Modification, Discourse Status, Argument Optionality, Precision Grammars & Corpus Data

Ling 567
February 6, 2018

Overview

- Software review
- Modification
- Discourse status
- Argument optionality
- Possessives
- Precision grammars & corpus data

Components

- HPSG: Theoretical foundations
- LKB
- Grammar (Matrix-provided, plus extensions)
- Emacs: editor, interaction with LKB
- [incr tsdb()]
- Version control: svn/git/...
- make_item

LKB

- tdl reader/compiler
- parser
- generator
- grammar exploration tools
  - parse chart
  - interactive unification
  - type and hierarchy exploration

Grammar

- A set of tdl files:
  - Grammar Matrix core
  - Additions from the customization system
  - Your additions
- Actually separated into:
  - Type definitions
  - Instances of grammar rules, lexical rules, lexical entries
  - Root symbols
  - Node label abbreviations
- Also includes: Lisp code for LKB interaction

[incr tsdb()]

- Pronounced “tee ess dee bee plus plus”
- Loading in test suites
- Running test suites (batch processing)
- Comparing multiple test suite runs:
  - Changes in which examples parse
  - Changes in number of analyses per item
  - Changes in representations per item
- Treebanking

Overview

- Software review
- Modification
- Discourse status
- Argument optionality
- Possessives
- Precision grammars & corpus data

Modification: Syntax

- Modifiers select the heads they modify through MOD
- Value of MOD is a list of *synsem*s
- Head-modifier rules are cross-classified according to order (head-adj, adj-head) and the non-scopal (“intersective”)/scopal distinction
- You likely already have head-modifier rules in your grammar (probably just instances in rules.tdl that inherit from types in matrix.tdl)

Semantics: Non-scopal modifiers

- Adjoined via a “head-compositional” phrase structure rule
- ARG1 is MOD’s INDEX (*individual*)
- LTOP = MOD’S LTOP (constraint on rule)

Semantics: Scopal modifiers

- Serve as semantic head daughters
  - What does this mean in tdl?
- Identify their own INDEX with their MOD's INDEX (why?)
- Take a handle-valued ARG1 (why?)
- Insert a qeq between their ARG1 and their MOD's LTOP (why?)

Scopal modifiers: Examples

- Kim did not read every book.
- Kim probably read every book.
- The most likely winner of every medal was disqualified.

Gate keeping

- The phrase structure rules for non-scopal and scopal modifiers need to be different
- Use subtypes of *local* to constrain which rule gets used
  - No other uses for subtypes of *local*
  - Modifiers constrain LOCAL inside their MOD value

Scopal mod phrase

```
scopal-mod-phrase := head-mod-phrase-simple &
  [ NON-HEAD-DTR.SYNSEM.LOCAL
                        [ CAT.HEAD.MOD < [ LOCAL scopal-mod ] >,
                          CONT.HOOK #hook ],
    C-CONT [ HOOK #hook,
             HCONS <! !> ] ].
```

Intersective mod phrase

```
isect-mod-phrase := head-mod-phrase-simple &
                    head-compositional &
  [ HEAD-DTR.SYNSEM.LOCAL.CONT [ HOOK.LTOP #hand ],
    NON-HEAD-DTR.SYNSEM.LOCAL
          [ CAT.HEAD.MOD < [ LOCAL intersective-mod ] >,
            CONT.HOOK.LTOP #hand ],
    C-CONT.HCONS <! !> ].
```

Modification: Your tasks

- (Instantiate head-adj-int and adj-head-int rules, as necessary)
- Create/edit lexical types for adverbs and adjectives (inheriting from matrix types)
- Constrain lexical types/lexical rules that are leading to overgeneration together with the head-adj / adj-head rules.

Overview

- Software review
- Modification
- Discourse status
- Argument optionality
- Possessives
- Precision grammars & corpus data

Discourse status: What's that?

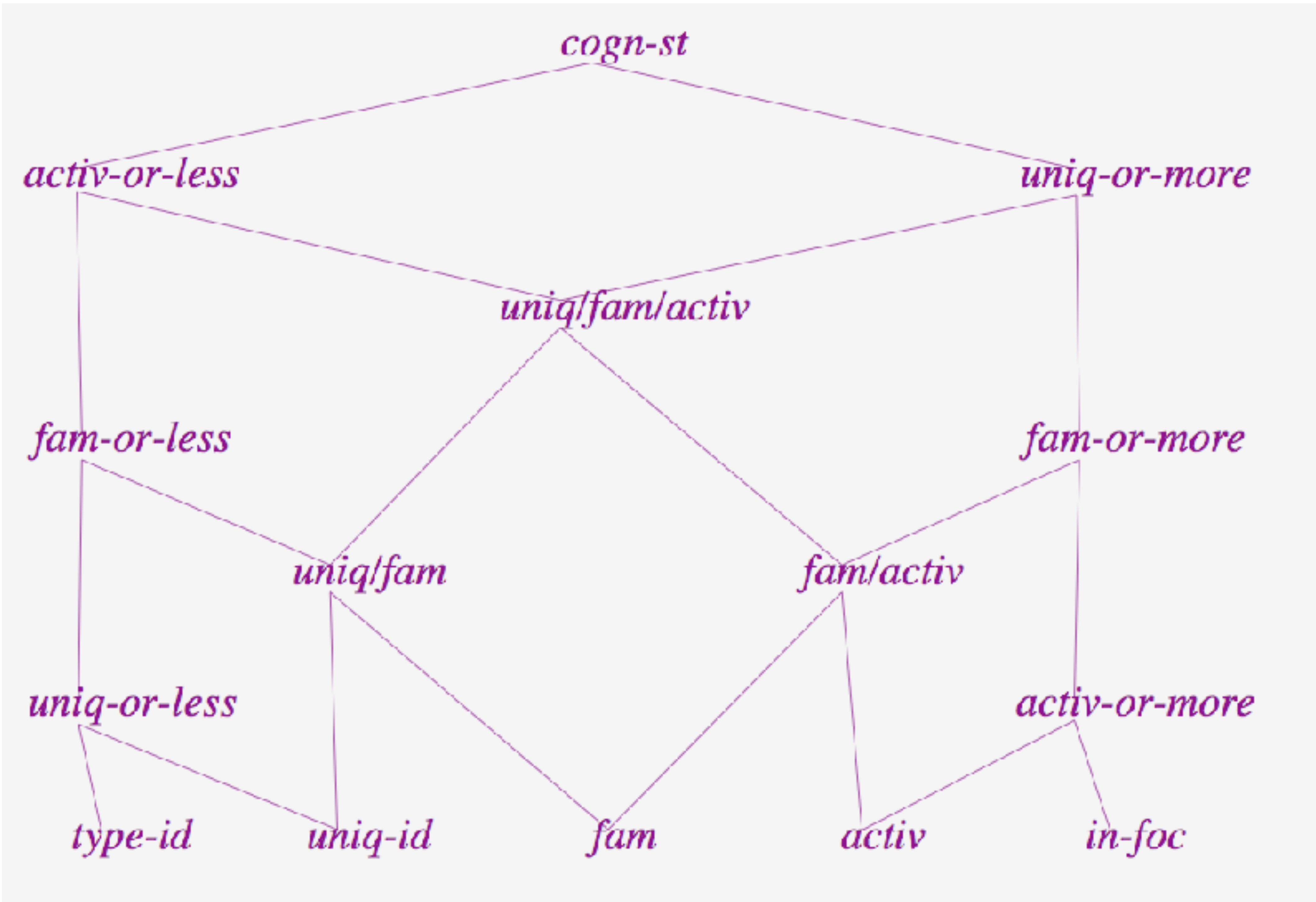
- A property of referents, describing their relationship to the common ground of a conversation
- Tends to be reflected syntactically in markers of "definiteness" as well as demonstratives and constraints on the availability of types of NPs in particular constructions.
- Closely related to (but distinct from) information structure
- The binary distinction "definite"/"indefinite" is not sufficient
- Furthermore, discourse status can be broken down into hearer-oriented "cognitive status" and speaker-oriented "specificity"

Givenness hierarchy
(Gundel et al 1993, Prince 1981)

Type id <	Referential <	Uniq. id. <	Familiar <	Activated <	In focus
a N	indefinite	*the N*	*that N*	*that, this*	*it*
	this N			*this N*	

NB: "In focus" != focus

Borthen & Haugereid’s proposal



Borthen & Haugereid’s proposal

$$
\begin{bmatrix}
\textsc{synsem.loc.cont.ref-prop} & \begin{bmatrix}
\textit{ref-prop} & \\
\textsc{index} & \begin{bmatrix}
\textit{ref} & \\
\textsc{per} & \textit{per} \\
\textsc{num} & \textit{num} \\
\textsc{gend} & \textit{gend}
\end{bmatrix} \\
\textsc{cogn-st} & \textit{cogn-st} \\
\textsc{speci} & \textit{bool} \\
\textsc{part} & \textit{bool} \\
\textsc{univ} & \textit{bool}
\end{bmatrix}
\end{bmatrix}
$$

Borthen & Haugereid’s proposal

- SPECI indicates specificity (speaker-oriented)
- Compatible with both “definite” and “indefinite” NPs:
    - *The fastest runner won*.
    - *The next customer will receive a reward*.
    - *I’m looking for a book*.
- Corresponds to overt syntactic phenomena in at least Norwegian (specificity adjectives) and Turkish (accusative case precludes specific interpretation)

Matrix-based proposal

$$\begin{bmatrix} \text{HOOK.INDEX} & \begin{bmatrix} \text{PNG} & \begin{bmatrix} \text{PER} & \textit{person} \\ \text{NUM} & \textit{number} \\ \text{GEND} & \textit{gender} \end{bmatrix} \\ \text{COG-ST} & \textit{cog-st} \\ \text{SPECI} & \textit{bool} \end{bmatrix} \end{bmatrix}$$

Your tasks (if applicable)

- Determine what overt marking of cognitive status and/or specificity occurs in your language
- Constrain the COG-ST value of pronouns, demonstratives, articles (if applicable), and any morphology that relates to cognitive status
- Constrain the COG-ST value of dropped arguments (see next section)

Overview

- Software review
- Modification
- Discourse status
- Argument optionality
- Possessives
- Precision grammars & corpus data

Optional arguments

- There are many cases in which an argument may be semantically present but syntactically absent
- Semantically, these cases can be categorized by how the missing argument is interpreted
- Syntactically, they can be categorized by how the missing argument is licensed

Semantic classification

- Indefinite null instantiation: *I ate*
  - The referent of the missing argument is indefinite, not (necessarily) recoverable from context
- Definite null instantiation: *I told you already*
  - The referent of the missing argument is definite, i.e., it is presumed to be recoverable from context
- Constructional null instantiation: *Eat!, I told Kim to eat., Stir until completely mixed.*
  - The referent of the missing argument is determined by the syntactic construction.
- Impersonal null instantiation:
  - The referent of the missing argument is “people in general”, or “one”

Syntactic classification

- Lexically licensed: The potential for an argument to be missing is determined by the lexical type of the selecting head
  - *eat* allows indefinite null instantiation of its object
  - *devour* does not
- Systematic: The potential for an argument to be missing is determined by its grammatical function, plus (possibly) inflection on the verb as well as other factors (PNG features on the argument, TAM features on the verb, ...)

Analysis in the Matrix

- Constructional null instantiation is handled by the analyses of imperatives, raising, etc.
- Distinction between definite and indefinite null instantiation handled by the COG-ST feature
  - Pronouns and arguments subject to DNI [COG-ST in-foc, SPECI +]
  - Arguments subject to INI are [COG-ST type-id, SPECI -]
- Posit opt-comp and opt-subj rules parallel to the bare NP rule
- Use a feature [OPT bool] to constrain possibility of argument drop/overt realization
- Use a feature [OPT-CS cog-st] to allow lexical items (and lexical rules!) to indicate COG-ST value of an argument in the event it is dropped

The features OPT and OPT-CS

- OPT and OPT-CS are both features of *synsem*
- However, nothing constrains its own OPT or OPT-CS value
- Rather, heads may constrain certain arguments to be [OPT -] (can't be dropped) or [OPT +] (must be dropped)
- OPT-CS is a "junk slot" to allow a lexical item to store information about how an argument will be interpreted if it's unexpressed
- The opt-comp rule will identify the OPT-CS value and HOOK.INDEX.COG-ST values of any argument it caches out as unrealized
- Because the HOOK.INDEX of every argument is identified with an ARGn position of the head's KEYREL, this information ends up in the MRS
- Note that we're not associating pronouns or quantifier relations with these argument positions

The Matrix opt-comp type

```
basic-head-opt-comp-phrase := head-valence-phrase & head-only &
                              head-compositional &
  [ INFLECTED #infl,
    SYNSEM canonical-synsem &
    [ ..CAT [ VAL [ SUBJ #subj, COMPS #comps, SPR #spr, SPEC #spec ],
              MC #mc, POSTHEAD #ph ],
      MODIFIED #mod ],
    HEAD-DTR [ INFLECTED #infl & +,
               ..CAT [ VAL [ SUBJ #subj, SPR #spr, SPEC #spec,
                             COMPS < unexpressed &
                                   [ OPT +, OPT-CS #def,
                                     ..INDEX.COG-ST #def ] . #comps >],
                                     MC #mc, POSTHEAD #ph ],
               ..CONT.HOOK.INDEX event,
               MODIFIED #mod ],
    C-CONT [ RELS <! !>, HCONS <! !> ] ].
```

Argument optionality: Your tasks (if applicable)

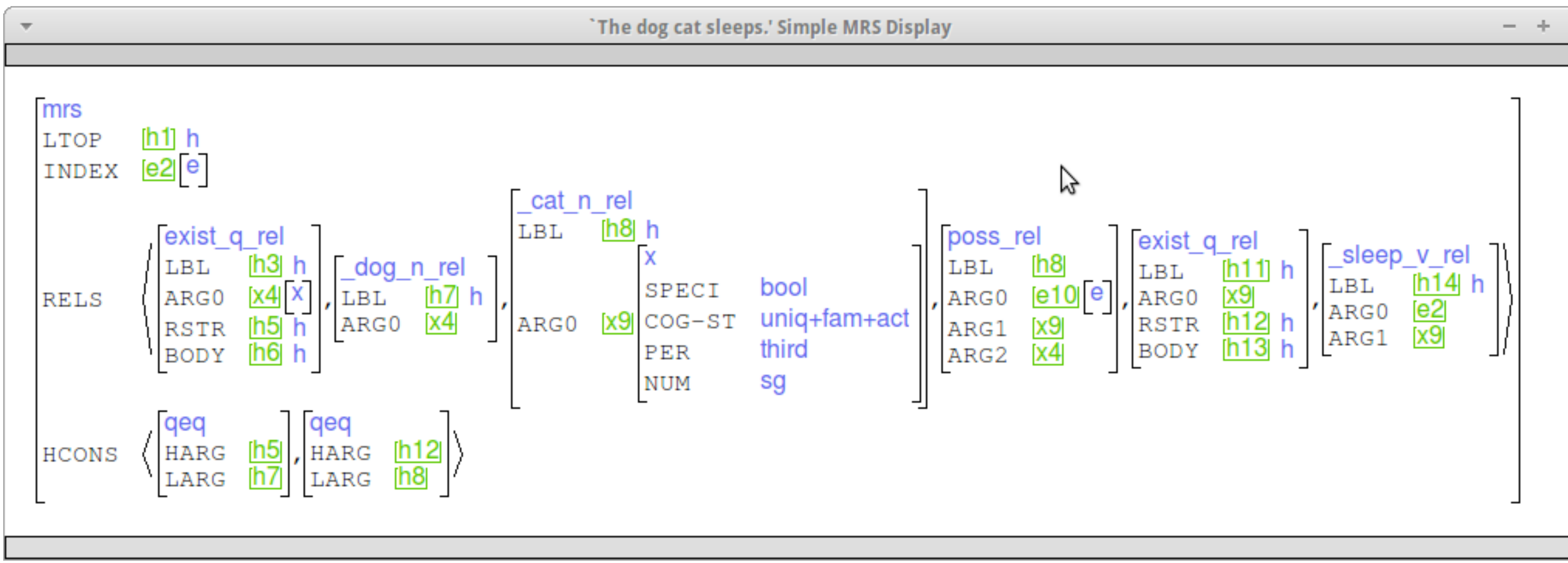
- Check that the analysis provided by the customization system covers the syntactic facts of your language
  - If it doesn't, post to Canvas
- Determine (as best you can) the semantic facts: How are arguments interpreted when they are dropped?
  - If object markers are usually required for object drop, do you get INI without object markers and DNI with?
- Constrain the OPT-CS value of droppable arguments to capture the semantic facts you discover. This may require making multiple subtypes of verbs.

Overview

- Software review
- Modification
- Discourse status
- Argument optionality
- Possessives
- Precision grammars & corpus data

Possessives

- Brand new library :)
- Target semantic representation:



Possessives: Your tasks

- Check whether the MRS looks good for the analysis you got from the customization system

- Implement any fixes that have come through Canvas

Overview

- Software review
- Modification
- Discourse status
- Argument optionality
- Possessives
- Precision grammars & corpus data

Precision grammars & corpus data: Overview

- Theoretical motivation
- “Beauty and the Beast” (Baldwin et al 2005) methodology
- Results (Baldwin et al 2005)
- Your grammars
- Precision grammars and NLP

Theoretical motivation

- Corpora as a sole source of evidence are inadequate:
  - Limited in size; may not reflect the full range of grammatical phenomena
  - Contain performance errors and may reflect other extragrammatical factors
  - Can only provide positive (attested) examples, and not contrasting negative ones
- Intuitions as a sole source of evidence are inadequate:
  - Grammaticality is neither homogeneous nor categorical
  - Grammaticality judgments are typically elicited in unnatural context vaccuums
  - Social/cultural biases color judgments
  - Relying solely on intuitions limits linguists to only the data they have the

Combine the two types of data for better results!

- Grammar engineering provides sophisticated ways of doing so
- Precision grammars encode a sharp notion of grammaticality
- User grammar as a representation of intuitions
- Use the corpus as a source of further data to explore
- Process the corpus with the grammar

“Beauty and the Beast” methodology

- Randomly select 20,000 “sentence tokens” from BNC
- Strip punctuation, POS-tag, tokenize proper names and number expressions, regularize to American spelling
- Select those strings with a full lexical span (32%)
- Process these strings with the ERG to isolate those that can’t presently be parsed
- Use treebanking technology/methodology to validate parses
- Propose paraphrases of the unparseable strings until the ERG is able to parse one

Results: Grammar coverage (2003)

- 57% of strings parsed
- 83% of those parsing included a preferred parse within the forest
- Average ambiguity for 10-20 word strings: 64 analyses

NB: 2017 ERG coverage is much higher, more like 85-90%

Results: Causes of parse failure

Cause of parse failure	Frequency	Category
Missing lexical entry	41%	grammar
Missing construction	39%	grammar
Fragment	4%	grammar
Preprocessor error	4%	neither
Parser resource limitations	4%	neither
Ungrammatical string	6%	corpus
Extragrammatical string	2%	corpus

Missing lexical entries

- Incomplete categorization of existing lexical items
    - *table* as a verb
    - ‘universal grinder’
- Syntactically-marked multi-word expressions
    - *take off,* verb + *up*
    - *off screen, at arm’s length*
    - These are high frequency: verb-particles constitute 1.6% of BNC word tokens

Missing lexical entries

- Drawbacks to introspection alone: subtle gaps like transitive *suffer*
- Drawbacks to corpus data alone: *tell* in the ‘discover’ sense:
    - @Not sure how you can tell
    - Can/could you tell?
    - Are you able to tell?
    - *They might/ought to tell.
    - How might you tell?
    - *How ought they to tell?

Missing constructions

- *@However pissed off* we might get from time to time, ...
- ERG specifically disallowed this
- -> Corpus data as a check on introspection.
- Further corpus investigations surprised us

Missing constructions

- @He's a good player and a *hell of a* nice guy, too.
- Baldwin et al present this as a semantic puzzle:
    - Apparent syntactic attachment to NP/N' because of definiteness restrictions
    - Semantic attachment to adjective (intensifier)
- Still complex, but less mysterious, in a world where definiteness is encoded as a feature of indices

Missing constructions

- @The price of train tickets can vary from *the reasonable* to *the ridiculous.*
- Exocentric NPs not limited to classes of people.
- What adjectives can appear here, and with what kinds of referents?

Missing constructions

- @This sort of response was also noted in the sample task for *criterion 2.*
- 'Title' (common noun) + series element
- Frequent in corpora (like dates, number names, quotatives)
- Not usually remarked on in syntactic theory

Extragrammatical strings

- Prime example: Structural markup
  - @There are five of these general arrest conditions: (a) the name of...
- Preprocessing requires interface to the grammar:
  - @(I) The Mrs Simpson could never be Queen.
  - @(I) rarely took notes during the thousands of informal conversational interviews.

Summary

- Methodology goes beyond merely using a corpus for inspiration
  - Encoding intuitions in the grammar
  - Use the grammar to process the corpus, twice:
    - filter out ‘easy’ cases
    - investigate where in a string the problems are
- Provides detailed feedback to grammar developers
- Turns up previously unnoted constructions, which might be too low frequency to be found otherwise

How about your grammars?

- Role of corpora so far?
- How to get from current state to something that could turn up unexpected constructions?
- What are you seeing in your test corpora?

Test corpora specs

- 10-20 sentences is all that's required
- Sentence tokenized
- Strings formatted as your grammar expects
- Imported into [incr tsdb()] skeleton
- Ideally you have IGT or at least a translation available so you have some way of knowing what's going on in the examples

If you're feeling ambitious

- Test corpora can be bigger, but don't bother unless
  - You have access to an external (already digitized) lexical resource
  - Your morphophonology is simple, or you have a morphophonological analyzer
- Any single [incr tsdb()] skeleton should have no more than 1,000 sentences

Precision grammars and NLP

- Baldwin et al: Notion of grammaticality cuts down on spurious ambiguity and is crucial in avoiding ill-formed output in generation

- Bender et al 2015; Flickinger et al 2017: Value of encoding grammatical constraints in producing consistent, elaborated annotations

- Cost: How does grammar development compare to treebank development?

Precision grammars & corpus data: Overview

- Theoretical motivation
- “Beauty and the Beast” (Baldwin et al 2005) methodology
- Results (Baldwin et al 2005)
- Your grammars
- Precision grammars and NLP

Overview

- Software review
- Modification
- Discourse status
- Argument optionality
- Possessives
- Precision grammars & corpus data