Distributional semantics and meaning

Aurelie Herbelot

Universität Potsdam
EB Kognitionswissenschaft

Freie Universität Berlin, 2013

Outline

1. Distributional semantics
   - Some historical pointers
   - Building distributions
   - What are distributions good for?
   - Some notes on the representation
2. Which semantics in distributional semantics?
3. Lexicalised Compositionality (LC)
   - Ideal distributions
   - From actual to ideal distributions
4. Conclusion

Outline

1. Distributional semantics
   - Some historical pointers
   - Building distributions
   - What are distributions good for?
   - Some notes on the representation
2. Which semantics in distributional semantics?
3. Lexicalised Compositionality (LC)
   - Ideal distributions
   - From actual to ideal distributions
4. Conclusion

The citations

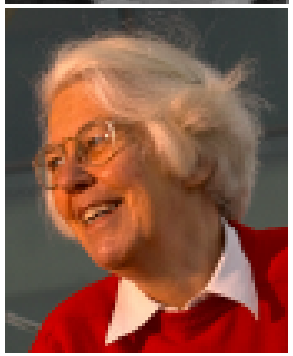
Harris (1954)

Words that appear in the same context are semantically similar.

Firth (1957)

'You shall know a word by the company it keeps.'

A continuous story



Ludwig Wittgenstein (1953): Words are defined by their usage.

Margaret Masterman (1955): Cambridge Language Research Unit (CLRU).

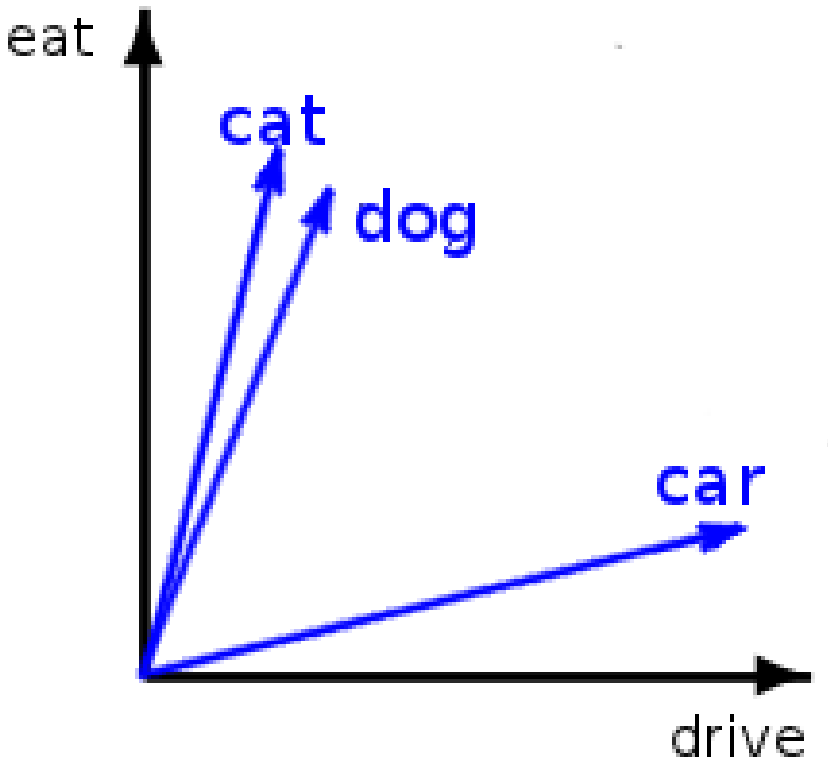
Karen Spärck-Jones (Late 1950s): With Harper, first experiments on distributional semantics.

The general intuition

- **Distributions** are vectors in a multidimensional semantic space, that is, objects with a magnitude (length) and a direction.
- The **semantic space** has dimensions which correspond to possible contexts, as taken from a given corpus.

A distributional space

- A mini-distributional space, with two possible contexts, *eat* and *drive*.



- In practice, many more dimensions are used:
  cat [...dog 0.8, eat 0.7, joke 0.01, mansion 0.2, zebra 0.1...]

Outline

1. Distributional semantics
   - Some historical pointers
   - Building distributions
   - What are distributions good for?
   - Some notes on the representation
2. Which semantics in distributional semantics?
3. Lexicalised Compositionality (LC)
   - Ideal distributions
   - From actual to ideal distributions
4. Conclusion

The notion of context

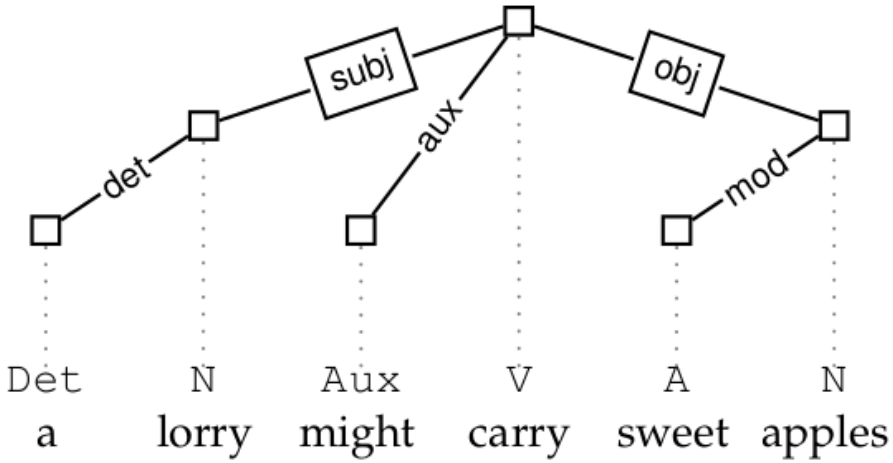
- **Context:** if the meaning of a word is given by its context, what does 'context' mean?
  - Word windows (unfiltered): *n* words on either side of the lexical item under consideration (unparsed text).
    Example: n=2 (5 words window):

    *... the prime **minister** acknowledged that ...*

  - Word windows (filtered): *n* words on either side of the lexical item under consideration (unparsed text). Some words are not considered part of the context (e.g. function words, some very frequent content words). The stop list for function words is either constructed manually, or the corpus is POS-tagged.
    Example: n=2 (5 words window):
    *... yesterday the prime **minister** acknowledged that he had ...*

The notion of context

- Dependencies: syntactic or semantic. The corpus is converted into a list of directed links between heads and dependents. Context for a lexical item is the dependency structure it belongs to. The length of the dependency path can vary according to the implementation (Padó and Lapata, 2007).




```
a        [Det,det,N]   lorry
lorry    [N,subj,V]    carry
might    [Aux,aux,V]   carry
apples   [N,obj,V]     carry
sweet    [A,mod,N]     apples
```

Parsed vs unparsed data: examples

word (unparsed)	**word (parsed)**
meaning_n	or_c+phrase_n
derive_v	and_c+phrase_n
dictionary_n	syllable_n+of_p
pronounce_v	play_n+on_p
phrase_n	etymology_n+of_p
latin_j	portmanteau_n+of_p
ipa_n	and_c+deed_n
verb_n	meaning_n+of_p
mean_v	from_p+language_n
hebrew_n	pron_rel_+utter_v
usage_n	for_p+word_n
literally_r	in_p+sentence_n

Context weighting

- Variations on the characteristic model: the weights given to the vector components express how *characteristic* a given context is for w. Functions used include:
  - Pointwise Mutual Information (PMI), with or without discounting factor.

$$pmi_{wc} = log(\frac{f_{wc} * f_{total}}{f_w * f_c}) \quad (1)$$

  - See Evert, 2004, for a summary.

What semantic space?

- Entire vocabulary.
  - + All information included – even rare, but important contexts
  - - Inefficient (100,000s dimensions). Noisy (e.g. *002.png|thumb|right|200px|graph_n*)
- Top *n* words with highest frequencies.
  - + More efficient (2000-10000 dimensions). Only 'real' words included.
  - - May miss out on infrequent but relevant contexts.

What semantic space?

- Singular Value Decomposition (LSA – Landauer and Dumais, 1997): the number of dimensions is reduced by exploiting redundancies in the data. A new dimension might correspond to a generalisation over several of the original dimensions (e.g. the dimensions for *car* and *vehicle* are collapsed into one).
  - + Very efficient (200-500 dimensions). Captures generalisations in the data.
  - - SVD matrices are not interpretable.

An example noun

- *language*:

0.541816::other+than_p()+English_n
0.525895::English_n+as_p()
0.523398::English_n+be_v
0.48977::english_a
0.481964::and_c+literature_n
0.476664::people_n+speak_v
0.468399::French_n+be_v
0.463604::Spanish_n+be_v
0.463591::and_c+dialects_n
0.452107::grammar_n+of_p()
0.445994::foreign_a
0.445071::germanic_a
0.439558::German_n+be_v
0.436135::of_p()+instruction_n
0.435633::speaker_n+of_p()
0.423595::generic_entity_rel_+speak_v
0.42313::pron_rel_+speak_v
0.42294::colon_v+English_n
0.419646::be_v+English_n
0.418535::language_n+be_v
0.4159::and_c+culture_n
0.410987::arabic_a
0.408387::dialects_n+of_p()
0.399266::part_of_rel_+speak_v
0.397::percent_n+speak_v
0.39328::spanish_a
0.39273::welsh_a
0.391575::tonal_a

An example adjective

- *academic*:

0.517031::Decathlon_n
0.512661::excellence_n
0.449711::dishonesty_n
0.445393::rigor_n
0.426142::achievement_n
0.421246::discipline_n
0.397311::vice_president_n+for_p()
0.391978::institution_n
0.38937::credentials_n
0.378062::journal_n
0.373727::journal_n+be_v
0.372052::vocational_a
0.371873::student_n+achieve_v
0.361359::athletic_a
0.356562::reputation_n+for_p()
0.354674::regalia_n
0.353712::program_n
0.351601::freedom_n
0.347751::student_n+with_p()
0.34621::curriculum_n
0.342008::standard_n
0.34151::at_p()+institution_n
0.340271::career_n
0.337857::Career_n
0.329923::dress_n
0.329358::scholarship_n
0.329281::prepare_v+student_n
0.328009::qualification_n

Outline

1. Distributional semantics
   - Some historical pointers
   - Building distributions
   - What are distributions good for?
   - Some notes on the representation
2. Which semantics in distributional semantics?
3. Lexicalised Compositionality (LC)
   - Ideal distributions
   - From actual to ideal distributions
4. Conclusion

Why use distributions?

- Modelling similarity:
  - Applications: document retrieval and classification, question answering, machine translation, etc.
  - Psychological phenomena: semantic priming, generating feature norms, etc.
- Semantic representation in tasks that require lexical information: compound noun classification, parsing, etc.
- Modelling composition at the lexical level (?)

Outline

1. Distributional semantics
   - Some historical pointers
   - Building distributions
   - What are distributions good for?
   - Some notes on the representation
2. Which semantics in distributional semantics?
3. Lexicalised Compositionality (LC)
   - Ideal distributions
   - From actual to ideal distributions
4. Conclusion

Corpus choice

- As much data as possible?
  - British National Corpus (BNC): 100 m words
  - Wikipedia: 897 m words
  - UKWac: 2 bn words
  - ...
- In general preferable, *but*:
  - More data is not necessarily the data you want.
  - More data is not necessarily realistic from a psycholinguistic point of view. We perhaps encounter 50,000 words a day. BNC = 5 years' text exposure.

Corpus choice

- Distribution for *unicycle*, as obtained from Wikipedia.

0.448051::motorized_a
0.404372::pron_rel_+ride_v
0.238612::for_p()+entertainment_n
0.235763::half_n+be_v
0.235407::unwieldy_a
0.230275::earn_v+point_n
0.216627::pron_rel_+crash_v
0.190785::man_n+on_p()
0.186325::on_p()+stage_n
0.185063::position_n+on_p()
0.168102::slip_v
0.162611::and_c+1_n
0.159627::autonomous_a
0.155822::balance_v
0.133084::tall_a
0.124242::fast_a
0.106976::red_a
0.0714643::come_v
0.0601987::high_a

Polysemy

- Distribution for *pot*, as obtained from Wikipedia.

0.566454::melt_v
0.442374::pron_rel_+smoke_v
0.434682::of_p()+gold_n
0.40773::porous_a
0.401654::of_p()+tea_n
0.39444::player_n+win_v
0.393812::money_n+in_p()
0.376198::of_p()+coffee_n
0.33117::amount_n+in_p()
0.329211::ceramic_a
0.326387::hot_a
0.323321::boil_v
0.313404::bowl_n+and_c
0.306324::ingredient_n+in_p()
0.301916::plant_n+in_p()
0.298764::simmer_v
0.292397::pot_n+and_c
0.284539::bottom_n+of_p()
0.28338::of_p()+flower_n
0.279412::of_p()+water_n
0.278914::food_n+in_p()
0.262501::pron_rel_+heat_v
0.260375::size_n+of_p()
0.25511::pron_rel_+split_v
0.254363::of_p()+money_n
0.2535::of_p()+culture_n
0.249626::player_n+take_v
0.246479::in_p()+hole_n
0.244051::of_p()+soil_n
0.243797::city_n+become_v

Fixed expressions

- Distribution for *time*, as obtained from Wikipedia.

0.462949::of_p()+death_n
0.448965::same_a
0.446277::1_n+at_p(temp)
0.445338::Nick_n+of_p()
0.423542::spare_a
0.418568::playoffs_n+for_p()
0.416471::of_p()+retirement_n
0.405288::of_p()+release_n
0.397135::pron_rel_+spend_v
0.389886::sand_n+of_p()
0.385954::pron_rel_+waste_v
0.382816::place_n+around_p()
0.37777::of_p()+arrival_n
0.376466::of_p()+completion_n
0.374797::after_p()+time_n
0.374682::of_p()+arrest_n
0.371589::country_n+at_p()
0.370736::age_n+at_p()
0.370626::space_n+and_c
0.370555::in_p()+career_n
0.370464::world_n+at_p()
0.363982::and_c+space_n
0.363241::generic_entity_rel_+mark_v
0.361872::of_p()+introduction_n
0.357929::in_p()+year_n
0.357565::of_p()+appointment_n
0.356229::of_p()+trouble_n
0.355658::of_p()+merger_n
0.354794::on_p()+ice_n
0.353891::practice_n+at_p()
0.351994::of_p()+birth_n
0.351556::full_a
0.348029::of_p()+accident_n
0.34785::state_n+at_p()
0.347753::to_p()+time_n
0.345147::of_p()+election_n
0.345088::area_n+at_p()
0.342571::and_c+money_n
0.342113::time_n+after_p()
0.341877::allotted_a

Outline

1. Distributional semantics
   - Some historical pointers
   - Building distributions
   - What are distributions good for?
   - Some notes on the representation
2. Which semantics in distributional semantics?
3. Lexicalised Compositionality (LC)
   - Ideal distributions
   - From actual to ideal distributions
4. Conclusion

What is a Wittgenstinian semantics?

- (Late) Wittgenstein: it makes no sense to ask what things are in the world (the preoccupation of metaphysics). Meaning only results from language games, not from the world.
- Is this compatible with doing semantics?
- Or phrased otherwise: as (traditional) semanticists, do we care about what *X* is? When *X* is...
  - life
  - red
  - very
  - not
  - can
  - some
- So are distributional semanticists not semanticists after all?

What is a Wittgenstinian semantics?

- (Late) Wittgenstein: it makes no sense to ask what things are in the world (the preoccupation of metaphysics). Meaning only results from language games, not from the world.
- Is this compatible with doing semantics?
- Or phrased otherwise: as (traditional) semanticists, do we care about what *X* is? When *X* is...
  - life
  - red
  - very
  - not
  - can
  - some
- So are distributional semanticists not semanticists after all?

What is a Wittgenstinian semantics?

- (Late) Wittgenstein: it makes no sense to ask what things are in the world (the preoccupation of metaphysics). Meaning only results from language games, not from the world.
- Is this compatible with doing semantics?
- Or phrased otherwise: as (traditional) semanticists, do we care about what *X* is? When *X* is...
  - life
  - red
  - very
  - not
  - can
  - some
- So are distributional semanticists not semanticists after all?

What is a Wittgenstinian semantics?

- (Late) Wittgenstein: it makes no sense to ask what things are in the world (the preoccupation of metaphysics). Meaning only results from language games, not from the world.
- Is this compatible with doing semantics?
- Or phrased otherwise: as (traditional) semanticists, do we care about what *X* is? When *X* is...
  - life
  - red
  - very
  - not
  - can
  - some
- So are distributional semanticists not semanticists after all?

What is a Wittgenstinian semantics?

- (Late) Wittgenstein: it makes no sense to ask what things are in the world (the preoccupation of metaphysics). Meaning only results from language games, not from the world.
- Is this compatible with doing semantics?
- Or phrased otherwise: as (traditional) semanticists, do we care about what *X* is? When *X* is...
  - life
  - red
  - very
  - not
  - can
  - some
- So are distributional semanticists not semanticists after all?

What is a Wittgenstinian semantics?

- (Late) Wittgenstein: it makes no sense to ask what things are in the world (the preoccupation of metaphysics). Meaning only results from language games, not from the world.
- Is this compatible with doing semantics?
- Or phrased otherwise: as (traditional) semanticists, do we care about what *X* is? When *X* is...
  - life
  - red
  - very
  - not
  - can
  - some
- So are distributional semanticists not semanticists after all?

What is a Wittgenstinian semantics?

- (Late) Wittgenstein: it makes no sense to ask what things are in the world (the preoccupation of metaphysics). Meaning only results from language games, not from the world.
- Is this compatible with doing semantics?
- Or phrased otherwise: as (traditional) semanticists, do we care about what *X* is? When *X* is...
  - life
  - red
  - very
  - not
  - can
  - some
- So are distributional semanticists not semanticists after all?

What is a Wittgenstinian semantics?

- (Late) Wittgenstein: it makes no sense to ask what things are in the world (the preoccupation of metaphysics). Meaning only results from language games, not from the world.
- Is this compatible with doing semantics?
- Or phrased otherwise: as (traditional) semanticists, do we care about what *X* is? When *X* is...
  - life
  - red
  - very
  - not
  - can
  - some
- So are distributional semanticists not semanticists after all?

Change meaning!

- If meaning is usage, the semanticist can study usage and still be a semanticist.
- But: can usage account for all observable phenomena in language? In particular, those phenomena which are never explicitely uttered in language but intuitively felt by speakers of a language.

Semantic content in distributions

- Joint work with Mohan Ganesalingam.
- Calculate semantic content: can we model *make* < *produce* < *weave* or *group* < *14* using distributions?
- Use Kullback-Leibler divergence to compare the (statistical) distribution of context words on their own and their distribution next to the target word.

$$D_{\mathrm{KL}}(P\|Q) = \sum_i \ln(\frac{P(i)}{Q(i)})P(i) \tag{2}$$

Semantic content in distributions

- Semantic content should model hyponymy: *chair* has more semantic content than *thing*, *produce* less than *weave*.
- Disappointing result: 80% precision on the task of correctly ordering 1279 hyponym-hypernym pairs.
- Main problem: Extension is not modelled properly. **Example:** *beverage* is strongly linked to *food, wine, tea, coffee, etc*.

Distributions are intensions...?

- The assumption amongst distributional semanticists is that distributions do not tell us much about extension. They are more intensional in nature. (Baroni et al, 2012; Erk, 2013)
- Intensions are usually defined as mapping from possible worlds to extensions. (We are drifting away from Wittgenstein...)
- It is debatable whether they are even intensional: some essential aspects of concepts do not appear in distributions.

The intension game

- Distribution for *concrete* (noun), as obtained from Wikipedia.

0.542296::and_c+steel_n
0.540451::steel_n+and_c
0.512329::slab_n+of_p()
0.466818::brick_n+and_c
0.463849::steel_n+or_c
0.453806::meter_n+of_p()
0.442502::and_c+glass_n
0.436364::stone_n+and_c
0.428527::and_c+brick_n
0.380303::be_v+material_n
0.374869::glass_n+and_c
0.374346::material_n+such+as_p()
0.374041::and_c+granite_n
0.367402::ton_n+of_p()
0.353181::or_c+stone_n
0.351596::yard_n+of_p()
0.342199::consistency_n+of_p()
0.340048::and_c+concrete_n
0.338328::or_c+metal_n
0.333411::centimeter_n+of_p()
0.331533::concrete_n+and_c
0.323514::exposed_a
0.317804::and_c+clay_n
0.31632::wood_n+and_c
0.31594::strength_n+of_p()
0.314691::foot_n+of_p()
0.312795::inch_n+of_p()
0.306334::Stone_n+and_c
0.304715::material_n+be_v

Are distributions doomed?

- Distributions do not model intension if we define intension as ‘a mapping from possible worlds to extension’.
- Distributions highlight a certain aspect of ‘discourse’ (we are vearing from philosophy of language to critical theory...)
- Discourse analysis is nice, but it is not semantics.
- Can we produce distributions which we can relate to both intension and extension in a traditional fashion?

Outline

1. Distributional semantics
   - Some historical pointers
   - Building distributions
   - What are distributions good for?
   - Some notes on the representation
2. Which semantics in distributional semantics?
3. Lexicalised Compositionality (LC)
   - Ideal distributions
   - From actual to ideal distributions
4. Conclusion

Outline

1. Distributional semantics
   - Some historical pointers
   - Building distributions
   - What are distributions good for?
   - Some notes on the representation
2. Which semantics in distributional semantics?
3. Lexicalised Compositionality (LC)
   - Ideal distributions
   - From actual to ideal distributions
4. Conclusion

Distributions are not extension

- Distributions do not encapsulate extension:
  - Because of pragmatic matters.

    My cat Kitty, who is a mammal, is 2 years old.
    My cat Kitty (a mammal) likes playing in the garden.
    Kitty, my cat – and a mammal –, is hungry.

  - Because sentences are uttered by real people (and therefore lead to inconsistencies, etc).
  - Because we don't know everything about the world (and therefore cannot say it).
- What if we consider a unique non-Gricean omniscient being?

LC: An idealised representation

- Joint work with Ann Copestake.
- **Ideal distributions** correspond to *complete distributional information* for a world *w*.
- They encapsulate information about *individual entities* and the *situations* in which those entities are found.
- They are hypothetical in the sense that they cannot be straightforwardly extracted from text.

Example: a microworld w_0

- Two small elephants playing and, in another place and at another time, a zebra eating.
- Let's assume a speaker whose vocabulary consists of the terms *small, elephant, zebra, play, eat* and the quantifiers *a/an* and *two*.

A small elephant plays. (x2)
Two small elephants play.
An elephant plays. (x2)
Two elephants play.
A zebra eats.

Logical forms for w_0

- Logical forms in predicate logic (implicit conjunctions):

$\text{elephant}'(x_1), \text{small}'(x_1), \text{play}'(e_1, x_1)$
$\text{elephant}'(x_2), \text{small}'(x_2), \text{play}'(e_2, x_2)$
$\text{elephant}'(x_1), \text{small}'(x_1), \text{play}'(e_1, x_1), \text{elephant}'(x_2), \text{small}'(x_2), \text{play}'(e_2, x_2)$
$\text{elephant}'(x_1), \text{play}'(e_1, x_1)$
$\text{elephant}'(x_2), \text{play}'(e_2, x_2)$
$\text{elephant}'(x_1), \text{play}'(e_1, x_1), \text{elephant}'(x_2), \text{play}'(e_2, x_2)$
$\text{zebra}'(x_3), \text{eat}'(e_3, x_3)$

- **Note:** plural quantifiers are expressed by repeating the appropriate logical form for each entity in the plural set.

Ideal context sets for w_0

$$
\begin{array}{lll}
\text{elephant}^\circ \equiv & \{ & < [x1][\text{small}^\circ(x1), \text{play}^\circ(e1, x1)], S_1 >, \\
& & < [x1][\text{play}^\circ(e1, x1)], S_1 >, \\
& & < [x2][\text{small}^\circ(x2), \text{play}^\circ(e2, x2)], S_1 >, \\
& & < [x2][\text{play}^\circ(e2, x2)], S_1 > \} \\
\text{zebra}^\circ \equiv & \{ & < [x3][\text{eat}^\circ(e3, x3)], S_2 > \} \\
\text{small}^\circ \equiv & \{ & < [x1][\text{elephant}^\circ(x1), \text{play}^\circ(e1, x1)], S_1 >, \\
& & < [x2][\text{elephant}^\circ(x2), \text{play}^\circ(e2, x2)], S_1 > \} \\
\text{play}^\circ \equiv & \{ & < [e1, x1][\text{elephant}^\circ(x1), \text{small}^\circ(x1)], S_1 >, \\
& & < [e1, x1][\text{elephant}^\circ(x1)], S_1 >, \\
& & < [e2, x2][\text{elephant}^\circ(x2), \text{small}^\circ(x2)], S_1 >, \\
& & < [e2, x2][\text{elephant}^\circ(x2)], S_1 > \} \\
\text{eat}^\circ \equiv & \{ & < [e3, x3][\text{zebra}^\circ(x3)], S_2 > \}
\end{array}
$$

Figure: Full context sets for w_0

Correspondence between LC and models

- There is a very straightforward correspondence between LC and the standard notion of extension (and of intension?)
- We only need to know the real world equalities between the constants corresponding to distributional arguments.

World w_1

- w_1 comprises one situation with two playing elephants, one eating elephant and one elephant that eats and plays.
  We omit the situation variable in what follows.

$$
\begin{aligned}
\text{elephant}^\circ = \quad \{ \quad & < [x1][\text{play}^\circ(e1, x1)] >, \\
& < [x2][\text{play}^\circ(e2, x2)] >, \\
& < [x3][\text{eat}^\circ(e3, x3)] >, \\
& < [x4][\text{play}^\circ(e4, x4)] >, \\
& < [x4][\text{eat}^\circ(e4, x4)] > \} \\
\text{play}^\circ = \quad \{ \quad & < [e1, x1][\text{elephant}^\circ(x1)] >, \\
& < [e2, x2][\text{elephant}^\circ(x2)] >, \\
& < [e5, x4][\text{elephant}^\circ(x4)] >, \\
& < [e5, x4][\text{eat}^\circ e4, x4] > \} \\
\text{eat}^\circ = \quad \{ \quad & < [e3, x3][\text{elephant}^\circ(x3)] >, \\
& < [e4, x4][\text{elephant}^\circ(x4)] >, \\
& < [e4, x4][\text{play}^\circ e5, x4] > \}
\end{aligned}
$$

Assume each lexeme co-occurs with itself

$$
\begin{aligned}
\text{elephant}^\circ(x) \equiv \quad \{ \quad & < [x1][\text{elephant}^\circ(x1)] >, \\
& < [x2][\text{elephant}^\circ(x2)] >, \\
& < [x3][\text{elephant}^\circ(x3)] >, \\
& < [x4][\text{elephant}^\circ(x4)] >, \\
& < [x1][\text{play}^\circ(e1, x1)] >, \\
& < [x2][\text{play}^\circ(e2, x2)] >, \\
& < [x3][\text{eat}^\circ(e3, x3)] >, \\
& < [x4][\text{play}^\circ(e4, x4)] >, \\
& < [x4][\text{eat}^\circ(e4, x4)] > \} \\
\text{play}^\circ(e, x) \equiv \quad \{ \quad & < [e1, x1][\text{play}^\circ(e1)] >, \\
& < [e2, x2][\text{play}^\circ(e2)] >, \\
& < [e5, x4][\text{play}^\circ(e5)] >, \\
& < [e1, x1][\text{elephant}^\circ(x1)] >, \\
& < [e2, x2][\text{elephant}^\circ(x2)] >, \\
& < [e5, x4][\text{elephant}^\circ(x4)] >, \\
& < [e5, x4][\text{eat}^\circ e4, x4] > \} \\
\text{eat}^\circ(e, x) \equiv \quad \{ \quad & < [e3, x3][\text{eat}^\circ(e3)] >, \\
& < [e4, x4][\text{eat}^\circ(e4)] >, \\
& < [e3, x3][\text{elephant}^\circ(x3)] >, \\
& < [e4, x4][\text{elephant}^\circ(x4)] >, \\
& < [e4, x4][\text{play}^\circ e5, x4] > \}
\end{aligned}
$$

Underspecify entities

$$
\begin{aligned}
\text{elephant}^{\circ}(x) \equiv \quad \{ \quad & < [x][\text{elephant}^{\circ}(x)] >, \\
& < [x][\text{elephant}^{\circ}(x)] >, \\
& < [x][\text{elephant}^{\circ}(x)] >, \\
& < [x][\text{elephant}^{\circ}(x)] >, \\
& < [x][\text{play}^{\circ}(e, x)] >, \\
& < [x][\text{play}^{\circ}(e, x)] >, \\
& < [x][\text{eat}^{\circ}(e, x)] >, \\
& < [x][\text{play}^{\circ}(e, x)] >, \\
& < [x][\text{eat}^{\circ}(e, x)] > \} \\
\text{play}^{\circ}(e, x) \equiv \quad \{ \quad & < [e, x][\text{play}^{\circ}(e)] >, \\
& < [e, x][\text{play}^{\circ}(e)] >, \\
& < [e, x][\text{play}^{\circ}(e)] >, \\
& < [e, x][\text{elephant}^{\circ}(x)] >, \\
& < [e, x][\text{elephant}^{\circ}(x)] >, \\
& < [e, x][\text{elephant}^{\circ}(x)] >, \\
& < [e, x][\text{eat}^{\circ} e, x] > \} \\
\text{eat}^{\circ}(e, x) \equiv \quad \{ \quad & < [e, x][\text{eat}^{\circ}(e)] >, \\
& < [e, x][\text{eat}^{\circ}(e)] >, \\
& < [e, x][\text{elephant}^{\circ}(x)] >, \\
& < [e, x][\text{elephant}^{\circ}(x)] >, \\
& < [e, x][\text{play}^{\circ} e, x] > \}
\end{aligned}
$$

Underspecified Generalised form

- The LC context sets have been converted into an underspecified generalised (UG) form.
- The UG form can be expressed as a (frequency-based) vector space:

	elephant $^\circ(x)$	play $^\circ(e, x)$	eat $^\circ(e, x)$
elephant $^\circ(x)$	4	3	2
play $^\circ(e, x)$	3	3	1
eat $^\circ(e, x)$	2	1	2

Obtaining truth values from UG distributional vectors

- Such a representation allows us to trivially answer questions such as

  Does one elephant eat?
  Do more than two elephants play?
  Do three elephants play or eat?
  How many elephants are there?
  Do all elephants play?
  Do most elephants eat?

Are ideal distributions distributions?

- Ideal distributions allow us to represent extension (and perhaps intension?)
- But: ideal distributions seem far from the idea that 'words are defined by their usage'. Arguably, they are just models looking like distributions.
- **Hypothesis:** ideal distributions are a generalisation over 'actual distributions'. They correspond to linguistic competence (as opposed to actual things said).
  - Having heard *All cats are mammals*, I can produce *Molly is a mammal*.
  - Having heard *Cats/dogs/sheep/horses are mammals*, I can produce *Goats are mammals*.
  - ...

Outline

1. Distributional semantics
   - Some historical pointers
   - Building distributions
   - What are distributions good for?
   - Some notes on the representation
2. Which semantics in distributional semantics?
3. Lexicalised Compositionality (LC)
   - Ideal distributions
   - From actual to ideal distributions
4. Conclusion

PUG distributions

- Moving to probabilistic model: we do not need to observe every bird in the world to utter *All birds have beaks*.
- We can generalise the truth-theoretic model to unobserved situations by assuming **probabilistic underspecified generalised** (PUG) distributions.
- In PUG distributions, the value of $w^\circ(x)$ along $d^\circ(x_{1\ldots n})$ is the probability for an individual in w° to fill the relevant argument in d°. We will initially assume that this probability is computed over the observed individuals in the full context set.

Example

- If w_1 corresponds to an observed world (and ignoring the data sparsity issue), we have the following PUG distribution.

	elephant$^\circ(x)$	play$^\circ(e, x)$	eat$^\circ(e, x)$
elephant$^\circ(x)$	1	0.75	0.5
play$^\circ(e, x)$	0.75	1	1
eat$^\circ(e, x)$	0.5	1	1

Figure: Vectors corresponding to probabilistic underspecified generalised context sets for w_1

The distributional dependency hypothesis (Herbelot, 2013)

- Let us assume a distributional space with n dimensions.
- Let us refer to the ideal distribution of A as A°.
- We hypothesise that the value of A° along a dimension d_k is dependent on the value of A° along all other dimensions $d_{1\ldots n}$ in that space.

Intuitively...

... the probability that a cat (habitually) eats is dependent on the probability of that cat to (habitually) sleep, run, communicate, to be made of stone or to write books. In other words, the ideal distribution of a typical cat x reflects its status as a living (non-human) being, which in turn implies a high probability of cat° along the dimension *eat*.

The distributional dependency hypothesis (Herbelot, 2013)

- Using the distributional dependency hypothesis, we can
  - infer that goats are mammals from the fact that other similar animals are mammals, because they all have the same kind of values along the relevant dimensions.
  - infer that dodos (normally) have wings, as long as we know that dodos are birds, because we know that a distribution with the value 1 on the feature *bird* has a value 0.9 on the feature *have_wings*.

The distributional dependency hypothesis (Herbelot, 2013)

- Problem: we need at least one relevant feature in the ideal distribution to make inferences. E.g.: we need to have heard *dodos are birds*, converted it into *all dodos are birds* and believed it to be true.
- Encyclopedic knowledge: where does it come from?
  - Explicit information: *All dodos are birds.*
  - Result of built-in generalisation process. See psycholinguistic research on generics: children do generics before quantification.
- Observing language being used and being told things explicitely allows us to generalise.

An experiment

- A small data set of 72 animal names, with their distributions *ant, bat, beaver, bee, cat, chicken*...
- 54 features (vector components): *be_v+bird_n, be_v+insect_n, be_v+mammal_n, domestic_a, graze_v, hibernate_v, lay_v+egg_n, poisonous_a*
- The task: classifying every {animal, feature} pair into quantificational classes *no, a few, some, most all*.
- A manual annotation is performed and the data separated into training and test data.

Incremental learning

	black	fly	mammal	dog chase	crawl
cat	0.03	0	0.00015	0.05	0.01
raven	0.002	0.2	0	0.0000006	0

corpus-based distributions

learn

feature	precision
carnivorous	0.14
live on land	0.03
bird	0.9

learnt classifiers (one per feature)

	black	fly	mammal	dog chase	crawl	bird (learnt)
cat	0.03	0	0.00015	0.05	0.01	no
raven	0.002	0.2	0	0.0000006	0	all

Baseline classifier for *aquatic_a*

```
J48 unpruned tree
------------------

part_n+of_p(1) <= 0.056681
|   fascinating_a <= 0
|   |   area_n+along_p() <= 0
|   |   |   in_p()+water_n <= 0.07208
|   |   |   |   complete_v+season_n <= 0
|   |   |   |   |   mediterranean_a <= 0.055136
|   |   |   |   |   |   preferable_a <= 0.061667: no (43.0/1.0)
|   |   |   |   |   |   preferable_a > 0.061667: few (3.0)
|   |   |   |   |   mediterranean_a > 0.055136: most (2.0/1.0)
|   |   |   |   complete_v+season_n > 0: all (2.0)
|   |   |   in_p()+water_n > 0.07208
|   |   |   |   variant_n+of_p() <= 0.047993: all (12.0/1.0)
|   |   |   |   variant_n+of_p() > 0.047993: few (3.0)
|   |   area_n+along_p() > 0: most (2.0)
|   fascinating_a > 0: some (2.0/1.0)
part_n+of_p(1) > 0.056681: some (2.0)
```

Improved classifier for *aquatic_a*

```
J48 unpruned tree
------------------

part_n+of_p(1) <= 0.056681
|   fascinating_a <= 0
|   |   area_n+along_p() <= 0
|   |   |   terrestrial_a:learnt = no: all (11.0)
|   |   |   terrestrial_a:learnt = few: no (0.0)
|   |   |   terrestrial_a:learnt = some: few (2.0)
|   |   |   terrestrial_a:learnt = most: few (1.0)
|   |   |   terrestrial_a:learnt = all
|   |   |   |   on_p()+river_n <= 0.079139
|   |   |   |   |   reach_v+length_n <= 0
|   |   |   |   |   |   growth_n+in_p() <= 0
|   |   |   |   |   |   |   in_p()+water_n <= 0.08147: no (43.0/1.0)
|   |   |   |   |   |   |   in_p()+water_n > 0.08147: few (2.0/1.0)
|   |   |   |   |   |   growth_n+in_p() > 0: few (2.0)
|   |   |   |   |   reach_v+length_n > 0: most (2.0/1.0)
|   |   |   |   on_p()+river_n > 0.079139: all (2.0)
|   |   area_n+along_p() > 0: most (2.0)
|   fascinating_a > 0: some (2.0/1.0)
part_n+of_p(1) > 0.056681: some (2.0)
```

Outline

1. Distributional semantics
   - Some historical pointers
   - Building distributions
   - What are distributions good for?
   - Some notes on the representation
2. Which semantics in distributional semantics?
3. Lexicalised Compositionality (LC)
   - Ideal distributions
   - From actual to ideal distributions
4. Conclusion

Conclusion

- Distributional models have evolved from a Wittgenstinian tradition.
- The Wittgenstinian view of meaning is not conducive to doing the tasks of ‘traditional’ semantics. Moreover, it prevents us from modelling phenomena which seem to have a place in *any* theory of meaning.
- The notion of an ideal distribution preserves the standard concepts of intension and extension.
- The notion of actual distribution preserves the idea that ‘meaning comes from usage’.
- We have to find out the processes that lead from actual to ideal distributions!

Thank you!
(also to my sponsor...)

Unterstützt von / Supported by

Alexander von Humboldt
Stiftung/Foundation