Integrated Semantic Framework

Le Tuan Anh, Francis Bond

A member of Francis Bond lab

05 August 2017

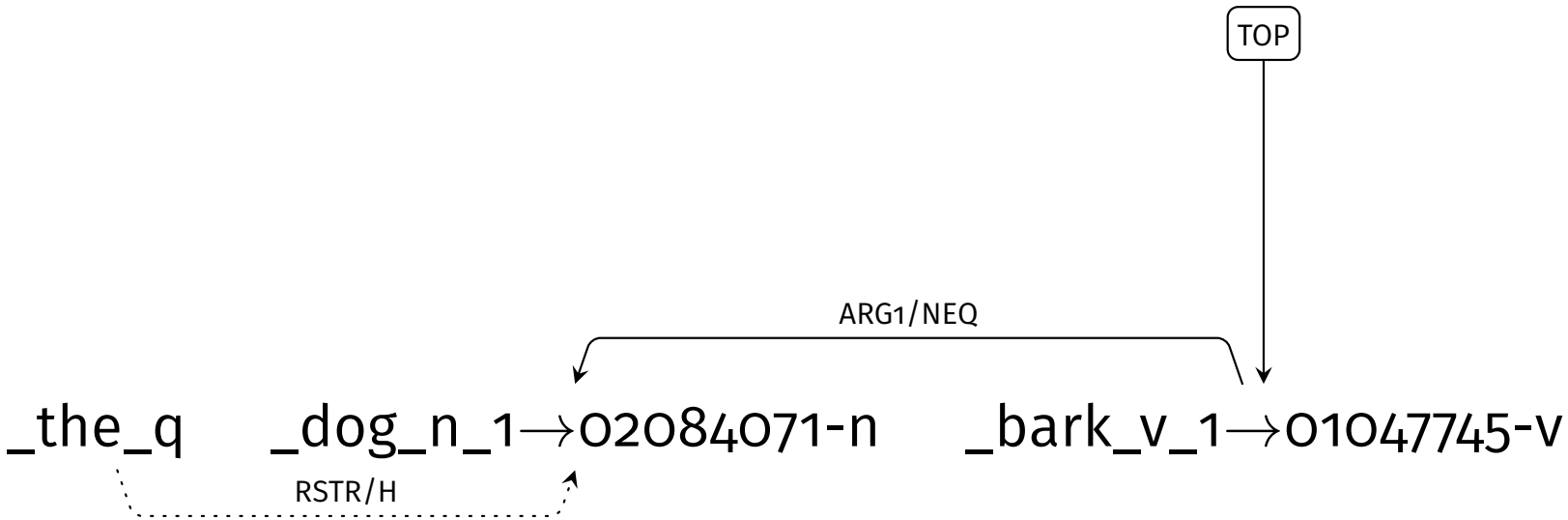
What ISF does

- Bring lexical semantics and structure semantics together
- Wordnet for lexical semantics
- DMRS for structure semantics
- Integrated Semantic Framework
  - parse to DMRS (ERG, Jacy) — take first best
  - map predicates to wordnet synsets
    - some MWEs (rewriting to single pred or decomposing)
  - disambiguate with LESK or UKB (or MFS)
  - compare to hand annotated corpora *The Adventure of the Speckled Band* (also DANC, CatB, gloss, ...)
    - most things match
    - hand munging anything that does not match
      want to see what we need

The easy cases: predicate == wordnet lemma

The dog barked.

- _dog_n_1 → 02084071-n
- _bark_v_1 → 01047745-v

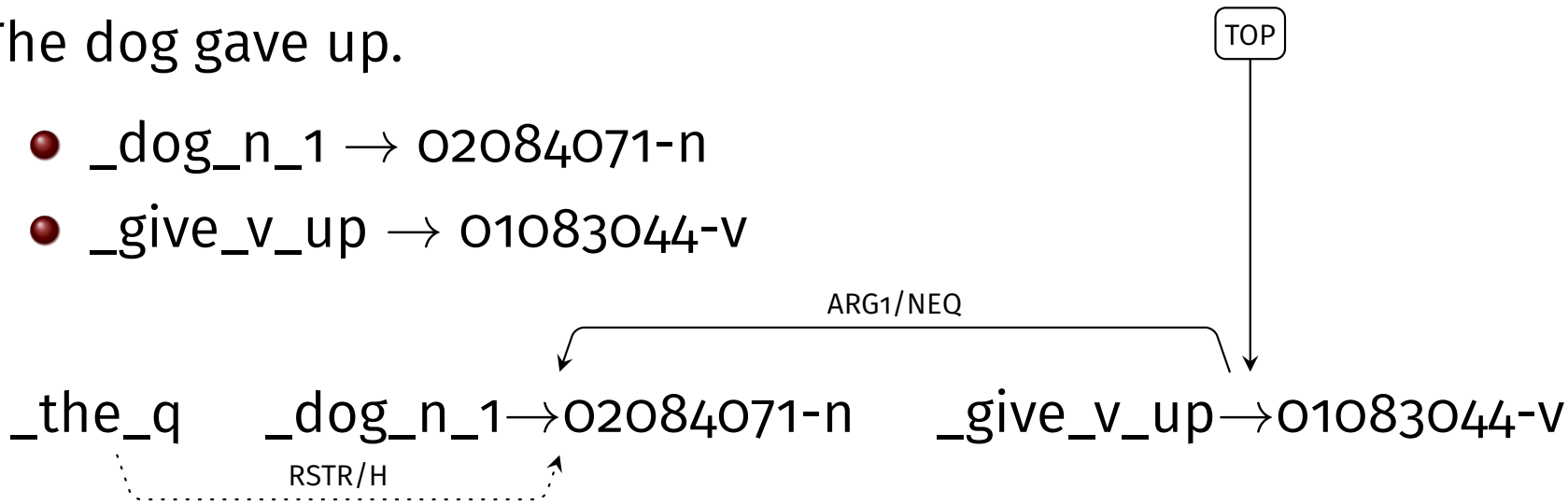


By far the most common case.

Almost easy: one predicate == wordnet lemma

The dog gave up.

- _dog_n_1 → 02084071-n
- _give_v_up → 01083044-v

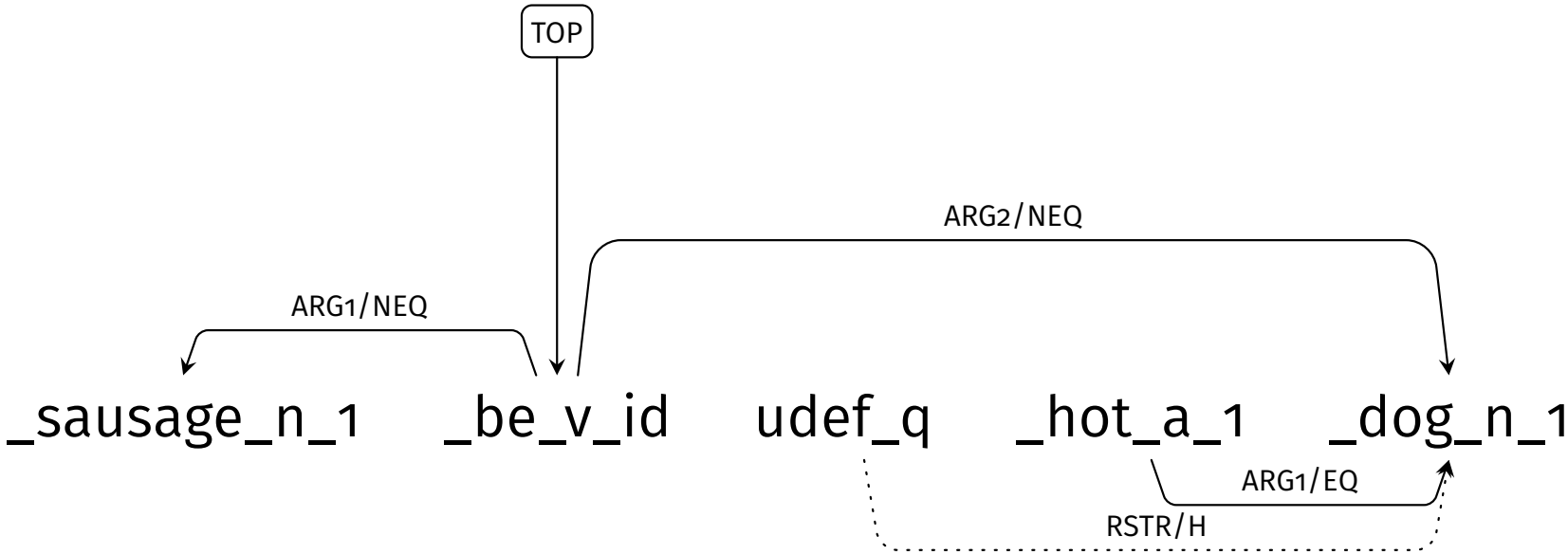


Can we add a feature to `give_up_v2`?
[HEADWORD "give up"]
headword for lookup in other resources, omit if same as ORTH

Harder: Concepts to structures

Sausages are hot-dogs.

- _sausage_n_1 → 07675627-n
- _hot_a_1 → ???
- _dog_n_1 → ???



Concepts to structures | Solution

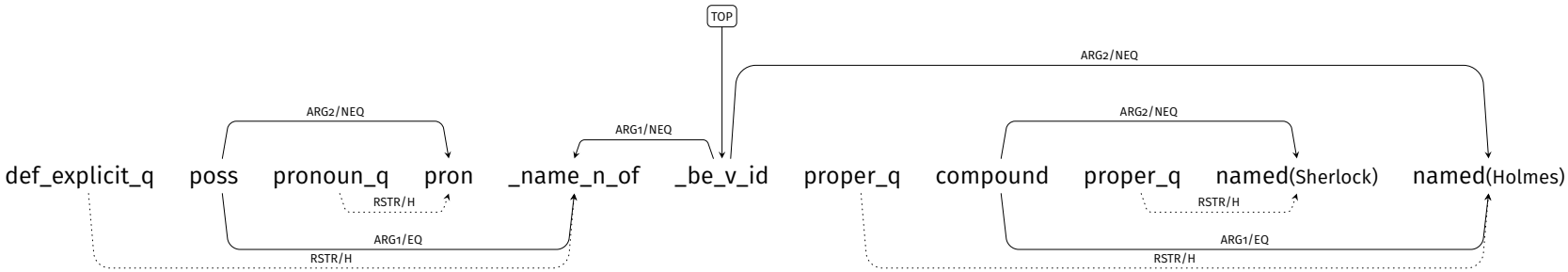
- Transform complex structures into a single predicate: udef_q + _hot_a_1 + _dog_n_1 $\rightarrow$ udef_q + _hot+dog_n_1
- Equivalent of adding a new lexical entry to ERG
- This maps to 07676602-n “frankfurter”

Compound names

My name is Sherlock Holmes.

- named(Sherlock) → 09604451-n
- named(Holmes) → 09604451-n

Similar solution is adopted: merge to a single Sherlock+Holmes.



Missing senses from Wordnet

- "Holmes, the Scotland Yard **Jack-in-office**!"
- Two days ago some repairs were started in the west wing of the building, and my bedroom wall has been pierced, **so that** I have had to move into the chamber in which my sister died, and to sleep in the very bed in which she slept.
- **At first** I thought that she had not recognised me, but as I bent over her she suddenly shrieked out in a voice which I shall never forget, 'Oh, my God!

Solution: Add new synsets to Wordnet

Empty Words

A month ago, **however**, a dear friend, whom I have known for many years, has done me the honour to ask my hand in marriage.

- ERG doesn't generate any predicate for *however*
- Wordnet annotators annotated *however* with 00028797-r: 'lit: by contrast; on the other hand'

More generally, the set of words that are thought of as interesting are different between WN and ERG — cases where they disagree are most often (in my opinion), places where they should be in both, with some principled disagreements: predicates with adjectives (WN only), prepositions (ERG only), ...

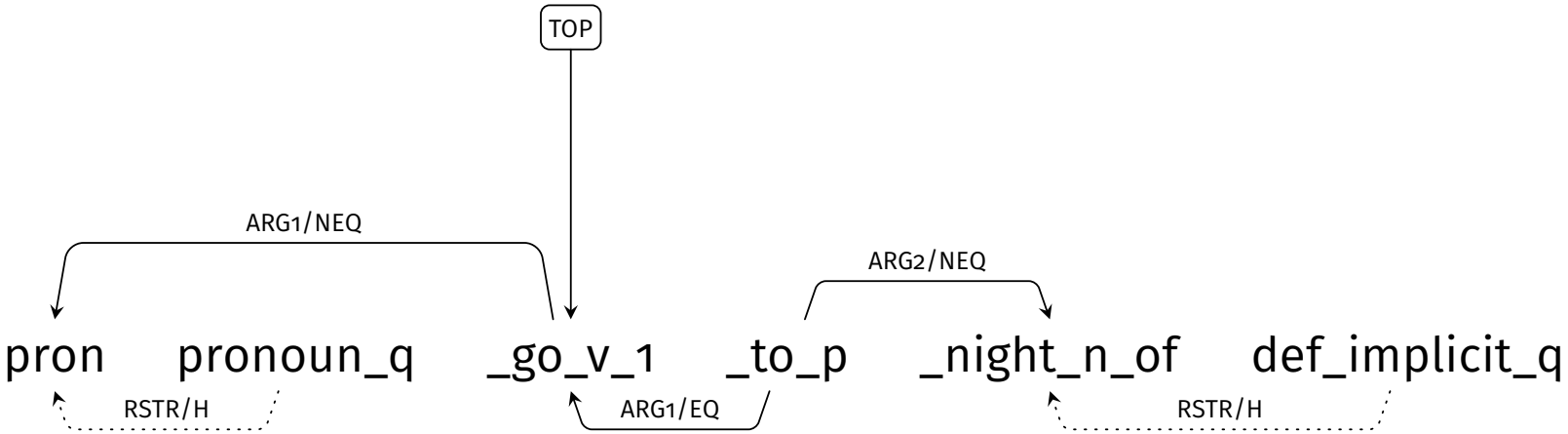
Surprising treebank choices (i)

I cannot tell where it came from–**perhaps** from the next room, **perhaps** from the lawn.

- ERG doesn't generate any predicate for *perhaps*
- Although we think it can and should

Surprising treebank choices (ii)

I go to-night.

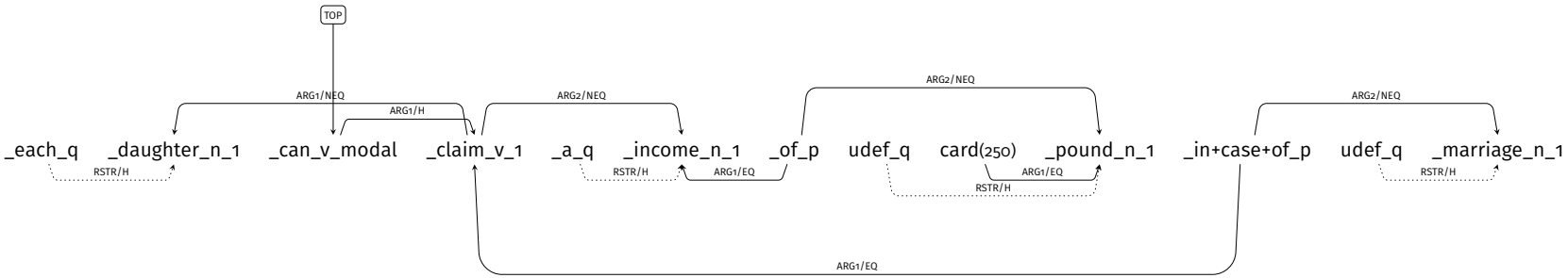


... to, eh, where?
Interesting tokenization issue, but we wondered if this interpretation should have been rejected.

with or without *of*

Each daughter can claim an income of 250 pounds, in case of marriage.

- Wordnet annotators annotated *in* and *case*
- ERG generated in+case+of
- But we can also just say *[just] in case [I need it]* with a similar meaning, ...



Other things we can do with wordnets

- Populate lexicons for specific tasks
  - *square* — 58 translations: Afrikaans–Yiddish and with the right sense
  - *triangle* — 53 translations: Afrikanns–Volapük
- Help grammarians with interesting sub-classes
  - Control verbs
    - translations of control verbs are often control verbs
    - identify which synsets are made of control verbs (match with the ERG)
    - use it to list candidates (for INDRA, Zhong, ...)
- Corpus of glosses

Discussion

- We still have issues with recognizing idioms (like bare PPs: *at school*) and larger (*make up ones mind*). The ERG recognizes them, but they are not marked in the MRS.
- We are still mainly working on our small corpora, but the end-to-end system works (github: `letuananh/intsem.fx` if you want to play)
- Surprisingly few MWEs (3–4%) but they really change the meaning
- Tokenization issues remain fraught with excitement: *white counter-paned*; *case-book*; *good-morning*