag of words)** | | | | 47.0 | *w = 5* |
| **Wandmacher et al. (LSA on term/term)** | 79.7% | 60.3% | **.353/.263** | 51.9 | *w = 75* |
| **FOO (first-order assoc.)** | **86.3%** | | .209/.170 | **30.0** | |
| | *t-score* | | *(MI)* | *t-score* | |
| **baseline** | 66.6% | 33.3% | | | |

*FOO model using Dice measure achieves mean rank 28.0 in prediction task; for 49% of cues, the correct target is among the first 5 suggestions.*

# Task 3: Property generation



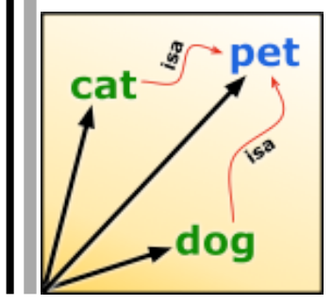



- Seems to be a difficult task for CSMs — very few results
- Shaoul (HAL): precision < 2%
- Barbu: precision 50%–80%, but not a proper CSM
  - direct property extraction with manually selected patterns
  - first-order associations work well for adjectives and verbs
  - but not evaluated against shared task gold standard!
- Marco's results on shared task data:

| | |
|---|---|
| 4.1% | SVD on term-term matrix (Rapp 2003, 2004) |
| 8.8% | Attribute-Value model (Almuhareb & Poesio 2004) |
| 14.1% | Dependency Vectors (Padó & Lapata 2007) |
| 23.9% | StruDEL (Baroni et al., to appear) |

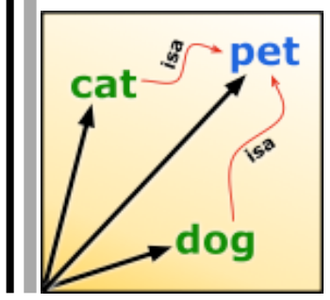
# Some important distinctions






- words vs. concepts
- distributed representation vs. distributional modelling
- theoretical discussion vs. experimental results
- key questions for distributional semantics

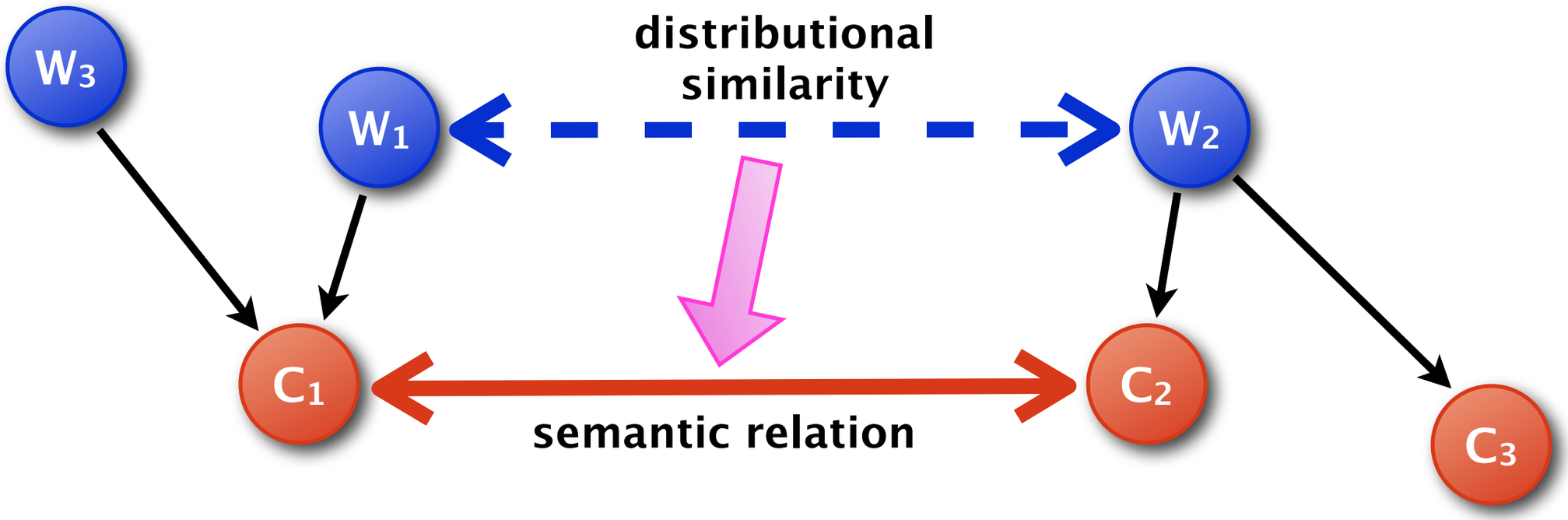
# Lexical semantics or conceptual meaning?



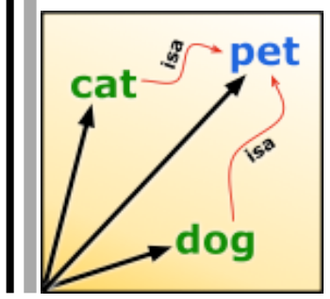



- Are we interested in the **meaning of words** or in **concepts**?
  - minimalist lexical semantics: **word = pointer to concept**
  - plus some genuinely linguistic meaning aspects (e.g. different usage and connotations of near-synonyms)
  - no function words (→ formal semantics)

- Word space hypothesis: distributional similarity between words reflects semantic relations between concepts



# Word Space: Holographic memory vs. CSM

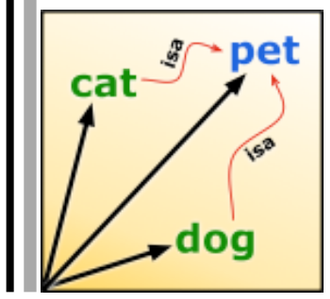





- Distributed, non-symbolic representation of meaning
  → **holographic memory**
  - Which facets of "meaning" (wrt. concepts, words, utterances, …) can be expressed in a distributed, non-symbolic representation?
  - Primarily addressed by **theoretical discussions**

- Infer meaning of word/concept from its distribution in text
  → **context-based semantic models (CSM)**
  - To what extent can the meaning of a word/concept be determined from its distribution in text?
  - Which models and parameters are best suited for this purpose?
  - Primarily addressed with **experimental methods** (→ shared task)

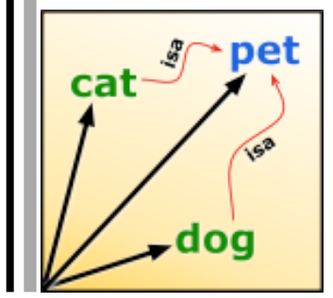
# Key questions for distributional semantics






1. What kind of information is encoded by a CSM?
   - problem: it is not clear what exactly we are looking for
2. Which aspects of lexical/conceptual meaning can be captured by holographic memory and CSMs?
   - problem: no good theory of concepts and lexical semantics
   - theoretical discussions needed to guide empirical research
3. What is the best CSM for a particular semantic task?
   - choice of model, base corpus, context definition, parameters, …
   - immediate result of shared task & workshop papers
4. Are linguistic data sufficient to build CSM representations?
   - which aspects of meaning can be learned from purely linguistic input, and which aspects require an embodied CSM?

# Key questions for distributional semantics






1. What kind of information is encoded by a CSM?
   - problem: it is not clear what exactly we

**empirical**

2. Which aspects of lexical/conceptual meaning can be captured by holographic memory and CSMs?
   - problem: no good theory of concepts and lex
   - theoretical discussions needed to guide empirical research

**theoretical**

3. What is the best CSM for a particular semantic task?
   - choice of model, base corpus, contex arameters, …
   - immediate result of shared task & worksh papers

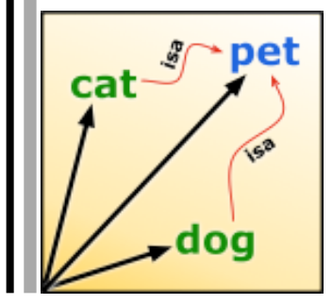
**technical**

4. Are linguistic data sufficient to build CSM representations?
   - which aspects of meaning ca ed from purely linguistic input, and which aspects req nbodied CSM?

**???**

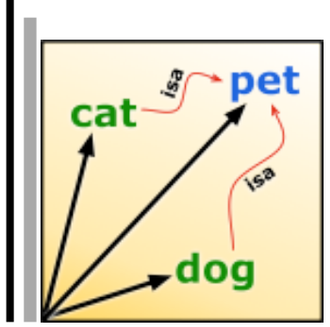
# Where to go from here






- Topic for the following discussion: the next steps
- Continue series of workshops on distributional semantics?
  - volunteers needed!
- Competitive (or friendly) evaluation campaign?
  - e.g. at SemEval 2010 (deadline for EoI: 21 Sep 2008)

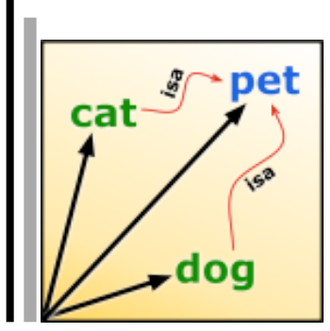
# Final discussion




Distributional Lexical Semantics Workshop @ ESSLLI 2008

Discuss!

# Stefan's position statement




Distributional Lexical Semantics Workshop @ ESSLLI 2008

- We need a battery of standardised tests for CSMs
  - cognitively plausible representation must work for all tasks!
  - shared tasks from this workshop could be part of this battery
  - add other types of tasks, different languages, etc.
  - large-scale evaluation campaign would make data sets available
- Develop common software platform to facilitate research
  - allows easy experiments with different CSMs and parameter settings, automatically running entire battery of tests
  - platform implements different uses of underlying representation for different types of tests (with automatic tuning)
  - SemanticVectors (Widdows), HIDeX (Shaoul), DependencyVectors (Padó), …
  - Python-based system for easy experimentation?