Two levels of lexical ambiguity and a unified categorical approach

Chenchen (Julio) Song[1], cjs021@zju.edu.cn

[1]Zhejiang University, Hangzhou, China

11th International Conference on Meaning and Knowledge Representation
July 3–5, 2024, Almería, Spain

Overview

Natural Language Understanding and Linguistics

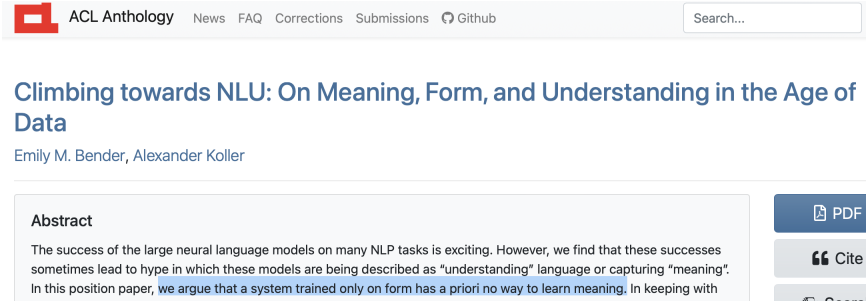
11TH INTERNATIONAL CONFERENCE ON MEANING AND KNOWLEDGE REPRESENTATION

Natural language understanding systems require a knowledge base provided with formal representations reflecting the structure of human beings' cognitive system. Although surface semantics can be sufficient in some other systems, the construction of a robust knowledge base guarantees its use in most natural language processing applications, thus consolidating the concept of resource reuse. This conference deals

[S]tatistical systems can accomplish natural language processing to a considerable degree, but they cannot achieve natural language understanding, which necessarily involves meaning, something which purely statistical approaches cannot capture. Something more is needed. To start down the path from NLP to NLU we have to go back to linguistics. *(Van Valin, 2016, p. 2)*

Natural Language Understanding and Linguistics

Human-analogous natural language understanding (NLU) is a grand challenge of artificial intelligence, which involves mastery of the structure and use of language and the ability to ground it in the world. While large neural LMs may well end up being important components of an eventual full-scale solution to human-analogous NLU, they are not nearly-there solutions to this grand challenge. *(Bender & Koller, 2020, p. 5185)*



So, to the extent that human-analogous NLU is a desirable goal, insights from (theoretical) linguistics are still useful.

The question is how to feed research results from linguistics into NLU/NLP, given the increasingly huge gap between the two fields.

This talk

I focus on the **formal representation** of language and use **lexical ambiguity** as a case to demonstrate how insights from theoretical linguistics can help us achieve more rigorous representation.

The formal tool I adopt is **Category Theory**, which is already in use in NLP research. Category Theory provides a nice bridge between lexical and compositional semantics.

Plan:

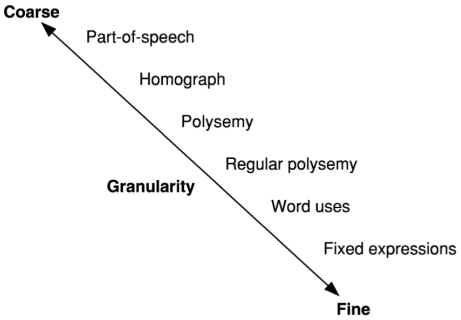
- Review the established types of lexical ambiguity in NLP.
- Distinguish two broad levels of lexical ambiguity in Linguistics.
- Formally represent the two levels and aim for a tentative unification.

Overview

Lexical ambiguity and disambiguation

Edmonds (2006):

- Lexical ambiguity is a fundamental defining characteristic of human language.
- Lexical disambiguation or **word sense disambiguation** (**WSD**) is one of the oldest problems in NLP.

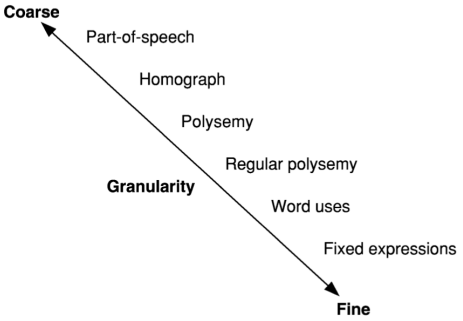


(Edmonds's spectrum of lexical ambiguity)

Lexical ambiguity and disambiguation

Part-of-speech ambiguity: e.g., *sharp*

- adj. having a thin edge
- n. a musical notation
- v. to raise in pitch
- adv. exactly



(Edmonds's spectrum of lexical ambiguity)

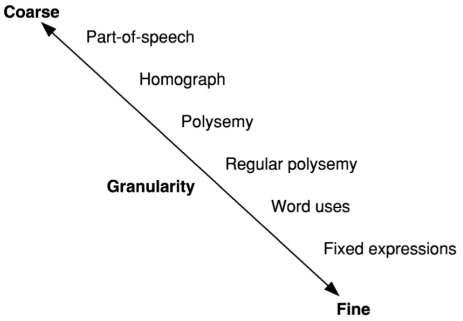
Edmonds's "word uses" and "fixed expressions" are not clearly defined, so I skip them.

Lexical ambiguity and disambiguation

Homographic ambiguity: e.g., *bow*

- /baʊ/ n. the front part of a boat or ship
- /boʊ/ n. an arrow-shooting weapon

(Same pronunciation ⇒ homonymy; e.g., *bank*)



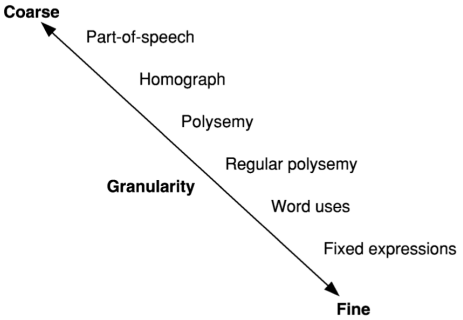
(Edmonds's spectrum of lexical ambiguity)

Edmonds's "word uses" and "fixed expressions" are not clearly defined, so I skip them.

Lexical ambiguity and disambiguation

Polysemy: e.g., *bank*

- n. the company or institution
- n. the building itself
- n. a money box (*piggy bank*)
- n. a place where a supply of something is held (*blood bank*)
- …



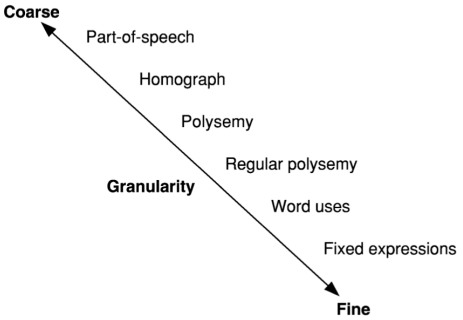
(Edmonds's spectrum of lexical ambiguity)

Edmonds's "word uses" and "fixed expressions" are not clearly defined, so I skip them.

Lexical ambiguity and disambiguation

Regular polysemy: e.g.,

- physical object vs. content (*book*, *CD*)
- institution vs. building (*bank*, *school*)



(Edmonds's spectrum of lexical ambiguity)

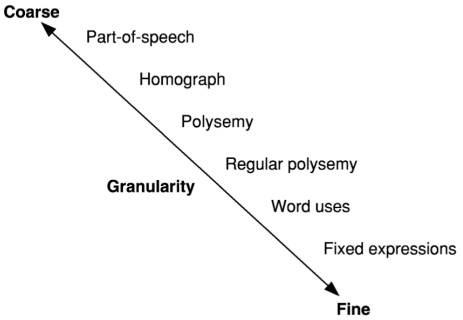
Edmonds's "word uses" and "fixed expressions" are not clearly defined, so I skip them.

Lexical ambiguity and disambiguation

POS and homographic ambiguity are considered solved problems in NLP, by

- POS-tagging
- clear contextual clues

The real challenge is **polysemy**, where contextual clues are less clear (Edmonds, 2006). (e.g., *book* in *I'm going to buy John a book* refers to both the physical object and the content)



(Edmonds's spectrum of lexical ambiguity)

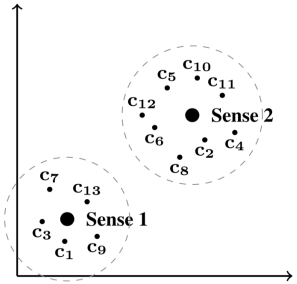
Edmonds's "word uses" and "fixed expressions" are not clearly defined, so I skip them.

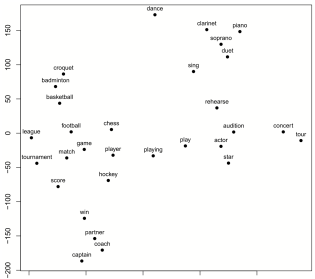
Lexical ambiguity and disambiguation

Polysemy is also somewhat solved in NLP, by **word sense induction** (**WSI**) techniques that represent word senses as

- clusters of contexts, or
- clusters of neighbors

Homograph/homonymy and polysemy are treated alike, while POS is considered a separate task.





(Lenci & Sahlgren, 2023, pp. 223–227)

Summary: lexical ambiguity in NLP

The NLP perspective is **task-oriented**:

- POS-tagging task, WSD task (homograph/homonymy, polysemy)

While this gets the job done, it does not align well with human intuition.

1. Homonymy/polysemy may also involve POS-level ambiguity.
   - Homonymy + POS: *stalk* 'n. part of a plant / v. follow a person'
   - Polysemy + POS: *ship* 'n. a large boat / v. send to a customer'
2. Homograph/homonymy and polysemy do have qualitative differences.
   - Homograph/homonymy usually arises by historical accident.
   - Polysemy usually arises by meaning extension or modification.

Overall, the NLP perspective on lexical ambiguity is convenient but impressionistic. It is not quite human-analogous.

Overview

Form-over-meaning vs. meaning-over-form

The NLP perspective on lexical ambiguity reflects a form-over-meaning mindset. Given a word form like *bank*, how to pin down its meaning?

1. What is its POS?
2. Is it homographic/homonymous? If so, which basic meaning is relevant?
3. Which specific sense of the basic meaning is involved?

This mindset is tied to the task of **parsing**. However, from the perspective of linguistics, human language works in more of a meaning-over-form fashion.

> *[L]anguage … is fundamentally a system of meaning. Aristotle's classic dictum that language is sound with meaning should be reversed. Language is meaning with sound (or some other externalization, or none).* *(Berwick & Chomsky, 2016, p. 101)*

Lexical ambiguity: a meaning-first view

To reach a more human-analogous treatment of lexical ambiguity:

1. We should treat homographic/homonymous words as separate lexical entries.
2. We should treat POS ambiguity as an integral part of polysemy.

Both principles should be reflected in our organization of meaning.

Example: *bank*

- $bank_1$ n. a mound, pile, or ridge raised above the surrounding level; the rising ground bordering a lake, river, or sea; …
  vt. to raise a bank about; to heap or pile in a bank; …
  vi. to rise in or form a bank; …
- $bank_2$ n. an establishment for the custody, loan, exchange, or issue of money; …
  vt. to deposit or store in a bank; …
  vi. to manage a bank; …

This is just the dictionary organization of word senses!

Toward a formal representation

- *bank*$_1$ n. a mound, pile, or ridge raised above the surrounding level; the rising ground bordering a lake, river, or sea; …
  vt. to raise a bank about; to heap or pile in a bank; …
  vi. to rise in or form a bank; …
- *bank*$_2$ n. an establishment for the custody, loan, exchange, or issue of money; …
  vt. to deposit or store in a bank; …
  vi. to manage a bank; …

Rationale behind dictionary organization:

- Word entries are built on separate **etymological roots**.
- Word senses sharing the same root are organized into **syntactic categories**.

The relationship between syntactic categories and word senses is mediated by the overarching roots: root (abstract) + syntactic category (abstract) = word sense (concrete)

Root syntax

The rationale we have seen is a standard part of current theoretical linguistics. It is a basic idea in the branch of generative syntax known as **root syntax**.

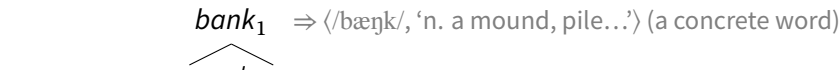
- Halle & Marantz (1993 et seq.): Distributed Morphology (DM)
- Borer (2005, 2013): Exoskeletal Syntax (XS)
- Chomsky (2019):

  If you accept … the Hagit Borer–Alec Marantz theory of root categorization, which I think is pretty strongly motivated, the roots in the lexicon are independent of category.

A theory-neutral definition of "root"

A root is a purely lexical unit in formal representation that is void of categorial information. Its syntactic category (and categorized meaning) is represented combinatorially.

Example (in DM style):



Root syntax

The “super-concept” (Fellbaum 2006) status of the root is most evident in Semitic languages.

Root-and-pattern morphology in Hebrew (Arad, 2005, p. 16)

√š-m-n	‘about some fatty substance’
šamen	‘adj. fat’
šuman	‘n. fat’
šaman	‘v. grow fat’
hišmin	‘v. fatten’
…	

√x-š-b	‘about some mental activity’
xašav	‘v. think’
xišev	‘v. calculate’
maxšava	‘n. thought’
maxšev	‘n. computer’
…	

[I]n an Arabic-English bilingual wordnet, the derivational root and form of each content word should be stored, since this way of semantically linking words is a basic expectation of a literate Arabic speaker. *(Black & ElKateb, 2004, p. 69)*

Root syntax

Root syntax is a fruitful research area in current theoretical linguistics. NB it is **not** a mere restatement of POS-tagging, because categorizers are not necessarily run-of-the-mill POS tags.

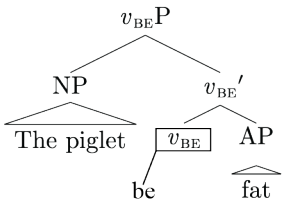
1. The categorizer may encode more subtle, language-specific information.
   - e.g., the “verbalizer” can encode event types, the “nominalizer” can encode genders
2. The categorizer may be a complex functional structure.
   - e.g., some verbs have a multi-layer structure

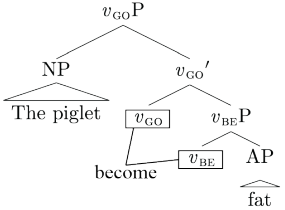
Root syntax

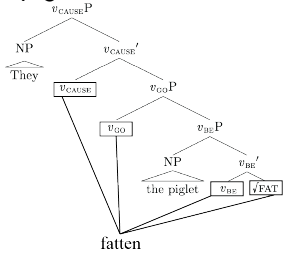
Root syntax is a fruitful research area in current theoretical linguistics. NB it is **not** a mere restatement of POS-tagging, because categorizers are not necessarily run-of-the-mill POS tags.

1. The categorizer may encode more subtle, language-specific information.
   - e.g., the "verbalizer" can encode event types, the "nominalizer" can encode genders
2. The categorizer may be a complex functional structure.
   - e.g., some verbs have a multi-layer structure

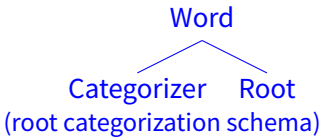
Examples: *The piglet is fat.* / *The piglet became fat.* / *They fattened the piglet.*







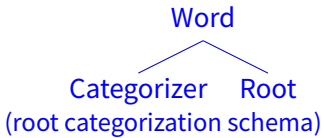
Back to lexical ambiguity



Two types of lexical ambiguity at the root categorization level:

- Homograph/homonymy: different roots
- POS-based polysemy: same root, different categorizers

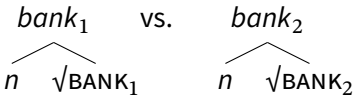
Back to lexical ambiguity

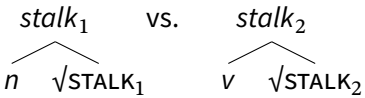


Two types of lexical ambiguity at the root categorization level:

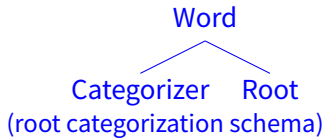
- Homograph/homonymy: different roots
- POS-based polysemy: same root, different categorizers

Example: homograph/homonymy





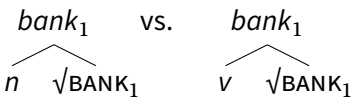
Back to lexical ambiguity



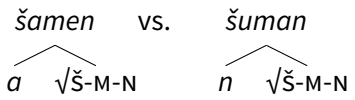
Two types of lexical ambiguity at the root categorization level:

- Homograph/homonymy: different roots
- POS-based polysemy: same root, different categorizers

Example: POS-based polysemy



(English)

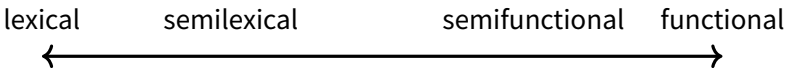


(Hebrew)

Generalized root syntax

Root syntax is not limited to content words but has been extended to certain function words too—primarily to **semilexical** or **semifunctional** (aka **semigrammatical**) items (Acedo-Matellán & Real-Puigdollers 2019; Song, 2019; Cavirani-Pots, 2020).

Semilexical or semifunctional items are linguistic elements with both lexical content and grammatical function. They are at an intermediate stage of grammaticalization.



(Continuum of lexicality in linguistic elements; Song, 2022)

I use “semilexicality” as a cover term for expository convenience.

Examples of semilexicality

Semilexicality is prevalent in analytic languages, where it often goes hand in hand with **multifunctionality**—the phenomenon where a word form has multiple grammatical functions.

(1) Classifiers

yī ***wèi*** / ***míng*** / ***gè*** *lǎoshī* [Mandarin Chinese]
one $\text{CL}_{\text{respectful}}$ $\text{CL}_{\text{professional}}$ $\text{CL}_{\text{neutral}}$ teacher
'a teacher'

(2) Conjunctions

a. *hālì bōtè* ***yǔ*** / ?***hé*** / ?***gēn*** *mófǎ shí* [Mandarin Chinese]
Harry Potter $\text{and}_{\text{literary}}$ $\text{and}_{\text{neutral}}$ $_{\text{colloquial}}$ magic stone
'*Harry Potter and the Philosopher's Stone*'

b. *xiǎomíng* / ?***yǔ*** / ***hé*** / ***gēn*** *xiǎohóng dōu zài shātān-shàng wán*
Xiaoming $\text{and}_{\text{literary}}$ $\text{and}_{\text{neutral}}$ $_{\text{colloquial}}$ Xiaohong both be.in beach-on play
'Both Xiaoming and Xiaohong are playing on the beach.'

Examples of semilexicality

Vietnamese negators (all meaning ‘not’)

không	default	*nào*	colloquial but elevated
chẳng	emphatic	*đếch*	mildly vulgar
chả	emphatic, informal	*đéo*	very vulgar
đâu	emphatic, colloquial	*cóc*	very informal

(3) *Em/Tao* ***không*** / ***đéo*** *cần* *anh/mày* *giúp.* [Vietnamese]
1SG.N/V $\text{NEG}_{\text{neutral}}$ $\text{NEG}_{\text{vulgar}}$ need 2SG.N/V help
‘I do not need your help.’ (Li Nguyen, p.c.)

Examples of semilexicality

Semilexicality is observed in synthetic languages too, though to a lesser extent.

(4) Alternative voice auxiliaries

La pasta ***va*** / ***viene*** *mangiata subito.* [Italian]
the pasta $\text{PASS}_{\text{obligatory}}$ $\text{PASS}_{\text{regular}}$ eaten immediately
'Pasta must be / is eaten immediately.' (Cardinaletti & Giusti, 2001, p. 392)

(5) Alternative aspect auxiliaries

a. *Ik heb de hele dag* ***zitten*** *te lezen.* [Dutch]
I have the entire day sit_{PROG} to read
'I have been reading the entire day.' (Cavirani-Pots, 2020, p. 1)

b. *Ek het gister baie (ge-)***loop** *(en) praat.* [Afrikaans]
I have yesterday a.lot $\text{walk}_{\text{PROG}}$ and talk
'I have been (walking and) talking a lot yesterday.' (ibid., p. 344)

Examples of semilexicality

Somewhat surprisingly, semilexicality is also prevalent in polysynthetic languages, though in a different guise (as affixes) and under a different name (called “lexical/field affixes”).

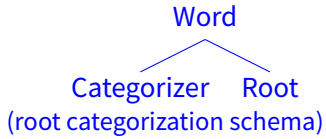
(6) Classifiers
ɬíx^{w}-***əqən*** *liśék* [Halkomelem]
three-CL$_{container}$ sack
‘three sacks’ (Gerdts & Hinkson, 1996, p. 10)

(7) a. *dikwh-**okwɬ*** *boɬak* [Yurok]
three-CL$_{salmon}$ salmon
‘three salmon’

b. *nahks-**oh*** *ha’aag*
three-CL$_{round}$ rock
‘three rocks’ (Conathan, 2004, pp. 26–27)

See Song (2021a) for more crosslinguistic data.

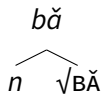
Back to lexical ambiguity



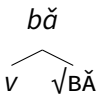
Semilexicality is a special type of POS-based polysemy:

- Categorizer: a functional/grammatical category
- Root: a normal root (chosen for various reasons)

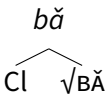
Example: Mandarin Chinese



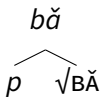
(n. handle)



(v. hold)



(cl. for holdable objects)



(p. for affected direct objects)

Summary: categorization-level ambiguity

In sum, two types of lexical ambiguity may arise at the root categorization level:

- Homograph/homonymy: different roots
- POS-based polysemy: same root, different categorizers
  - Content words (a solved problem in NLP)
  - Semigrammatical words (not yet dealt with in NLP)

Theoretical linguistics provides a way to systematically represent these ambiguity types (via the theory of generalized root syntax).

Lexical ambiguity beyond the categorization level

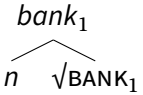
The only remaining type of lexical ambiguity we have not addressed yet is **word sense polysemy**; e.g., *bank*

- n. the company or institution
- n. the building itself
- n. a money box (*piggy bank*)
- n. a place where a supply of something is held (*blood bank*)
- …

We only consider ambiguity within the same POS here—that is, **post-categorization** ambiguity. The separation of POS-based ambiguity into two levels is a crucial feature of our theoretical linguistic perspective on lexical ambiguity, which distinguishes it from the NLP perspective.

The limit of root syntax

Root syntax is not designed to tackle issues at the post-categorization level. For example, any nominal sense of $bank_1$ is represented as



In root syntax, all categorized senses of a word are lumped together in an area in the Lexicon called the **Encyclopedia**. The internal structure of this area is left unstudied.

The categorical approach I will introduce offers a way to formally represent word sense ambiguity at the post-categorization level.

Overview

Category Theory

Category Theory is a branch of mathematics dedicated to the representation and reasoning of abstract, complex structures.

> *Category theory is […] unmatched in its ability to organize and layer abstractions, to find commonalities between structures of all sorts, and […] it has also been branching out into science, informatics, and industry.* *(Fong & Spivak, 2019)*

Category Theory has also been applied to the study of natural language, especially to NLP. The most representative application is that of the Lambek-Coecke school (Lambek, 1988; Coecke et al., 2010; et seq.).

Fong & Spivak (2019) is also an accessible modern textbook on the subject.

Categorical linguistics: the Lambek-Coecke school

The focus of this school of categorical linguistics, called "DisCoCat," is a principled integration of lexical semantics and compositional semantics in NLP. The main idea is to treat semantic interpretation as a **functor** from syntax to semantics (see Coecke et al., 2013 for an overview).

- Syntax: a free category $\mathcal{C}$ based on Lambek's (1999) grammatical type system
- Semantics: the category $\mathcal{FV}ect$ of finite-dimensional vector spaces

$$\mathcal{C} \xrightarrow{I} \mathcal{FV}ect$$

I maps grammatical types to vector spaces that represent meanings, but thanks to the shared structure of the two categories, meaning vectors now live in vector spaces of different "types" and can compose according to grammatical relations.

Lexical ambiguity in DisCoCat

Lexical ambiguity has not received a lot of attention in the DisCoCat framework. The few studies dedicated to it (Piedeleu, 2014; Kartsaklis, 2014; Piedeleu et al., 2015) are all inspired by quantum mechanics:

- ambiguous words $\Rightarrow$ mixed states
- nonambiguous words $\Rightarrow$ pure states (represented in a "mixed" way)

Technical details aside, this modeling of lexical ambiguity mainly focuses on polysemy, while homonymy is separately treated in a pre-compositional disambiguation step. Two problems:

- The classification of ambiguity types is a coarse one.
  - With just the coarsely defined polysemy and homonymy
- There is no built-in way to disambiguate word senses.
  - The mixed states are weighted summations of pure states $\sum_i p_i |s_i\rangle\langle s_i|$.

What this model does is mainly **give ambiguous words a place** in DisCoCat.

Alternative categorical approaches to natural language

DisCoCat is not the only application of Category Theory to linguistics.

- Asudeh & Giorgolo (2020): conventional implicature via monad
  - e.g., the negative speaker attitude in words like *Yank* and *cur*
  - The monad tool keeps "at-issue" and "side-issue" aspects of meaning separate.
- Song (2021b): an extension of the monadic approach to root syntax
  - The arbitrary meaning of words is essentially also a matter of conventionalization.
  - e.g., the meaning of *bank* has a non-arbitrary part (its POS) and an arbitrary part (its root)
  - The POS is an "at-issue" for compositional semantics, while the root content is a "side-issue."
- Asher (2011), Babonnaud (2019, 2021, 2022): "dot type" polysemy via topos
  - This corresponds to regular polysemy in the WSD task.
  - e.g., *book* has the type $\textsc{p}\cdot\textsc{i}$, with a physical and an informational aspect
  - A dot type can be categorically reduced to its individual aspects (i.e., be disambiguated).

Alternative categorical approaches to natural language

DisCoCat is not the only application of Category Theory to linguistics.

- Asudeh & Giorgolo (2020): conventional implicature via monad
- Song (2021b): an extension of the monadic approach to root syntax
- Asher (2011), Babonnaud (2019, 2021, 2022): “dot type” polysemy via topos

The categorical environment in Song (2021b) is a topos too, so these alternative approaches could potentially be combined. Such a unified approach would have two advantages:

- It would preserve the fine-grained classification of lexical ambiguity from the theoretical linguistic perspective.
- It would support meaning composition along the way (though not using distributional semantics, which is still a unique feature of DisCoCat).

Overall, this potential unified approach could provide a better channel to connect theoretical linguistics and NLU (e.g., the topos environment naturally supports ontology building).

Conventional implicature via monad

Monad is a general concept from Category Theory, but the particular type of monad Asudeh & Giorgolo (2020) use to model conventional implicature—the **writer** monad—is borrowed from functional programming (which in turn has a category-theoretic basis).

> *[T]he writer monad [is] used for logging or tracing the execution of functions. It's also an example of a more general mechanism for embedding [side] effects in pure computations.*
>
> *(Milewski, 2019, p. 49)*

type Writer a = (a, **String**)
(*Writer a* creates a log area for any type *a*)

(8) ⟦Yank⟧ = (American, {The speaker has a negative attitude.})

Knowledge-enhanced pre-trained language models (PLMs)

The two-track representation in monadic semantics is reminiscent of a common method in **knowledge-enhanced PLMs** (e.g., Zhang et al., 2020; Wang et al., 2021; Sun et al., 2020, 2021).

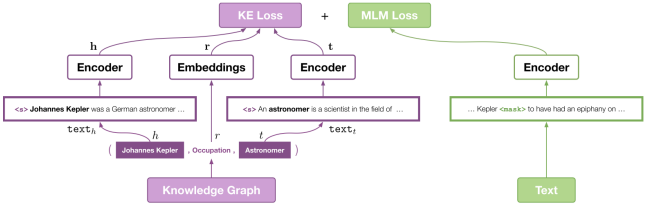


Figure 2: The KEPLER framework. We encode entity descriptions as entity embeddings and jointly train the knowledge embedding (KE) and masked language modeling (MLM) objectives on the same PLM.

$$\mathcal{L} = \mathcal{L}_{KE} + \mathcal{L}_{MLM}$$

(Wang et al., 2021)

Jointly optimizing the two objectives can implicitly integrate knowledge from external KGs into the text encoder, while preserving the strong abilities of PLMs for syntactic and semantic understanding. *(Wang et al., 2021, p. 179)*

Root categorization via monad

A monad is an **endofunctor** with two natural transformations. Abstracting away from technical details, the core idea is to represent meanings with non-pure-function content as a pair

(pure-function meaning, {non-pure-function content})

This representation is formally obtained via **three tools** provided by the monad:

- A "wrapper" lifting a pure-function meaning to the Writer type: $\eta(x) = (x, e) :\ a \rightarrow \text{Writer}\ a$
- A "writer" logging non-pure-function content into the Writer type: $\text{write}(s) = (1, s)$
- A "bind" operation enabling composition:
  (x, s_1) >>= $\lambda u.(f(u), s_2) = (f(x), s_2 +\!\!+ s_1) :$ Writer $a \rightarrow (a \rightarrow$ Writer $b) \rightarrow$ Writer b

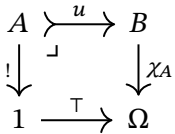
(9) a. $[\![[_{\text{N}}\ n\ \sqrt{\text{BANK}_1}\ ]]\!] = \text{write}((n, \sqrt{\text{BANK}_1}))$ >>= $\lambda y.\, \eta[\![n]\!] = ([\![n]\!], \{(n, \sqrt{\text{BANK}_1})\})$
(an entity that is idiosyncratically characterized by $\sqrt{\text{BANK}_1}$)

b. $[\![[_{\text{Cl}}\ \text{Cl}\ \sqrt{\text{BǍ}}\ ]]\!] = \text{write}((\text{Cl}, \sqrt{\text{BǍ}}))$ >>= $\lambda y.\, \eta[\![\text{Cl}]\!] = ([\![\text{Cl}]\!], \{(\text{Cl}, \sqrt{\text{BǍ}})\})$
(a classifier that is idiosyncratically characterized by $\sqrt{\text{BǍ}}$)

Subtyping in a topos

A topos is a special category that, in addition to the basic categorical setting, has some extra features, including products, "pullbacks," exponentials, and most importantly a **subobject classifier** (usually denoted by Ω).

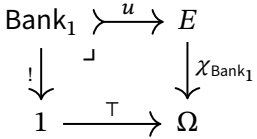
The above features together make the topos a richly structured, quasi-set-theoretic environment that is particularly suitable for knowledge and meaning representation. Specifically, the Ω tool can be used to represent **subtypes** and thereby to build **type ontologies**.



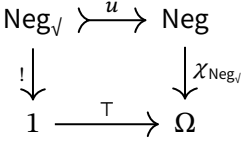
In this "pullback" square (which exists for all subtype relations),

- Ω represents the truth-value type,
- A represents any subtype of B,
- χ_A is the characteristic map of A.

Linguistic subtyping

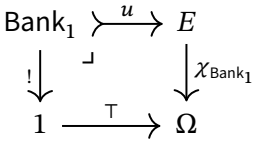


(content word)

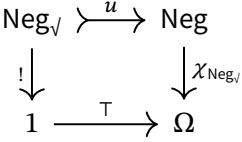


(semigrammatical word)

Linguistic subtyping



(content word)



(semigrammatical word)

Toposes already emerged in Asher's (2011) categorical model for TCL as suitable (and even necessary) for interpreting dot types, and Babonnaud (2019) further argues that toposes could be the best categorical models to interpret on a unified basis a large variety of semantic frameworks with subtyping. *(Babonnaud, 2021, p. 19)*

“Dot type” polysemy via topos

Asher (2011) uses “dot types” to represent words with **inherent polysemy** (aka regular polysemy in WSD). For instance, *book* has the type P·I, which has both a PHYSICAL and an INFORMATIONAL aspect. This idea can be traced back to Pustejovsky (1996).

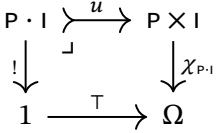
> *For each sense pair [like* PHYSICAL *and* INFORMATIONAL*], there is a relation which “connects” the senses in a well-defined way. … This relation must be seen as part of the definition of the semantics for the dot object … to be well-formed.* *(Pustejovsky, 1998, p. 335)*

“Dot type” polysemy via topos

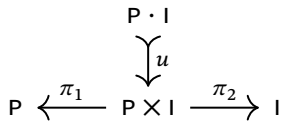
Asher (2011) uses “dot types” to represent words with **inherent polysemy** (aka regular polysemy in WSD). For instance, *book* has the type P·I, which has both a PHYSICAL and an INFORMATIONAL aspect. This idea can be traced back to Pustejovsky (1996).

> *For each sense pair [like* PHYSICAL *and* INFORMATIONAL*], there is a relation which “connects” the senses in a well-defined way. … This relation must be seen as part of the definition of the semantics for the dot object … to be well-formed.* *(Pustejovsky, 1998, p. 335)*

Babonnaud (2019, 2021, 2022) treats dot types as subtypes of product types. On this view, the individual aspects of dot types can simply be retrieved via **categorical projections**.



(P·I is a relation between P and I)



(π_1, π_2 are projections)

Asher (2011) holds a more complex view involving power objects.

Post-categorization ambiguity: where “dot types” meet root syntax

Now let’s combine the “dot type” view and the root syntax view. For a word with “inherent polysemy” like *book*, we can assign it the following syntactic and semantic representations:

(10) a. Syntax: $[_{\mathrm{N}}\ n\ \sqrt{\textsc{book}}\ ]$

b. Semantics: $(\llbracket n \rrbracket, \{(n, \sqrt{\textsc{book}})\})$ (an entity idiosyncratically characterized by $\sqrt{\textsc{book}}$)

Root syntax does not specify what exactly this “idiosyncratic characterization” is. In fact, when it is just a single word,

(11) $\mathrm{Type}(\llbracket n \rrbracket, \{(n, \sqrt{\textsc{book}})\}) = \mathrm{Type}(\llbracket n \rrbracket) \times \mathrm{Type}(\{(n, \sqrt{\textsc{book}})\}) \cong \mathrm{Type}(\llbracket n \rrbracket) \times 1 \cong \mathrm{Type}(\llbracket n \rrbracket)$

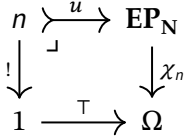
This demonstrates that as far as pure functional composition is concerned, the root-contributed idiosyncrasy does not matter.

Post-categorization ambiguity: where “dot types” meet root syntax

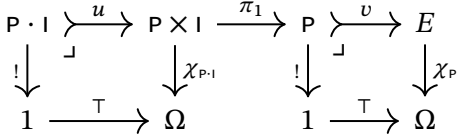
But that root-contributed idiosyncrasy is precisely what gives rise to **inherent polysemy**. Hence, the dot type corresponds to the root-tagging “log.” This brings us to the overall representation:

(12) $\mathrm{Type}((\llbracket n \rrbracket, \{(n, \sqrt{\textsc{book}})\})) = \mathrm{Type}(\llbracket n \rrbracket) \times \mathsf{P} \cdot \mathsf{I}$

The first component of this product is relevant to pure functional composition, while the second component is relevant to lexical semantic interpretation. NB both components may be part of some ontological structure (one grammatical and the other lexical):



(grammatical ontology)



(semantic ontology)

$\mathbf{EP_N}$ represents a certain well-defined set of grammatical types that *n* is part of.

Overview

Conclusion

In this study, I have

- **revisited** the established types of lexical ambiguity in NLP/NLU
  - POS ambiguity, homograph/homonymy, polysemy
- **reorganized** those types into two theoretical linguistic levels
  - root categorization level: homograph/homonymy, POS-based polysemy
  - post-categorization level: word sense polysemy (within the same POS)
- formally **represented** the two ambiguity levels in the language of Category Theory
  - respectively via monad and topos
  - with a tentative unification

Takeaway: Theoretical linguistics can still be useful for next-generation NLU. Category Theory may be a useful channel.

Future research: more details about the above unification

Thank you!

Selected references I


Asher, N.
Lexical meaning in context
CUP, 2011


Asudeh, A. & G. Giorgolo
Enriched meanings
OUP, 2020


Babonnaud, W.
A topos-based approach to building language ontologies
Formal Grammar: 24th International Conference, 18–34, 2019


Babonnaud, W.
On the dual interpretation of nouns as types and predicates in semantic type theories
Proceedings of the ESSLLI 2021, 15–24, 2021

Selected references II

Borer, H.
Structuring sense: Taking form (Vol. 3)
OUP, 2013

Coecke, B., E. Grefenstette & M. Sadrzadeh
Lambek vs. Lambek
Annals of Pure and Applied Logic 164, 1079–1100, 2013

Coecke, B., M. Sadrzadeh & S. Clark
Mathematical foundations for a compositional distributional model of meaning
Manuscript, 1–34, 2010

Edmonds, P.
Disambiguation, lexical
Elsevier Encyclopedia of Language & Linguistics, 2nd Ed. 607–623, 2006

Selected references III

Fellbaum, C.
WordNet(s)
Elsevier Encyclopedia of Language & Linguistics, 2nd Ed. 665–670, 2006

Fong, B. & D. Spivak
An invitation to Applied Category Theory
CUP, 2019

Halle, M. & A. Marantz
Distributed Morphology and the pieces of inflection
The View from Building 20, 111–176, MIT Press, 1993

Lambek, J.
Categorial and categorical grammars
Categorial grammars and natural language structures, 297–318
D. Reidel, 1988

Selected references IV

Lambek, J.
Type grammar revisited
Logical aspects of computational linguistics, 1–27, Springer, 1999

Piedeleu, R., D. Kartsaklis, B. Coecke & M. Sadrzadeh
Open system categorical quantum semantics in natural language processing
CALCO'15 Proceedings, 267–286, 2015

Pustejovsky, J.
The semantics of lexical underspecification
Folia Linguistica 32(3–4), 323–347, 1998

Song, C.
On the formal flexibility of syntactic categories
University of Cambridge dissertation, 2019

Selected references V

Song, C.
A typology of semilexicality and the locus of grammatical variation
Talk at ICFL9, 2021a

Song, C.
On the semantics of root syntax
LENLS18 Proceedings, 61–74, 2021b

Van Valin, R.
From NLP to NLU
Manuscript, 1–7, 2016

Wang, X., T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li & J. Tang
KEPLER: A unified model for knowledge embedding and pre-trained language representation
Transactions of the Association for Computational Linguistics 9, 176–194, 2021