Automatic clustering and the lexical semantics of cooking adjectives

Curt Anderson, Oliver Hellwig, and Wiebke Petersen
Heinrich-Heine-Universität Düsseldorf

14 September 2018
Cognitive Structures 2018 (CoSt 2018)
Düsseldorf

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

SFB 991

Outline

- Motivate the use of the bidirClus-Algorithm
- Explore the lexical semantics of one lexical field: adjectives and nouns related to food and cooking
- Show more generally how clustering methods can be used for detailed lexical semantic study

Motivation for the BidirClus-algorithm

- background: project on adjectival modification
- question: Can adjectives be clustered purely on the basis of the nouns they modify?
- primarily aim: not the best clustering algorithm but clusters that give an insight of modificational patterns

Experiment: Which noun is modified by these adjectives?

- new
- electric
- other
- classic
- small
- rental
- first
- own
- several
- fast
- european
- old
- many
- used
- expensive
- second
- private

- black
- same
- compact
- local
- german
- australian
- japanese
- american
- big
- good
- armored
- armoured
- white
- large
- fatal
- underground
- british

?

Experiment: Which noun is modified by these adjectives?

- new
- electric
- other
- classic
- small
- rental
- first
- own
- several
- fast
- european
- old
- many
- used
- expensive
- second
- private

- black
- same
- compact
- local
- german
- australian
- japanese
- american
- big
- good
- armored
- armoured
- white
- large
- fatal
- underground
- british

?

car

Experiment: Which noun is modified by these adjectives?

- new
- electric
- other
- classic
- small
- rental
- first
- own
- several
- fast
- european
- old
- many
- used
- expensive
- second
- private

- black
- same
- compact
- local
- german
- australian
- japanese
- american
- big
- good
- armored
- armoured
- white
- large
- fatal
- underground
- british

?

car

Experiment: Which noun is modified by these adjectives?

- overnight
- new
- irish
- sunken
- fast
- last
- philippine
- local
- high-speed
- regular
- other
- short
- free
- main
- daily
- little
- additional

- extra
- long
- private
- small
- many
- old
- weekly
- ancient
- direct
- expensive
- first
- major
- north
- several
- special
- turkish
- wooden

?

Experiment: Which noun is modified by these adjectives?

- overnight
- new
- irish
- sunken
- fast
- last
- philippine
- local
- high-speed
- regular
- other
- short
- free
- main
- daily
- little
- additional

- extra
- long
- private
- small
- many
- old
- weekly
- ancient
- direct
- expensive
- first
- major
- north
- several
- special
- turkish
- wooden

?

ferry

Experiment: Which noun is modified by these adjectives?

- high
- main
- busy
- residential
- narrow
- same
- suburban
- nearby
- one-way
- side
- quiet
- two-way
- few
- massive
- commercial
- local
- violent

- back
- crowded
- new
- several
- empty
- fourth
- expensive
- many
- british
- bustling
- dusty
- entire
- other
- whole
- deadly
- huge
- public

?

Experiment: Which noun is modified by these adjectives?

- high
- main
- busy
- residential
- narrow
- same
- suburban
- nearby
- one-way
- side
- quiet
- two-way
- few
- massive
- commercial
- local
- violent

- back
- crowded
- new
- several
- empty
- fourth
- expensive
- many
- british
- bustling
- dusty
- entire
- other
- whole
- deadly
- huge
- public

?

street

Experiment: Which adjective modifies these nouns?

- service
- health
- sector
- school
- interest
- transport
- opinion
- offering
- safety
- relation
- servant
- office
- spending
- policy
- inquiry
- appearance
- support

- life
- money
- official
- debt
- place
- comment
- finances
- eye
- debate
- statement
- space
- figure
- company
- hearing
- fund
- confidence
- affair

?

Experiment: Which adjective modifies these nouns?

- service
- health
- sector
- school
- interest
- transport
- opinion
- offering
- safety
- relation
- servant
- office
- spending
- policy
- inquiry
- appearance
- support

- life
- money
- official
- debt
- place
- comment
- finances
- eye
- debate
- statement
- space
- figure
- company
- hearing
- fund
- confidence
- affair

?

public

Experiment: Which adjective modifies these nouns?

- road
- day
- street
- schedule
- time
- week
- life
- summer
- intersection
- period
- shopping
- market
- year
- weekend
- night
- area
- city

- man
- highway
- people
- work
- morning
- lifestyle
- motorway
- route
- month
- stretch
- thoroughfare
- junction
- holiday
- season
- airport
- traffic
- mother

?

Experiment: Which adjective modifies these nouns?

- road
- day
- street
- schedule
- time
- week
- life
- summer
- intersection
- period
- shopping
- market
- year
- weekend
- night
- area
- city

- man
- highway
- people
- work
- morning
- lifestyle
- motorway
- route
- month
- stretch
- thoroughfare
- junction
- holiday
- season
- airport
- traffic
- mother

?

busy

The bidirClus-Algorithm

Input: adjective-noun co-occurrence matrix

	bike	boat	bus	car	ferry	ship	train
actual	0	1	0	1	0	0	1
additional	0	1	1	1	1	1	1
advanced	0	0	0	1	0	1	0
air-conditione	0	0	0	1	0	0	1
american	0	1	0	1	0	1	1
amphibious	0	0	0	1	0	1	0
ancient	0	1	0	0	1	1	0
annual	1	1	0	1	0	0	0
antique	0	0	0	1	0	0	1
argentine	0	0	0	0	0	1	1
armed	0	0	0	1	0	1	0
armoured	0	0	0	1	0	0	0
asian	0	0	0	1	0	0	0
australian	0	1	0	1	0	1	1
automatic	0	0	0	1	0	0	1
available	1	0	0	1	0	1	0
average	0	1	1	1	0	0	1
beautiful	1	0	0	1	0	1	0
belgian	0	0	0	1	0	1	0
beloved	1	1	0	1	0	0	0
big	1	1	1	1	0	1	1
black	1	0	1	1	0	1	1
blue	1	1	1	1	0	0	1
brand-new	1	0	0	1	0	0	0
british	1	1	1	1	0	1	1
broken	1	1	0	1	0	0	0
broken-down	0	0	1	1	0	0	1
burnt-out	0	0	1	1	0	0	0
busy	0	0	1	1	0	0	1
canadian	0	0	0	1	0	0	1
central	0	0	1	0	0	0	1

The bidirClus-Algorithm

Cluster step: cluster rows by similarity

	n1	n2	n3	n4	n5	n6	n7	n8
a1	1	0	0	1	1	0	0	0
a2	0	1	0	0	1	1	0	1
a3	1	0	0	1	1	1	0	0
a4	0	0	0	0	0	0	1	1
a5	1	0	1	0	0	0	1	0
a6	0	0	0	1	1	0	0	0
a7	0	1	1	0	0	1	0	0
a8	1	1	0	0	1	0	0	1

The bidirClus-Algorithm

Cluster step: cluster rows by similarity

	n1	n2	n3	n4	n5	n6	n7	n8
a1	1	0	0	1	1	0	0	0
a2	0	1	0	0	1	1	0	1
a3	1	0	0	1	1	1	0	0
a4	0	0	0	0	0	0	1	1
a5	1	0	1	0	0	0	1	0
a6	0	0	0	1	1	0	0	0
a7	0	1	1	0	0	1	0	0
a8	1	1	0	0	1	0	0	1
ca1(a1,a3,a6)	1	0	0	1	1	0	0	0

The bidirClus-Algorithm

bidirectional: transform matric after each clustering step

	a2	a4	a5	a7	a8	ca1
n1	0	0	1	0	1	1
n2	1	0	0	1	1	0
n3	0	0	1	1	0	0
n4	0	0	0	0	1	1
n5	1	0	0	0	1	1
n6	1	0	0	1	0	0
n7	0	1	1	0	0	0
n8	1	1	0	0	1	0

The bidirClus-Algorithm

bidirectional: transform matric after each clustering step

	a2	a4	a5	a7	a8	ca1
n1	0	0	1	0	1	1
n2	1	0	0	1	1	0
n3	0	0	1	1	0	0
n4	0	0	0	0	1	1
n5	1	0	0	0	1	1
n6	1	0	0	1	0	0
n7	0	1	1	0	0	0
n8	1	1	0	0	1	0
cn1(n1,n4,n5)	0	0	0	0	1	1

The bidirClus-Algorithm

stop condition: similarity treshold not reached or compression treshold reached

	a2	a4	a5	a7	a8	ca1
n2	1	0	0	1	1	0
n3	0	0	1	1	0	0
n6	1	0	0	1	0	0
n7	0	1	1	0	0	0
n8	1	1	0	0	1	0
cn1	0	0	0	0	1	1

The bidirClus-Algorithm

Advantages:

- transparent: information on which modificational pairs led to clustering is avaiable
- transparent: clustering steps can be recorded in a history
- research question specific: the only used information is the information about modificational pairs
- good results: with some enhancements standard clustering methods are outperformed

(Petersen & Hellwig IWCS 2017, Hellwig & Petersen Coling 2016)

Methods

- Extracted pairs of `JJ-[NN|NE|NNS]` from the 2013 English news dump from `www.statmt.org` (Bojar et al., 2014).
- 1,048,653 A+N pairs with 8,392 adjective lexemes, 17,560 noun lexemes, and 430,256 unique combinations of A+N were obtained
- Manual exploration of the space of clusters.
- Many clusters related to food and cooking were discovered, although they were not targeted by the algorithm in any way.
- Manually extracted these clusters for further analysis.
- Not task specific data! Interesting to see from a general corpus whether specific conceptually related adjectives cluster well.

Why cooking?

- Adjectives and nouns related to food preparation and food generally are a useful space to begin this sort of inquiry.
- Relatively clear judgements about when words form clusters or not, and about how they relate to each other.
- Mainly looking at deverbal adjectives; already have insight into their structure.
- Because our aims are partially methodological, removing other confounding factors is useful.
- Some work already done on cooking as a semantic field. See Lehrer 1969, who develops detailed lexical semantic analysis of groups of cooking verbs.

Output of algorithm

- Output of clustering algorithm are sets of clusters of adjectives and nouns.
- Example (abbreviated):

```
_class_442 [ southerly northerly westerly easterly] {breeze (4),
course (4), direction (4), flow (4), gale (4), tip (4), wind (4),
component (3), current (3), island (3), outpost (3), region (3),
track (3), air (2), airflow (2), bit (2), farm (2), garden (2),
isle (2), population (2), storm (2), swell (2), weather (2),
airport (1), archipelago (1), blast (1), bluster (1), boundary (1),
campsite (1), category (1), cathedral (1), charge (1), cinema (1),
climate (1), clime (1), colleague (1), constituency (1), county
(1), edge (1), … }
```

Clusters based on paradigms

- Often find adjective clusters based on related words:

(1)
a. southerly, northerly, westerly, easterly
b. low-rise, high-rise
c. zimbabwean, haitian, cambodian, honduran, tanzanian, singaporean, bolivian, nepalese, panamanian, malian, ugandan, ...
d. anti-jewish, anti-catholic, anti-islam, anti-religious, anti-israel, anti-immigrant, ...

- Hypothesis: adjectives that differ in the value of a single attribute.

Example

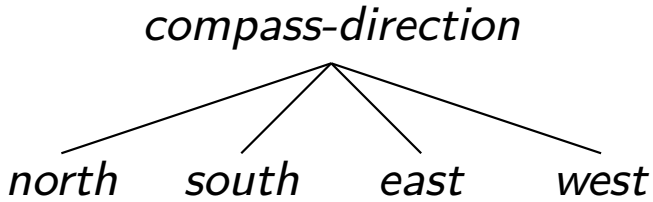
- Analyze *{south, north, west, east}erly* as specifying a DIRECTION attribute of the noun they modify.
- Evidence for this comes from the nominals these modify.

  (2) breeze, course, direction, flow, gale, tip, wind, component, current, island

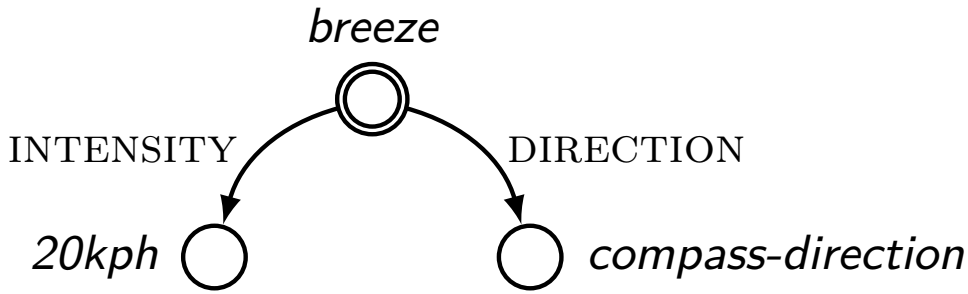
- Many of these nouns can be thought of as having directed motion (*breeze*, *course*, *flow*).

Example

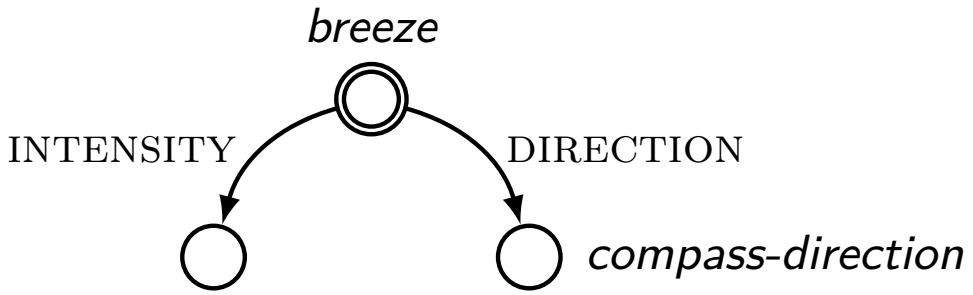
- Assume a type *compass-direction*, with subtypes for particular cardinal points.

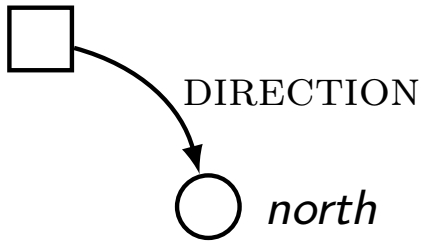


- *Breeze* (for instance) can be assumed to have a DIRECTION attribute in its frame.

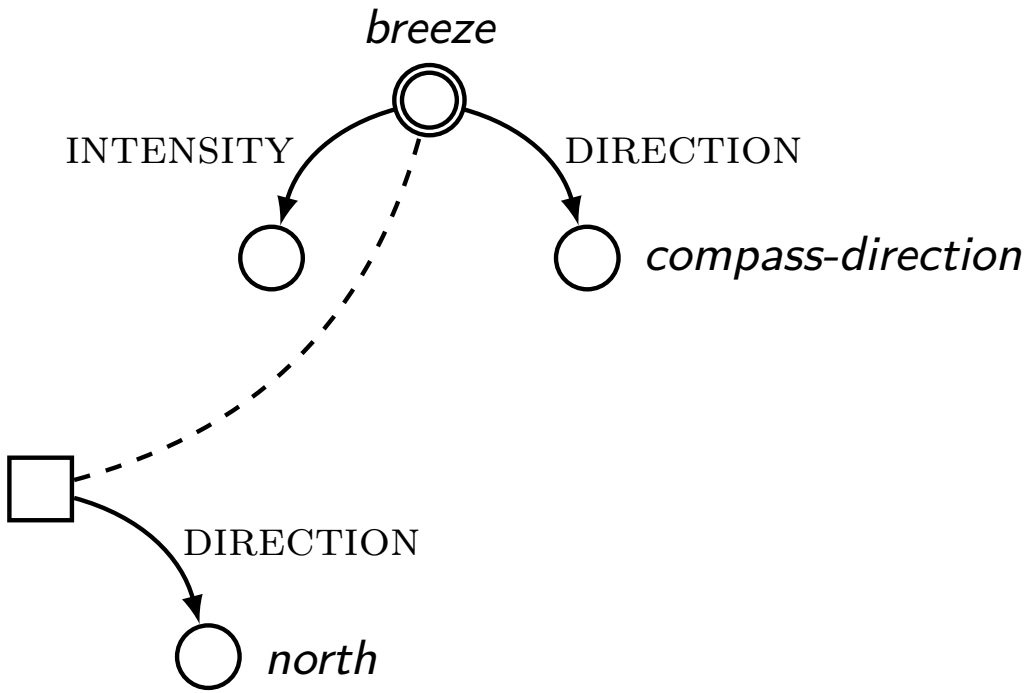


- *Northerly* requires something with a DIRECTION attribute
- In the course of modification unifies with the frame of the modifiee frame, constrained by the type information in each node.
- Referential node of the resulting frame will usually be the referential node of the modified NP (cf. headedness).
- Example: *northerly* frame subsumes the *breeze* (*northerly breeze*)





- *Northerly* requires something with a DIRECTION attribute
- In the course of modification unifies with the frame of the modifiee frame, constrained by the type information in each node.
- Referential node of the resulting frame will usually be the referential node of the modified NP (cf. headedness).
- Example: *northerly* frame subsumes the *breeze* (*northerly breeze*)



Notes

- Understanding examples like *northerly* helps understand other examples.
- Adjectives derived from cooking and food preparation verbs can also be considered to form clusters based on a single attribute.

Food preparation I

- One set of food preparation clusters involves changes of state in the food being cooked (e.g., not cooked to cooked).

Type	Adjectives
SIMMER	stewed, curried
FRY/BROIL/BAKE	undercooked, pan-fried
	grilled, baked, fried, roasted, deep-fried, crispy

Table: Result state of cooking clusters

- FRY/BROIL/BAKE differ in result state, while SIMMER differ in a specification of the liquid.

Food preparation I

- SIMMER class involves simmering some food item in a liquid as part of the cooking process.

  (3) a. stewed goat
  b. curried lentils

- Primary way *stewed* and *curried* differ is in what something is cooked in.

Food preparation I

- The FRY/BROIL/BAKE class differs in being a specification of the result of the cooking.
- Involve application of heat to the food item, and thus cooking.
- Different result states of this process (fried, baked, etc.)

Food preparation II

- Another set of adjective clusters also related to food preparation.

Type	Adjectives
PREPARE	pickled
subtype: NON-INTEGRAL	sliced, diced
CONTAINER	tinned, canned

Table: Food preparation clusters

- These clusters also involve a change of state, but of a different kind.
- Specify change of location of item (*tinned*, *canned*) or a change in a property of the affected object.
- These are not derived from frames for cooking per se, but other changes of state. Do not always entail the object has cooked!

Event structure in adjectives

- We adopt the basic form for change of state verbs in Kallmeyer & Osswald 2014.
- CoS verbs have a bipartite frame structure inspired by event decomposition approaches to verb meaning (Rappaport Hovav & Levin, 1998).

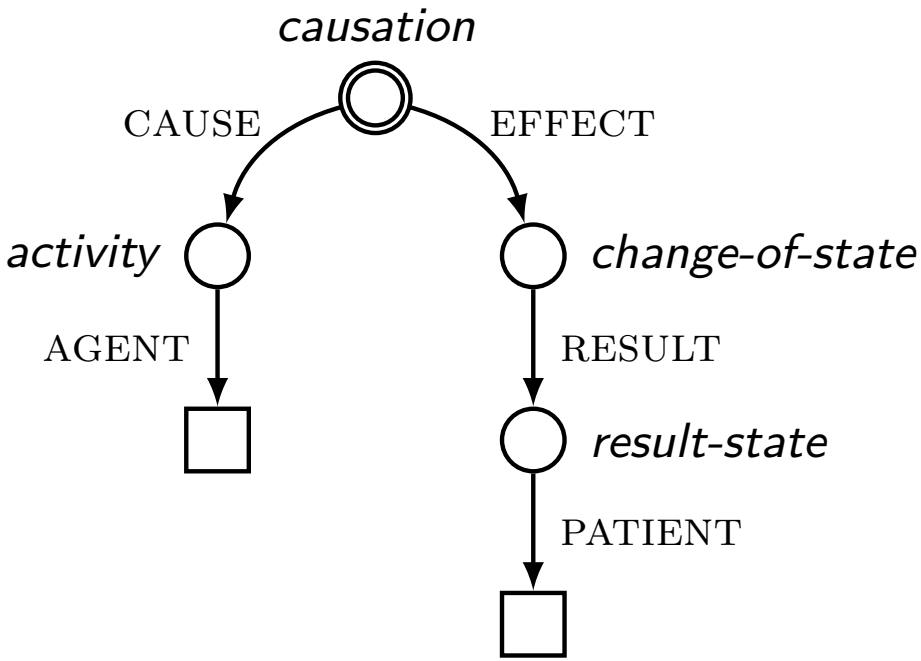
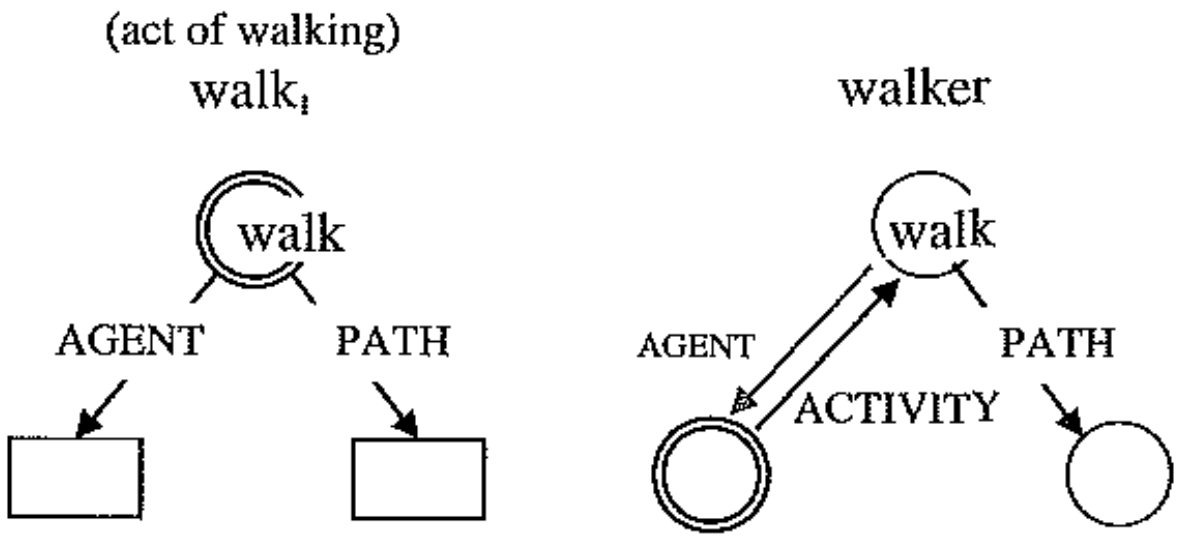


Figure: Frame for change of state verb

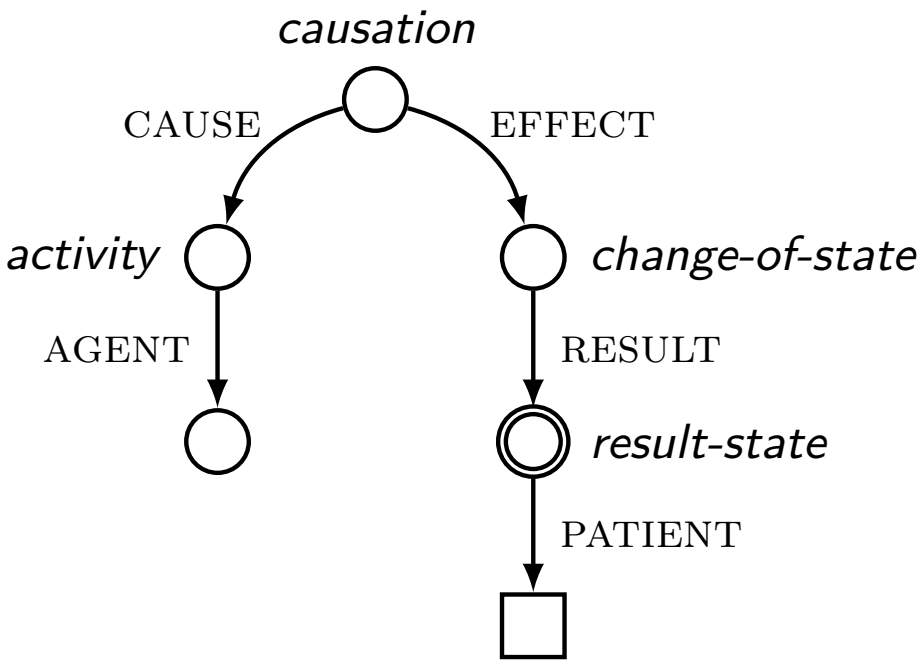
Deverbal adjectives

- Following other work in frame semantics, changing of a category can be seen as a shift in the root node of the frame.
- Example: The nominalizing suffix *-er* causes a shift from the event node to a participant node.



Deverbal adjectives

- View deverbal adjectives as having undergone a shift to the result state node in the verbal frame.
- Follows views in the formal semantics literature that passive participles have undergone a process of stativization (e.g., Kratzer 2000).



Analysis of clusters

- We present an analysis of several prototypical examples

(4)	pan-fried aubergine	(FRY/BROIL/BAKE)
(5)	crispy chicken	(FRY/BROIL/BAKE)
(6)	stewed goat	(SIMMER)
(7)	canned peaches	(CONTAINER)

- Show how composition between modifier and modifiee occurs.
- Demonstrates how members of clusters can vary along a single attribute.

Example: *pan-fried aubergine* (FRY/BROIL/BAKE)

- *Pan-fried aubergine* shows two processes: specification of the result state of cooking, and an additional specification of an instrument.
- Derived from the verb *pan-fry*, where the incorporated nominal *pan* restricts the type of the instrument (on the assumption that *pan* $\sqsubseteq$ *cooking-utensil*).
- We also note that in the course of deriving the deverbal adjective, the AGENT node is generally unavailable for further composition.

  (8) *pan-fried chef

- Type information on the PATIENT node permits identification of *aubergine* with the PATIENT.

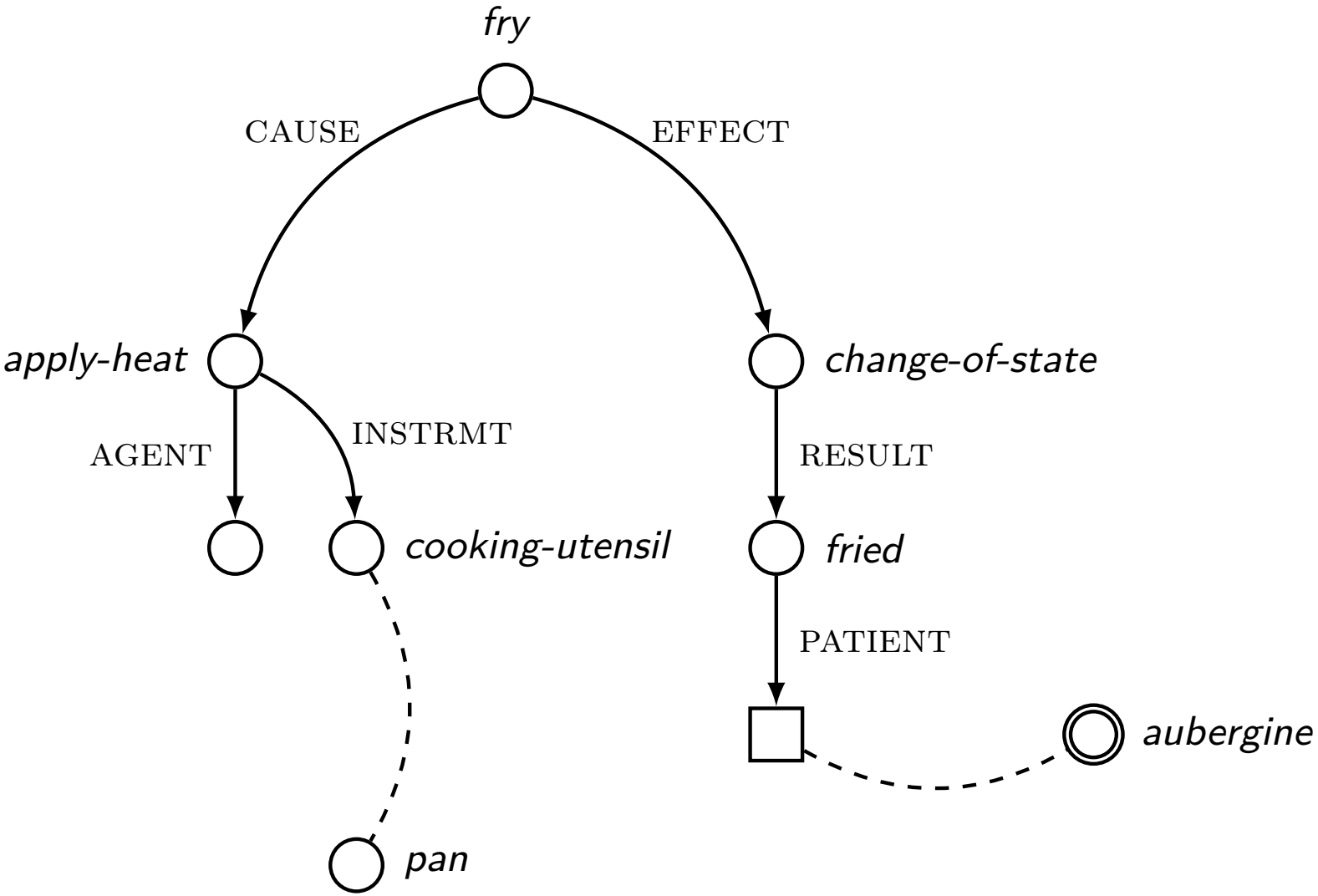


Figure: Frame for *pan-fried aubergine*

Example: *crispy chicken* (FRY/BROIL/BAKE)

- *Crispy chicken* differs from other FRY/BOIL/BAKE verbs in that the manner (*fry*, *bake*, etc.) is left unspecified.
- However, where it is the same is in overtly encoding a specification of the result of the cooking (being crispy).
- Additional support for *crispy* encoding an event from BNC. *crispy+N* combinations usually food-related, and with cooked food.

(9) crispy pancakes, batter, crumb, duck, crust, edges, foliage, garlic, lettuce, mustard, pancake, prawn, pretzels, ratatoui, seaweed, shell, surface, toast, veg

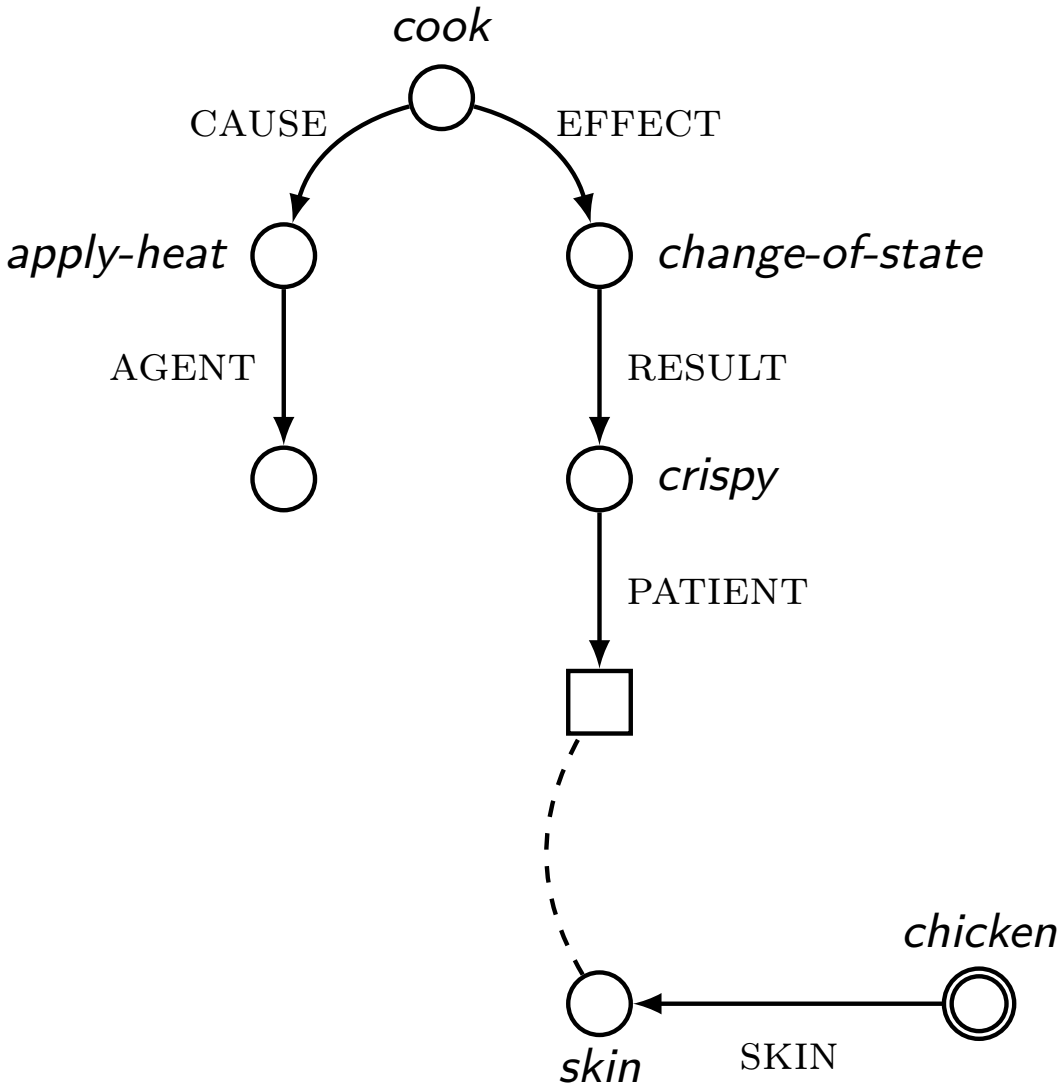


Figure: Frame for *crispy chicken*

Example: *stewed goat* (SIMMER)

- *Stewed goat*, as part of the SIMMER class, has in its frame that the food is cooked in a liquid.
- Informal survey of recipes finds reference to simmering in a liquid (often tomato-based for *stewed goat*)
- Additional complication, similar to *crispy chicken*: it is the meat of the goat that is stewed, not the whole goat.
- Introduce an attribute MEAT for the edible portion of the goat (cf. Anderson & Andreou, this conference).

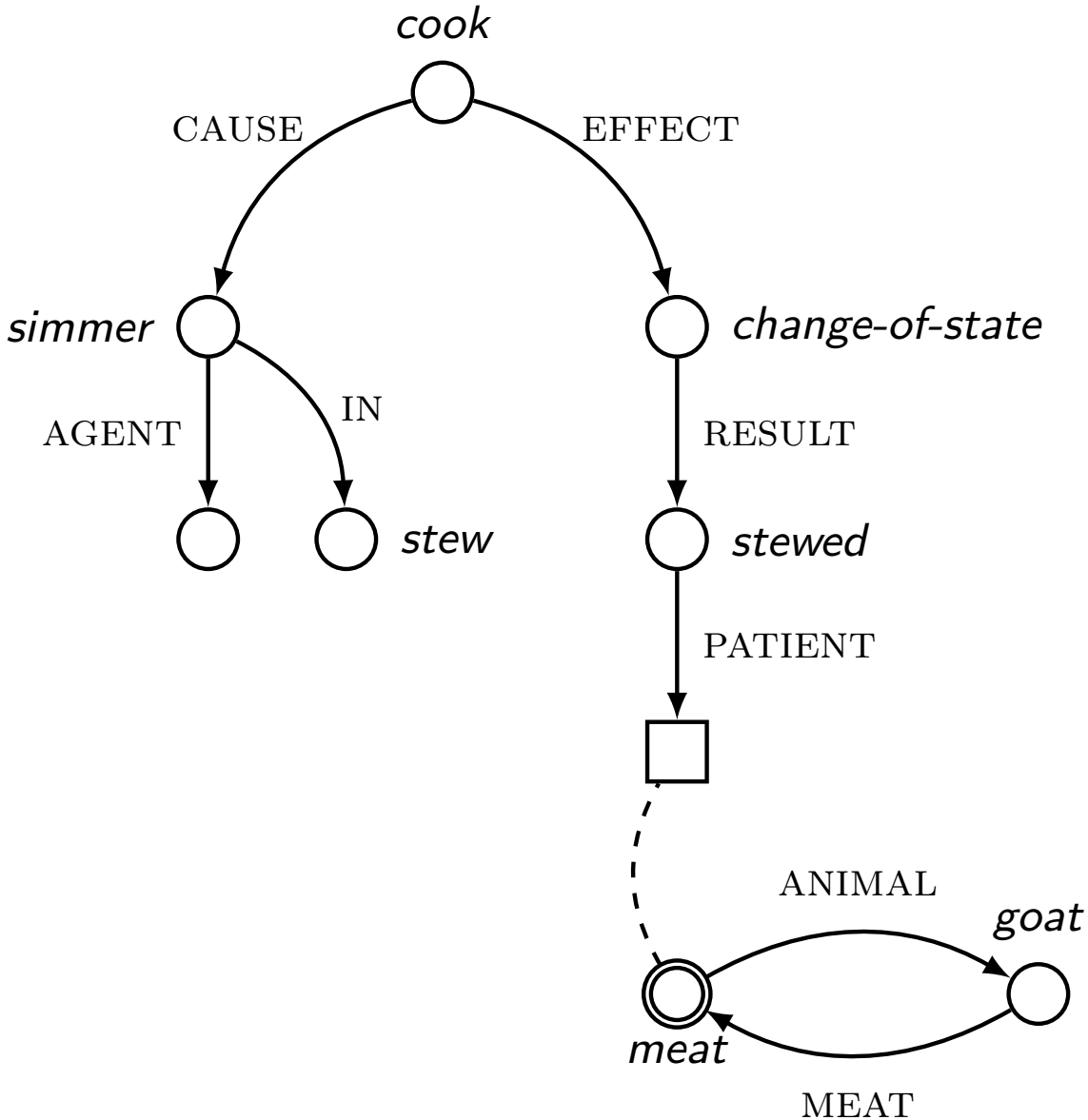


Figure: Frame for *stewed goat*

Example: *canned peaches* (CONTAINER)

- Members of the CONTAINER class specify a container into which the food item goes.
- These adjectives do not always entail that the food has been cooked as part of the process...

  (10) When canning peaches you can either raw or hot pack. (Google)

- ...although there are implications for cooking depending on the food (such as *canned tuna*)

  (11) Tinned fish is prepared, packaged, sealed and then pressure-cooked in the can (Google)

- Relevant attribute for clustering is what kind of container the food is in.

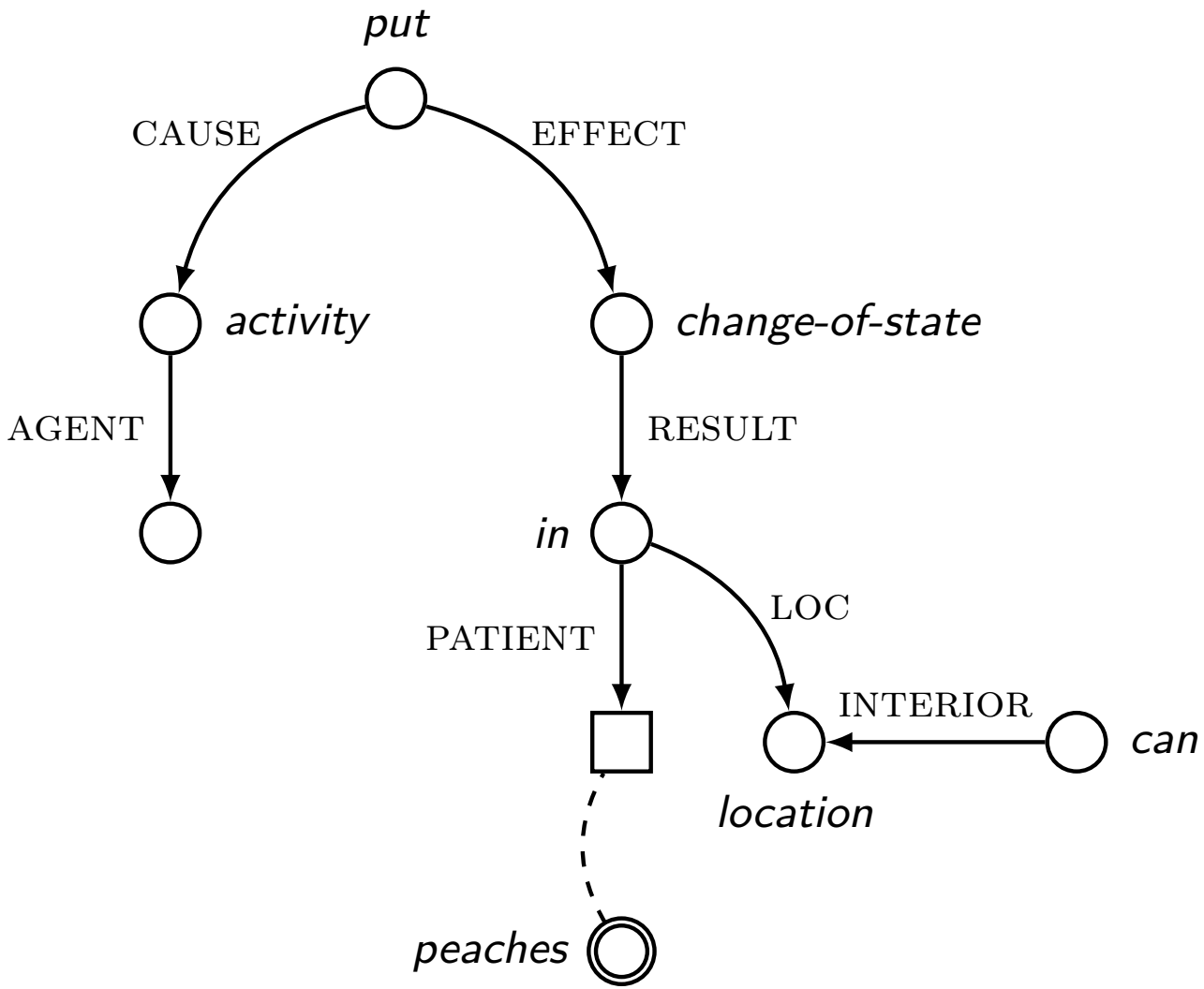


Figure: Frame for *canned peaches*

Conclusion

- Corpus methods already commonly used in theoretical linguistics for discovery of example sentences, support for acceptability judgements, and so on
- Clustering methods provide a type of data especially useful for lexical semantic exploration, as clusters are often based on the valuation of a single attribute.
- Being able to identify the attribute in common between words within a cluster allows for precise investigation of the lexical semantics of classes of words.
- Type information on the nodes plays a role in constraining interpretations.
- Motivation for the BidirClus-Algorithm.

Acknowledgements

Thank you!

This research is supported by DFG CRC 991 “The Structure of Representations in Language, Cognition, and Science,” project C10.

`https://frames.phil.uni-duesseldorf.de/c10/`

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

SFB 991

References I

Bojar, Ondrej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, 12–58.

Kallmeyer, Laura & Rainer Osswald. 2014. Syntax-driven semantic frame composition in lexicalized tree adjoining grammars. *Journal of Language Modelling* 1(2). 267–330.

Kratzer, A. 2000. Building statives. In *Berkeley Linguistic Society 26*, 385–399.

Lehrer, Adrienne. 1969. Semantic cuisine. *Journal of Linguistics* 5(1). 39–55.

Rappaport Hovav, Malka & Beth Levin. 1998. Building verb meanings. In Miriam Butt & Wilhelm Geuder (eds.), *The projection of arguments: Lexical and compositional factors*, 97–134. Stanford, CA: CSLI Publications.