Informational nouns: Nominal complementiser clauses and polysemy

Peter R. Sutton
Universität Potsdam

peter-sutton.github.io/talks

Informational Nouns

Informational Nouns

- nouns with at least one sense that denotes a piece, pieces, a body or bodies of information
  - which could be modelled as e.g., a proposition, some collection of propositions, Record Types...

Examples of informational nouns in English are given in 1

(1) *allegation*, *belief*, *book*, *fact*, *information*, *knowledge*, *newspaper*, *report*, *statement*

Polysemy and Informational Nouns

Many Informational Nouns are polysemous

- have multiple interrelated senses

Examples:

- *book* is (at least) 2-ways polysemous
  - the physical book made of paper, and an informational sense, the body of information that is the contents of the book (see e.g., Pustejovsky 1995; Asher 2011)
- *statement* is (at least) 3-ways polysemous
  - a stating event or a physical artefact (e.g., a written document) and the informational contents of each of these
- *fact* is not polysemous
  - it intuitively seems to have only an informational sense
  - although there is debate about what facts are (Austin vs. Strawson)

Noun-related propositional complementiser clauses (NCCS)

I adopt the neutral term *noun-related complementiser clause* NCC from Müller 2023

Not all nouns can be used with NNCs

- Call the construction in which a noun is used with an NNC, the COMP *environment:*

(2) The (COMP) environment. Examples:

a. the allegation/belief/evidence/fact/information/report/statement that Bilbo found the ring
b. #the house that Bilbo met Gandalf
c. #the event/celebration that Gandalf set off fireworks
d. #the book/brochure/document/newspaper that Bilbo found the ring

Some evidence of cross-linguistic robustness:

- Finnish, German, Italian, Spanish

Puzzle

NNCs clearly relate to propositions/informational entities etc.

- Naïve hypothesis: Informational Nouns can be felicitously used with NNCs in the COMP environment
- Correctly predicts *#the house/event/celebration that...*
- Problem: does not predict *#the book/brochure/document/ newspaper that...*

What is the right restriction for licensing NNCs?

- Elliott (2020): NNCs are only defined for "contentful entities"
- "The natural conclusion to draw from such contrasts is that the linguistic notion of a contentful entity is distinct from the intuitive notion." (p.63)

The locative strategy for referring to informational contents

Some languages use locative PPs to target the informational contents of nouns like *book*:

(3) Context: Andrew is a character in the book:

a. #the book that Andrew lied
b. the book in which Andrew lied

(4) el libro en el que Andrés mintió
the.M book in REL.M that Andrés lied
"the book in which Andrés lied"

(5) das Buch in dem Andreas gelogen hat
the.N book in REL.N.DAT Andreas lied has
"the book in which Andreas lied"

The locative strategy as an alternative to COMP

The information locative PP environment has a near complementary distribution with COMP

(6) The information locative PP environment (LOC). Examples:

a. the information in this book/brochure/document/newspaper/report/statement

b.??the information in this allegation/belief/evidence/fact/information

Only COMP nouns

- allegation, belief, evidence, fact, information

Only LOC nouns

- book, brochure, document, newspaper

Both COMP and LOC nouns

- report, statement

Hypothesis

(H1) The additional senses (if any) that an informational noun has, determine whether that noun is felicitous in COMP, LOC, or both COMP and LOC.

(7) If a noun is, minimally INF-PHYS polysemous it can be used in LOC. This includes:

a. *book*, *newspaper*, which are INF-PHYS polysemous;
b. *statement*, *report*, which are INF-PHYS-EV polysemous

(8) If a noun minimally has an INF sense, but is not INF-PHYS polysemous, it can be used in COMP but not LOC. This includes:

a. purely informational nouns such as *fact*
b. *allegation*, *belief*, *statement* which are INF-EV polysemous

Main questions

(Q1) To what extent do different types of nominal polysemy predict the licensing conditions for the COMP and LOC environments?

(Q2) What implications does the answer to Q1 have on the semantics of NCCs, and on how different informational nouns should be semantically classified?

Outline

- Corpus study
  1. Define a class of nouns that are felicitous in COMP and/or LOC
  2. Annotate these nouns for senses (e.g., physical entity, eventuality)
  3. Assess whether 2 predicts 1 (Conditional Inference Tree (CIT) analysis)
- Refine the CIT model with additional parameters
- Formal semantic analysis
  - How polysemous nouns of different sorts and NNCs compose

Corpus information

The UK Web Annotated Corpus (ukWaC)

(Ferraresi et al., 2008; Baroni et al., 2009)

- 1,547,594,305 tokens
- POS tagged, but not dependency parsed
- Accessed via SketchEngine (`https://www.sketchengine.eu`)
  - The UkWaC corpus is also available on request `https://wacky.sslmit.unibo.it`

Methodology Overview

1. API to pull (noisy) corpus data via CQL queries for COMP and LOC environments.
2. Dependency parsing using the *spaCy* Python package (Honnibal and Montani, 2017). Use a Python script to identify the nouns in these constructions. Manual checking and cleaning
3. Sense annotation of nouns using listed senses scraped from `www.wiktionary.org`
   - All nouns assumed to have an informational sense
   - Annotation only for ±P(hysical entity) and ±E(ventuality)
4. Conditional Inference Tree modelling
   - Probabilistic model that can be used to identify which nouns are correctly/incorrectly predicted to be felicitous in the COMP and LOC environments by the hypothesis

API requests

COMP

- Any sentence containing *N that DP*
  - Nb. Rel clauses filtered later after dependency parsing

LOC

- Any sentence containing *information in this/that* followed by an NP, with optional adjective and that could be a compound

For details, see the Appendix

Dependency parsing

COMP

- 1,078,043 corpus hits
- a random sample of 30,000 of these sentences was dependency parsed

LOC

- 1653 corpus hits (after duplicates removed)
- All sentences dependency parsed

Cleaning: COMP

Not possible to differentiate the following structures from the dependency parse information

(9) a. Comment on the fact that they quit. (good case)
b. Stop in the event that they quit. (bad case)

- the script was adjusted to exclude all cases where the target noun is the object of a preposition

Other frequent false positives removed:

(10) a. so much/little N that ...
b. such an A N that ...

After duplicates were removed and the algorithm applied:

- 6935 sentences, 607 unique noun lemmas
- 75 most frequent carried forward for further analysis

Cleaning: LOC

Output of python script

- 204 unique noun lemmas, the most frequent 75 were further analysed

Excluded lemmas included:

- *section*, *chapter*, *column*, *paragragh*, *page*, *part*
  - Labour-saving measure: Not bad, per se, but all parts of larger informational entities (e.g., books)
- *area*, *category*, *class*, *connection*, *context*, *format*, *manner*, *matter*, *regard*, *respect*, *way*.
  - Classified types of information or relations between information (e.g. *the information in this connection was useful*

The cleaning process left 50 unique noun lemmas and 732 sentences

Re-running the corpus search

Lemmas from each environment searched for in the other

- Since I only took a sample of most frequent lemmas, nouns that appear in both COMP and LOC could be missed
- Re-running the process on all identified lemmas ensured finding all that are felicitous in both environments

An issue with *issue*

- One noun identified in both environments as a result was *issue*
- But clearly a case of lexical ambiguity
  - *issue*: problem/topic (*Sachverhalt*, *Problem*, *Thema*)
  - *issue*: volume/tome (*Ausgabe*)
- These listed as *issue-1* (felicitous in COMP) and *issue-2* (felicitous in LOC)

Lemma lists

(11) *Only*-COMP nouns:
advantage, advice, agreement, allegation, argument, assertion, assumption, assurance, belief, chance, claim, complaint, concern, conclusion, confidence, confirmation, consensus, contention, conviction, danger, decision, declaration, doubt, evidence, expectation, fact, fear, feeling, guarantee, hint, hope, hypothesis, idea, impression, indication, information, issue-1, knowledge, likelihood, news, notice, notion, opinion, order, perception, point, possibility, principle, probability, proof, proposal, question, realisation, reassurance, recognition, recommendation, reminder, requirement, risk, rumour, sense, sign, signal, speculation, suggestion, suspicion, theory, thought, view, warning

(12) *Only*-LOC nouns:
appendix, article, book, booklet, box, briefing, brochure, catalogue, database, directory, document, edition, factsheet, field, file, form, guide, handbook, issue-2, leaflet, list, manual, newsletter, pack, paper, prospectus, publication, series, site, table, text, volume, webpage, website

(13) *Both*-COMP&LOC nouns:
announcement, application, bulletin, case, email, entry, guidance, letter, message, note, (press) release, report, review, statement, story

Sense Annotation

An independent means of annotating each noun for having either an eventuality-denoting sense, a physical entity-denoting sense, or both was needed

- I used a free version of the Lingua Robot API which provides access to JSON-formatted English wiktionary.org entities
- Where the Wiktionary entry was not clear, or where there seemed to be at least one sense missing, the relevant entry in the Oxford English Dictionary (`https://www.oed.com`) was checked

Sense Annotation: criteria

Physical entity denoting senses

- referring to objects (as with *box*, above) or animate individuals
- expressions such as *written* or *in writing*
- digital entities (*email*, *website*): treated as physical entities.

Eventuality denoting senses

- reference to acts, actions, events, situations, circumstances, processes or (mental) states
- the use of other psych nouns and light verb constructions were taken to indicate an eventuality qua mental state

Results of sense annotation

Assuming that all of the nouns have an informational sense:

(14) [+I, -E, -P] Inf nouns without eventuality or physical entity senses:
advice, case, chance, danger, fact, hint, hypothesis, information, issue-1, likelihood, news, point, principle, probability, requirement, rumour, theory

(15) [+I, +E, -P] Inf nouns with an eventuality sense, but no physical entity sense:
advantage, allegation, argument, assertion, assumption, assurance, belief, briefing, claim, complaint, concern, conclusion, confidence, consensus, contention, conviction, decision, doubt, expectation, fear, feeling, guidance, hope, idea, impression, knowledge, notion, opinion, perception, possibility, question, realisation, reassurance, recognition, recommendation, risk, sense, speculation, story, suggestion, suspicion, thought, view

(16) [+I, -E, +P] Inf nouns with a physical entity sense, but no eventuality sense:
appendix, article, book, booklet, box, brochure, catalogue, database, directory, document, edition, email, entry, factsheet, field, file, form, guide, handbook, issue-2, leaflet, letter, list, manual, newsletter, note, pack, paper, press release, prospectus, publication, release, series, site, table, text, volume, webpage, website

(17) [+I, +E, +P] Inf nouns with both a physical entity sense and an eventuality sense:
agreement, announcement, application, bulletin, confirmation, declaration, evidence, guarantee, indication, message, notice, order, proof, proposal, reminder, report, review, sign, signal, statement, warning

Conditional Inference Tree Analysis

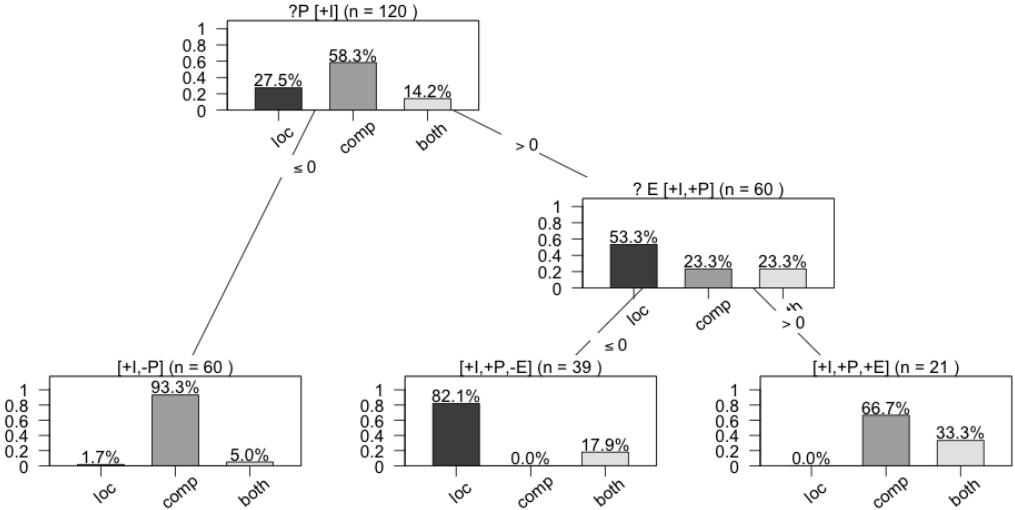
CIT tree analyses are a form of recursive partitioning method

- The algorithm searches for ways of splitting the data (in this case the set of nouns), given the available predictors, in this case $\pm E$ and $\pm$P
- Splits are only chosen if they exceed a default significance value of $p = 0.05$ for rejecting the global null hypothesis (which is derived via randomly permuting the values of the response variables).
- The final output leaves terminal nodes of distributions over predicted response variables (in our case only-COMP, only-LOC or both-COMP&LOC).
- Predictions for the model are derived by taking the category with the highest probability.

Stats package information

- Tree (CIT) analysis was applied using the PARTYKIT package in R (Hothorn and Zeileis, 2015)
- The model was run with the following inputs (all other parameters taken as defaults):

(18) `data.cit <- ctree(diagnostic ~ ev + phys, data = data)`


?P [+I] (n = 120)
27.5%
58.3%
14.2%
loc
comp
both
≤ 0
> 0
? E [+I,+P] (n = 60)
53.3%
23.3%
23.3%
loc
comp
> 0
≤ 0
[+I,-P] (n = 60)
1.7%
93.3%
5.0%
[+I,+P,-E] (n = 39)
82.1%
0.0%
17.9%
[+I,+P,+E] (n = 21)
0.0%
66.7%
33.3%

Hypothesis recap

(H1) The additional senses (if any) that an informational noun has, determine whether that noun is felicitous in COMP, LOC, or both COMP and LOC.

(19) If a noun is, minimally INF-PHYS polysemous it can be used in LOC. This includes:

a. *book*, *newspaper*, which are INF-PHYS polysemous;
b. *statement*, *report*, which are INF-PHYS-EV polysemous

(20) If a noun minimally has an INF sense, but is not INF-PHYS polysemous, it can be used in COMP but not LOC. This includes:

a. purely informational nouns such as *fact*
b. *allegation*, *belief*, *statement* which are INF-EV polysemous.

Identifying bad predictions

Table: Actual category vs. predicted category in the model

		Actual category			
		both	comp	loc	Errors
Predicted category	both	0	0	0	0
	comp	10	70	1	11
	loc	7	0	32	7
	Errors	17	0	1	18

Summary

- Total failure to predict when a noun is felicitous in both COMP and LOC (17/18 errors)
- Mostly a result of over-extending the *only*-COMP prediction (10/18)
- Suggests that [+I, +P, +E] is not a sufficiently good predictor of this

Next steps

- Deeper dive into where the model is going wrong
- Can we identify any other predictors?

Where did the model go wrong?

(21) Nouns which are [+I,+P, +E] and felicitous in both the COMP- and LOC-environments, but falsely predicted to be only COMP.
announcement, *application*, *bulletin*, *message*, *report*, *review*, *statement*.

(22) Nouns which are [+I,+P, -E] and felicitous in both the COMP- and LOC-environments, but falsely predicted to be only LOC.
email, *entry*, *letter*, *note*, *(*(press) release), *text*

(23) Nouns which are [+I,-P,±E] and felicitous in both the COMP- and LOC-environments, but falsely predicted to be only COMP.
case, *guidance*, *story*.

Nouns in (22):

- Possibly excepting *(diary) entry*, these nouns denote artefacts that are used as messages/missives.
- That these artefacts are used to inform somebody of something (e.g., *message*) could be eventuality-related
- Some overlap with some of the nouns in (21), e.g., *bulletin*, *message*

New predictor: Missive Nouns

Missive Nouns:

- (An) N we was sent out in order to inform x of y.

(24) Missive nouns (nouns that are [+M]):
announcement, *bulletin*, *declaration*, *email*, *entry*, *letter*, *message*, *note*, *notice*, *order*, (*press*) *release*, *proposal*, *reminder*, *report*, *review*, *statement*, *text*, *warning*

Another source of error

Original motivation for the hypothesis:

- Nouns such as *message* and *statement* can be informational, eventuality-denoting, or denote physical objects
- These nouns are felicitious in both COMP and LOC
- The following data was unexpected:

(25) [+I, +E, +P] nouns that are felicitous in the COMP environment, but are not felicitous in the LOC environment:
agreement, *confirmation*, *declaration*, *evidence*, *guarantee*, *indication*, *notice*, *order*, *proof*, *proposal*, *reminder*, *sign*, *signal*, *warning*

Possible pattern: Some of these nouns

- *confirmation*, *evidence*, *guarantee*, *indication*, *proof*, *reminder*, *sign*, *signal*, *warning*

Seem to be of a different nature that nouns such as *message* and *statement*

Examples: *proof, sign*

- An object/eventuality can be proof/a sign that supports some proposition.
  - e.g., a fingerprint or lying can be proof/a sign that someone is guilty
- So-called *subject nominals* such as *proof*, *sign* have been identified as outliers in the NNC literature (see e.g., Safir 1985; Moulton 2009; Müller 2023)

(26) The best proof that John was not lying is that he was here last night. (Higgins 1973)

> *"It is clearly not plausible that both that-clauses refer to the same object or describe the content of the same object [...] They refer to two different things: one thing that is considered to be the proof, that is the thing by which something is proven, and another thing that is proven by that" (Müller 2023 p.93)*

Nouns such as *proof* and *sign* are arguably examples of Saussurian signs which have a *signifier* and a *signified*

Signifiers

To paraphrase Müller, that by which something is proven, and that which that thing proves

- This contrast can be seen in (27) and (28):

(27) a. The fingerprint is proof/a sign/an indication/a confirmation/a reminder of Alex's innocence.
b. [The] Oklahoma [bombing] was a warning that cancer was spreading in the body politic. [(ukWaC, 939890)]

(28) a. ?The handshake was an agreement between them.
b. ?Alex saying *p* was a statement that *p*.
c. ?The plaque on the fence was a notice that the owner has a dog.

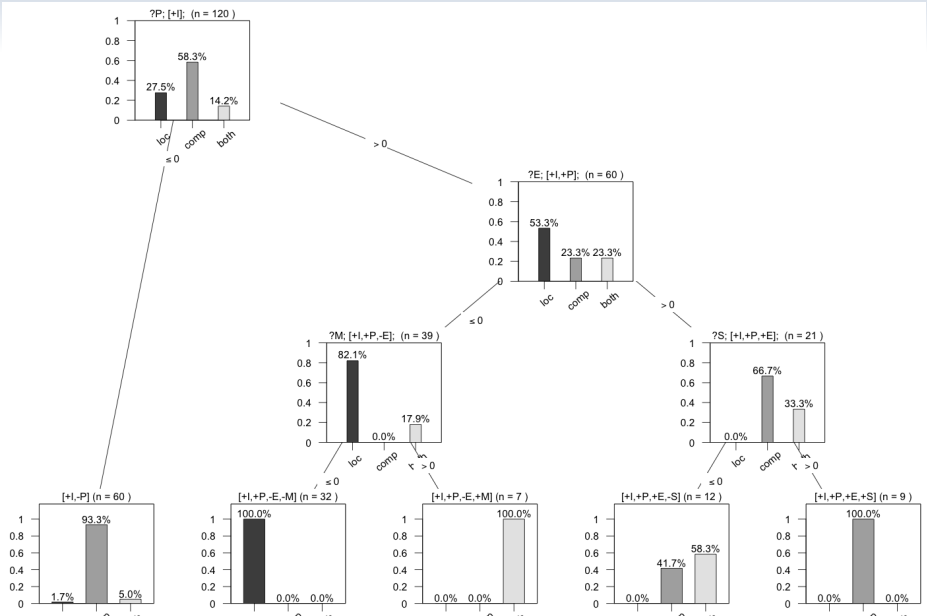
Annotating nouns as signifiers:

(29) Nouns that denote *signifiers* (nouns that are [+S]):
confirmation, *danger*, *evidence*, *guarantee*, *hint*, *indication*, *proof*, *reminder*, *sign*, *signal*, *warning*

Re-running the CIT analysis with extra predictors

The CIT model was re-run taking ±S(ignifier) and ±M(issive) as extra predictors:

(30)
```
data.cit <- ctree(diagnostic ~ ev + phys + signifier
                          + missive, data = data)
```


?P; [+I]; (n = 120)
27.5%
58.3%
14.2%
loc
comp
both
≤ 0
> 0
?E; [+I,+P]; (n = 60)
53.3%
23.3%
23.3%
?M; [+I,+P,-E]; (n = 39)
82.1%
0.0%
17.9%
?S; [+I,+P,+E]; (n = 21)
66.7%
33.3%
0.0%
[+I,-P] (n = 60)
1.7%
93.3%
5.0%
[+I,+P,-E,-M] (n = 32)
100.0%
0.0%
[+I,+P,-E,+M] (n = 7)
100.0%
0.0%
[+I,+P,+E,-S] (n = 12)
41.7%
58.3%
0.0%
[+I,+P,+E,+S] (n = 9)
100.0%
0.0%

Identifying bad predictions

Table: First Model: Actual category vs. predicted category in the model

		Actual category			
		both	comp	loc	Errors
Predicted category	both	0	0	0	0
	comp	10	70	1	11
	loc	7	0	32	7
	Errors	17	0	1	18

Table: Second Model: Actual category vs. predicted category in the model

		Actual category			
		both	comp	loc	Errors
Predicted category	both	14	5	0	5
	comp	3	65	1	4
	loc	0	0	32	0
	Errors	3	5	1	9

Summary

- The CIT model with extra predictors outperforms the first run of the model
- Most under-prediction of *both* COMP&LOC resolved
- Now some under-prediction of *only*-COMP

Discussion

Table: Details of remaining errors in the second model

noun	diagnostic	ev	phys	signifier	missive	cit prediction
briefing	loc	1	0	0	0	comp
guidance	both	1	0	0	0	comp
story	both	1	0	0	0	comp
case	both	0	0	0	0	comp
agreement	comp	1	1	0	0	both
declaration	comp	1	1	0	1	both
notice	comp	1	1	0	1	both
order	comp	1	1	0	1	both
proposal	comp	1	1	0	1	both

Blame the annotation?

- Most nouns in the upper section of the Table may in fact be used with a physical entity denoting sense, even if dictionaries do not list this
  - E.g., *story* can be used an an abbreviation of *story book*
- Nouns in the lower section of the Table may in fact be found in LOC constructions in a larger corpus

Summary: Empirical Landscape

Original Hypothesis

- Polysemy explains which nouns can be used with NNCs
  - Can be [+I, +P, +E] (*statement*), or [+I, +E] (*assertion*), or [+I, -P, -E] (*fact*)
  - Cannot be [+I, +P] (*book*)
- Found evidence that this hypothesis covers a large amount of the data ($\approx 85\%$)

Extra parameters needed

- Missive nouns (*letter*, *message*)
  - Nouns denoting written artefacts that are sent as missives
- Signifier nouns (*evidence*, *signal*)
  - Nouns denoting written artefacts that are sent as missives

Analysis overview

- Background: Semantics of Noun Related Complementiser clauses NCCs
- Background: Semantics of polysemous common nouns
- Proposal: how to unite the two

Background: The semantics of NCCs

A property of content bearing individuals (e.g., Kratzer 2006; Moulton 2015; Elliott 2020)

- Assuming a vanilla semantics for now, suppose that content bearing entities are a sub-sort of type e, signified e_c. Then:

(31) a. $[\![\text{claim}]\!] = \lambda w.\lambda x_{:e_c}.CLAIM(w)(x)$
b. $[\![\text{Alex lied}]\!] = \lambda w'.LIE(w')(alex)$
c. $[\![\text{that Alex lied}]\!] = \lambda w.\lambda x.CONT(w)(x) = [\![\text{Alex lied}]\!]$

- and via (intensionalised) predicate modification:

(32) $[\![\text{claim that Alex lied}]\!]$
$= \lambda w.\lambda x.CLAIM(w)(x) \wedge CONT(w)(x) = [\![\text{Alex lied}]\!]$

Background: Polysemous nouns and the challenge for semantics

On different occasions, a single use of a polysemous noun, can be used to denote entities that are of one, another or multiple semantic types:

(33) The library has 101,293 books.

a. The library has 101,293 physical tomes (inf duplicates possible)
b. The library has 101,293 informational distinct books (phys duplicates possible)
c. The library has 101,293 informationally and physically distinct books

Simple type theory:

(34) For a set of basic types **B** (e.g., $\{e, t, s, v\}$, the set of types for the interpretations of natural language expressions, **T** is defined inductively such that:

a. If $\sigma \in \mathbf{B}$, then $\sigma \in \mathbf{T}$
b. For all $\sigma, \tau \in \mathbf{T}$, $\langle \sigma, \tau \rangle \in \mathbf{T}$

Problem:

- no function expressible within (34) that can capture the truth conditions in (33a-c)
  - *book* can denote entities of type σ, τ, or of both simultaneously

Background: Proposals for modelling nominal polysemy

No consensus, but some of the main proposals include:

(35) Enriching the type theory, minimally by allowing for additional ways to combine types. This includes:
   a. Product types (Pustejovsky, 1995)
   b. Dot types, possibly as part of a richer Type theory (Chatzikyriakidis and Luo, 2015; Asher and Pustejovsky, 2006; Asher, 2011)
   c. Record types within a richly typed situation theoretic semantics (Cooper, 2007, 2011; Sutton, 2022, 2024)

(36) Mereology, polysemous expressions denote sums of entities, of putatively different types or sorts (Gotham, 2014, 2017)

(37) Property inheritance. Claim that any use of a polysemous noun denotes entities of one type, and derive readings via metaphysical property inheritance relations (Liebesman and Magidor, 2017, 2019)

I'll set (37 aside)

A generalised semantics for polysemous nouns

Noun lexical entry schema

(38) $\lambda \mathbf{x}.\psi$

Where **x** stands for something more complex either in terms of makeup or has a complex type. E.g.,

- $\mathbf{x} : \tau_1 \bullet \tau_2$ (An entity of a dot type)
- $\mathbf{x} : [x_1 : \tau_1; \ldots \; x_n : \tau_n]$ (A record of some type)
- $\mathbf{x} = a \sqcup b$ (A mereological sum of entities, possibly of different sorts)

Access to structure

- For each approach to polysemy, there is some means of accessing subparts of **x**, 'aspects' of **x**, or fields in **x** (e.g., projection functions, paths, mereological operations)
- I will also assume some such mapping

For types:

- p - physical entity
- i - informational entity
- v - eventuality

We have schematic lexical entries for different sorts of polysemous nouns

(39) $[\![\text{fact}]\!] = \lambda \mathbf{x}_{\mathbf{x} \mapsto [\phi : i]}.FACT(\mathbf{x})$

(40) $[\![\text{book}]\!] = \lambda \mathbf{x}_{\mathbf{x} \mapsto [x : p;\ \phi : i]}.BOOK(\mathbf{x})$

(41) $[\![\text{assertion}]\!] = \lambda \mathbf{x}_{\mathbf{x} \mapsto [e : v;\ \phi : i]}.ASSER(\mathbf{x})$

Where:

- *fact* is not polysemous (is purely informational)
- *book* is informational+physical entity polysemous
- *assertion* is informational+eventuality polysemous

For three-ways polysemous nouns e.g., *statement*

- There are restrictions on access to some combinations of senses (Sutton, 2022):

(42) a. ?The statement in the envelope lasted half an hour.
b. The statement in the envelope is inaccurate.
c. The inaccurate statement lasted half an hour.

Following Sutton 2022, we can capture this in terms of a join type:

(43) $[\![\text{statement}]\!] = \lambda \mathbf{x}_{\mathbf{x} \mapsto [x:p \sqcup v;\ \phi:i]}.STATEMENT(\mathbf{x})$

Capturing the original hypothesis

Recall the original motivating data:

- *fact/assertion/statement that* ϕ
- *??book that* ϕ

Proposal: add restrictions onto the *CONT* function assumed in the NCC literature

(44) $CONT_{NNC}$ is a function of type $\langle \sigma, i \rangle$ such that:

a. If $a : i$, $CONT_{NNC}(a) = a$ (the identity function)
b. If $a : v$, $CONT_{NNC}(a) = Theme(a)$ (the Theme function)
c. If $a : p$, $CONT_{NNC}(a) = \perp$ (undefined for physical entities)

Proposal: analysis of complementiser *that*

(45) $CONT_{NNC}$ is a function of type $\langle \sigma, \phi \rangle$ such that:

a. If $a : i$, $CONT_{NNC}(a) = a$ (the identity function)
b. If $a : v$, $CONT_{NNC}(a) = Theme(a)$ (the Theme function)
c. If $a : p$, $CONT_{NNC}(a) = \perp$ (undefined for physical entities)

Accessing parts/aspects/fields:

- If $\mathbf{x} \mapsto [x_i : \sigma_i, ..., x_n : \sigma_n]$, then:
  - $x_i^{\mathbf{x}} := x_i$
  - $\{x : x \in \mathbf{x}\} = \{x_i, ...x_n\}$

A semantics for the nominal *that* complementiser.

- $[\![\text{that}_{comp}]\!]$ universally quantifies over parts/aspects/fields and applies $CONT_{NNC}$ to each:

(45) $[\![\text{that}_{comp}]\!] = \lambda\phi_{:i}\lambda\mathbf{x}.\forall x \in \mathbf{x}.CONT_{NNC}(x) = \phi$

Example: *fact*

(39) $[\![\text{fact}]\!] = \lambda \mathbf{x}_{\mathbf{x} \mapsto [\phi:i]}.FACT(\mathbf{x})$

(46) $[\![\text{that}_{comp}\ \text{Alex lied}]\!]$
$= \lambda \mathbf{x}.\forall x \in \mathbf{x}.CONT_{NNC}(x) = [\![\text{Alex lied}]\!]$

Via predicate modification

(47) $[\![\text{fact that Alex lied}]\!]$
$= \lambda \mathbf{x}_{\mathbf{x} \mapsto [\phi:i]}.FACT(\mathbf{x}) \wedge \forall x \in \mathbf{x}.CONT_{NNC}(x) = [\![\text{Alex lied}]\!]$
$= \lambda \mathbf{x}_{\mathbf{x} \mapsto [\phi:i]}.FACT(\mathbf{x}) \wedge \phi^{\mathbf{x}} = [\![\text{Alex lied}]\!]$

A property of informational entities that are facts and identical to $[\![\text{Alex lied}]\!]$

Example: *assertion*

(41) $[\![\text{assertion}]\!] = \lambda \mathbf{x}_{\mathbf{x} \mapsto [e:v,\ \phi:i]}.ASSER(\mathbf{x})$

(46) $[\![\text{that}_{comp}\ \text{Alex lied}]\!]$
$= \lambda \mathbf{x}.\forall x \in \mathbf{x}.CONT_{NNC}(x) = [\![\text{Alex lied}]\!]$

Via predicate modification

(48) $[\![\text{assertion that Alex lied}]\!]$
$= \lambda \mathbf{x}_{\mathbf{x} \mapsto [e:v,\ \phi:i]}.ASSER(\mathbf{x}) \wedge \forall x \in \mathbf{x}.CONT_{NNC}(x) = [\![\text{Alex lied}]\!]$
$= \lambda \mathbf{x}_{\mathbf{x} \mapsto [e:v,\ \phi:i]}.ASSER(\mathbf{x}) \wedge \phi^{\mathbf{x}} = [\![\text{Alex lied}]\!] \wedge Theme(e^{\mathbf{x}}) = [\![\text{Alex lied}]\!]$

A property of informational+eventuality entities/situations etc. that are assertions, the contents of which is identical to $[\![\text{Alex lied}]\!]$

Example: *book*

(40) $[\![\text{book}]\!] = \lambda \mathbf{x}_{\mathbf{x} \mapsto [x:p,\ \phi:i]}.BOOK(\mathbf{x})$

(46) $[\![\text{that}_{comp}\ \text{Alex lied}]\!]$
$= \lambda \mathbf{x}.\forall x \in \mathbf{x}.CONT_{NNC}(x) = [\![\text{Alex lied}]\!]$

Via predicate modification

(49) $[\![\text{??book that Alex lied}]\!]$
$= \lambda \mathbf{x}_{\mathbf{x} \mapsto [x:p,\ \phi:i]}.BOOK(\mathbf{x}) \wedge \forall x \in \mathbf{x}.CONT_{NNC}(x) = [\![\text{Alex lied}]\!]$
$= \lambda \mathbf{x}_{\mathbf{x} \mapsto [x:p,\ \phi:i]}.BOOK(\mathbf{x}) \wedge \phi^{\mathbf{x}} = [\![\text{Alex lied}]\!] \wedge CONT_{NNC}(x^{\mathbf{x}}) =$
$[\![\text{Alex lied}]\!]$
$=$ Undefined, since $CONT_{NNC}(x^{\mathbf{x}}) = \perp$

Example: *statement*

(43) $[\![\text{statement}]\!] = \lambda \mathbf{x}_{\mathbf{x} \mapsto [x:p \sqcup v,\ \phi:i]}.STATEMENT(\mathbf{x})$

(46) $[\![\text{that}_{comp}\ \text{Alex lied}]\!]$
$= \lambda \mathbf{x}.\forall x \in \mathbf{x}.CONT_{NNC}(x) = [\![\text{Alex lied}]\!]$

The interpretation of *statement that Alex lied* depends on the type of x:

- Good if $x : v$
- Bad if $x : p$

Prediction:

- Degraded interpretation for *statement that* ϕ if *statement* is interpreted in context as a physical content-bearing artefact.

This prediction is arguably borne out, although the judgment is admittedly subtle [my judgement]:

(50) a. Billie's statement that Alex lied took but a few moments (but its effects on their relationship were permanent).
b. ?Billie's statement that Alex lied is in the filing cabinet.
c. Billie's statement in which he claims that Alex lied is in the filing cabinet.

I also consulted two native speakers of German who detected a similar contrast

- Consultant's comment: the (b) variant, but not the (a) variant makes it clear that the statement is written, not spoken

(51) a. ?Berthas Stellungnahme, dass Alex gelogen hat, ist in dem Aktenschrank.
b. Berthas Stellungnahme, in der sie behauptet, dass Alex gelogen hat, ist in dem Aktenschrank.

Extension to *missive* nouns (e.g., *letter, note*)

The problem for the original hypothesis:

- Missive nouns are $+$*physical-entity*, $-$*eventuality*
- Predicted to be infelicitous in comp
- But *I got a letter that I should pay a fine* is okay

Observation: variation in felicity depending on the content of the NNC [my judgements]:

(52) a. ?The broadcaster received a letter that a viewer complained about the programme.
b. The broadcaster received a letter that the presenter's comment was inappropriate.

Routinised coercion

(52) a. ?The broadcaster received a letter that a viewer complained about the programme.
b. The broadcaster received a letter that the presenter's comment was inappropriate.

Such subtle differences are reminiscent of cases of routinised coercion

- E.g., eventuality-selecting verbs such as *start* can be felicitously combined with DPs such as *the book* (53-a)
- but that access to this eventuality is restricted (53-b)

(53) a. Mary began the book. (Pustejovsky 1995)
b. ?War and Peace is a six-month book.
(Chatzikyriakidis et al. 2025, p.30)

Missive Ns in COMP result from coercion

(52) a. ?The broadcaster received a letter that a viewer complained about the programme.
b. The broadcaster received a letter that the presenter's comment was inappropriate.

- The TELIC-*quale* of a missive noun such as *letter* is plausibly an eventuality type such as *communication eventuality e, such that e communicates that p*.
- For (53-b) ?*a letter communicating that a viewer complained about the programme* is degraded

No coercion possible for *book*-like nouns

Coercion for a noun such as *book* is predicted to be inaccessible

- Eventuality types for qualia fields for *book*
  - e in which x reads b
  - e in which x writes b
- the eventuality types for reading and writing a book in Pustejovsky's analysis have their themes already saturated with *book*

Extension to *signifier* nouns (*proof*, *evidence*)

Safir (1985), Moulton (2009):

- "subject nominals" such as *proof* and *evidence* are relational ($\langle st, \langle e, t \rangle \rangle$)
- *The proof (that Bob was guilty)* the type *st* argument is filled by the CP (or contextually), and the type *e* argument can be a situation, physical entity, or proposition
- Moulton (2009) generally analyses CPs as type $\langle e, t \rangle$
  - Assumes 'late merge' to reconcile these types

With the semantics outlined here, CPs can be analysed in situ

Example: *proof*

(46) $[\![\text{that}_{comp}\ \text{Alex lied}]\!] = \lambda\mathbf{x}.\forall x \in \mathbf{x}.CONT_{NNC}(x) = [\![\text{Alex lied}]\!]$

(54) $PROOF : \langle i, \langle \sigma, t \rangle \rangle$

PROOF expresses a relation between e.g., [[Alex lied]] and an entity that is the proof of this

- So [[proof]] must map e.g., (46) to [[Alex lied]]

(55) $[\![\text{proof}]\!] = \lambda P_{\langle i,t \rangle}.\lambda\mathbf{x}.PROOF(\iota\phi_i[\exists \mathbf{y}_{\mathbf{y}\mapsto[\phi:i]}[P(\mathbf{y})]])(\mathbf{x})$

Because $CONT_{NNC}$ expresses the identity function when applied to informational entities:

(56) $\iota\phi_i[\exists \mathbf{y}_{\mathbf{y}\mapsto[\phi:i]}[[\![\text{that Alex lied}]\!](\mathbf{y})]]$
$= \iota\phi_i.\exists \mathbf{y}_{\mathbf{y}\mapsto[\phi:i]} CONT_{NNC}(\phi^{\mathbf{y}}) = [\![\text{Alex lied}]\!] \quad = \iota\phi_i.\phi = [\![\text{Alex lied}]\!]$

Deriving, in situ, the right truth conditions:

(57) $[\![\text{proof that Alex lied}]\!] = \lambda\mathbf{x}.PROOF(\iota\phi_i.\phi = [\![\text{Alex lied}]\!])(\mathbf{x})$

Summary: Formal Semantic Analysis

Original Hypothesis covers part of the data

- Nouns that can be used with NNCs in the COMP construction
  - Can be [+I, +P, +E] (*statement*), or [+I, +E] (*assertion*), or [+I, -P, -E] (*fact*)
  - Cannot be [+I, +P] (*book*)
- Implemented in a generalised semantics for polysemous nouns
  - Advantages of theory-specific implementations to be investigated

Summary: Formal Semantic Analysis

For (most of) the remaining data:

- *Missive* nouns
- *Signifier* nouns

Missive nouns with NNCs (e.g., *letter/message that*)

- Attributed to a form of routinised coercion
- $\approx$ *letter/message* COMMUNICATING *that*

Signifier nouns with NNCs (e.g., *evidence/sign that*)

- Relational semantics
- CP interpreted in situ, as opposed to a *late merge* analysis

Conclusions

Talking about informational content

- Complementiser strategy vs. locative strategy
- Are there two different types/sorts of content?
  - Metaphysical/Ontological distinction?
  - Or more a reflection of extending a container+contents metaphor?

Future work (currently underway):

- Replication corpus study in Finnish

Thanks and Acknowledgements

Thanks for your attention!

Acknowledgements:

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), project number 552248395. Peter Sutton additionally received funding from the University of Potsdam via a postdoctoral bridge stipend.

Thank you to Carla Bombi, Frank Grüneisen, and Alex Hogrebe for Spanish and German judgements.

References I

Asher, N. (2011). *Lexical Meaning in Context: A Web of Words*. Cambridge University Press.

Asher, N. and J. Pustejovsky (2006). A type composition logic for generative lexicon. *Journal of Cognitive Science*, 1–38. reprinted in *Advances in Generative Lexicon Theory*, Kluwer Academic Publishers, 2013. doi:https://doi.org/10.1007/978-94-007-5189-7_310.1007/978-94-007-5189-7_3.

Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation 43*(3), 209–226.

Chatzikyriakidis, S., R. Cooper, E. Gregoromichelaki, and P. R. Sutton (2025). *Theories of Types and the Structure of Meaning*. Cambridge University Press. Part of the Cambridge Elements in Semantics series.

Chatzikyriakidis, S. and Z. Luo (2015, July). Individuation criteria, dot-types and copredication: A view from modern type theories. In *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 2015)*, Chicago, USA, pp. 39–50. Association for Computational Linguistics.

Cooper, R. (2007). Copredication, dynamic generalized quantification and lexical innovation by coercion. In P. Bouillon, L. Danlos, and K. Kanzaki (Eds.), *Proceedings of GL 2007, Fourth International Workshop on Generative Approaches to the Lexicon*, pp. 143–184.

References II

Cooper, R. (2011). Copredication, quantification and frames. In S. Pogodalla and J.-P. Prost (Eds.), *Logical Aspects of Computational Linguistics. Number 6736 in Lecture Notes in Computer Science*, pp. 64–79. Springer.

Elliott, P. D. (2020). *Elements of Clausal Embedding*. Ph. D. thesis, University College London.

Ferraresi, A., E. Zanchetta, M. Baroni, and S. Bernardini (2008). Introducing, evaluating ukWaC, a very large web-derived corpus of English. In S. Evert, A. Kilgarriff, and S. Sharoff (Eds.), *Proceedings of the 4th Web as Corpus Workshop*, Marrakech, Morocco, pp. 47–54. European Language Resources Association.

Gotham, M. (2014). *Copredication, Quantification and Individuation*. Ph. D. thesis, University College London.

Gotham, M. (2017, 08). Composing Criteria of Individuation in Copredication. *Journal of Semantics 34*(2), 333–371.

Higgins, F. R. (1973). *The pseudocleft construction in English*. Ph. D. thesis, MIT.

Honnibal, M. and I. Montani (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Hothorn, T. and A. Zeileis (2015). partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research 16*, 3905–3909.

References III

Kratzer, A. (2006). Decomposing attitude verbs. https://semanticsarchive.net/Archive/DcwY2JkM/attitude-verbs2006.pdf.

Liebesman, D. and O. Magidor (2017). Copredication and property inheritance. *Philosophical Issues 27*, 131–166.

Liebesman, D. and O. Magidor (2019). Copredication, counting, and criteria of individuation: A response to gotham. *Journal of Semantics 36*, 549–561.

Moulton, K. (2009). *Natural Selection and the Syntax of Clausal Complementation*. Ph. D. thesis, University of Massachusetts at Amherst.

Moulton, K. (2015). Cps: Copies and compositionality. *Linguistic Inquiry 46*(2), 305–342.

Müller, K. (2023). On noun-related complementizer clauses. In K. Axel-Tober, L. Gunkel, J. Hartmann, and A. Holler (Eds.), *On the nominal nature of propositional arguments*, pp. 93–120. Buske.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.

Safir, K. (1985). *Syntactic Chains*. Cambridge University Press.

Sutton, P. R. (2022). Restrictions on copredication: a situation theoretic approach. *Semantics and Linguistic Theory (SALT) 32*, 335–355.

Sutton, P. R. (2024). Individuation criteria and copredication: modification in context. *Proceedings of Sinn und Bedeutung 28*, 876–894.

CQL request: the COMP diagnostic

(58) `year|week|day|today|yesterday|tomorrow|century|time|thing|something|everything|anything|everyone|everything|anyone|someone|nothing|friend`

(59) `<s/> containing [tag="NN|NNS"&lemma!="(58)"] [word="that"][tag="PP.*|NP.*|DT"]`[1]

Any sentence containing *N that DP* (excluding the above mentioned cases).

CQL request: The LOC diagnostic

(60) `<s/> containing`
`[lemma="information"][tag="IN"&lemma="in"]`
`[tag="DT"&lemma="this|that"][tag="JJ"]*`
`[tag="NN"]*[tag="NN|NNS"]`
Any sentence containing *information in this/that* followed by an NP, with optional adjective and that could be a compound.