

Princeton Glosstag Corpus

Alexandre Rademaker, IBM Research and FGV/EMAp, DELPH-In Summit 2021

The PWN glosses

- The definitions of Princeton WordNet were used in many projects
 - <https://wordnet.princeton.edu/download/standoff-files>
 - Semantically annotated gloss corpus
 - Logical Forms of the glosses
 - XWN - eXtended WordNet (University of Texas at Dallas)
 - UXWN (Italy)
- word sense disambiguation
 - UKB, <http://ixa2.si.ehu.eus/ukb/>
 - Lesk algorithms

Motivation

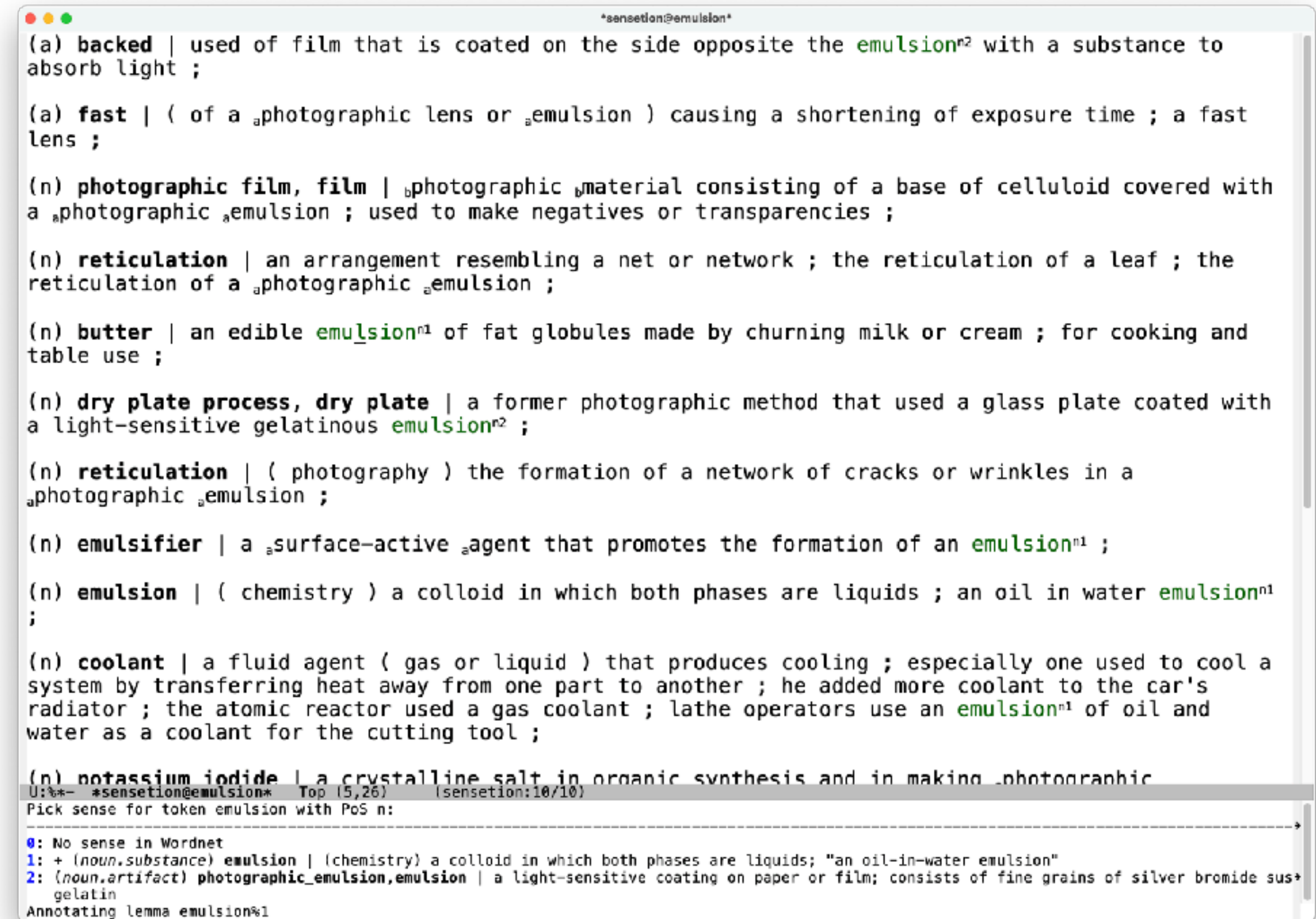
- WN as a lightweight ontology
 - <http://wn.mybluemix.net/synset?id=07848338-n>
- dataset for sense disambiguation
 - In the semantic representation
 - In the syntactic structure
- applications: QA, text entailment etc
- Mapping ERG predicates, Wordnet, Propbank (Unifying LR)

Crouch, Dick, and Tracy Holloway King. "Unifying lexical resources." Proceedings of the interdisciplinary workshop on the identification and representation of verb features and verb classes. 2005.

- Improve WN
 - <http://wn.mybluemix.net/synset?id=04617562-n> (word not in example)
 - soundness and completeness

Gloss corpus sense annotation

- Glosstag corpus is partially annotated
 - <https://wordnetcode.princeton.edu/glosstag.shtml> (206711 entries)
 - Considered more consistent than SemCor by Christiane Fellbaum
- <http://github.com/own-pt/glosstag>
 - <https://aclanthology.org/2019.gwc-1.48/>
 - Emacs mode for annotation <https://github.com/own-pt/sensation.el>
 - Release format (XML vs JSON)



```
*sensation@emulsion*
(a) backed | used of film that is coated on the side opposite the emulsionn2 with a substance to
absorb light ;

(a) fast | ( of a photographic lens or emulsion ) causing a shortening of exposure time ; a fast
lens ;

(n) photographic film, film | photographic material consisting of a base of celluloid covered with
a photographic emulsion ; used to make negatives or transparencies ;

(n) reticulation | an arrangement resembling a net or network ; the reticulation of a leaf ; the
reticulation of a photographic emulsion ;

(n) butter | an edible emulsionn1 of fat globules made by churning milk or cream ; for cooking and
table use ;

(n) dry plate process, dry plate | a former photographic method that used a glass plate coated with
a light-sensitive gelatinous emulsionn2 ;

(n) reticulation | ( photography ) the formation of a network of cracks or wrinkles in a
photographic emulsion ;

(n) emulsifier | a surface-active agent that promotes the formation of an emulsionn1 ;

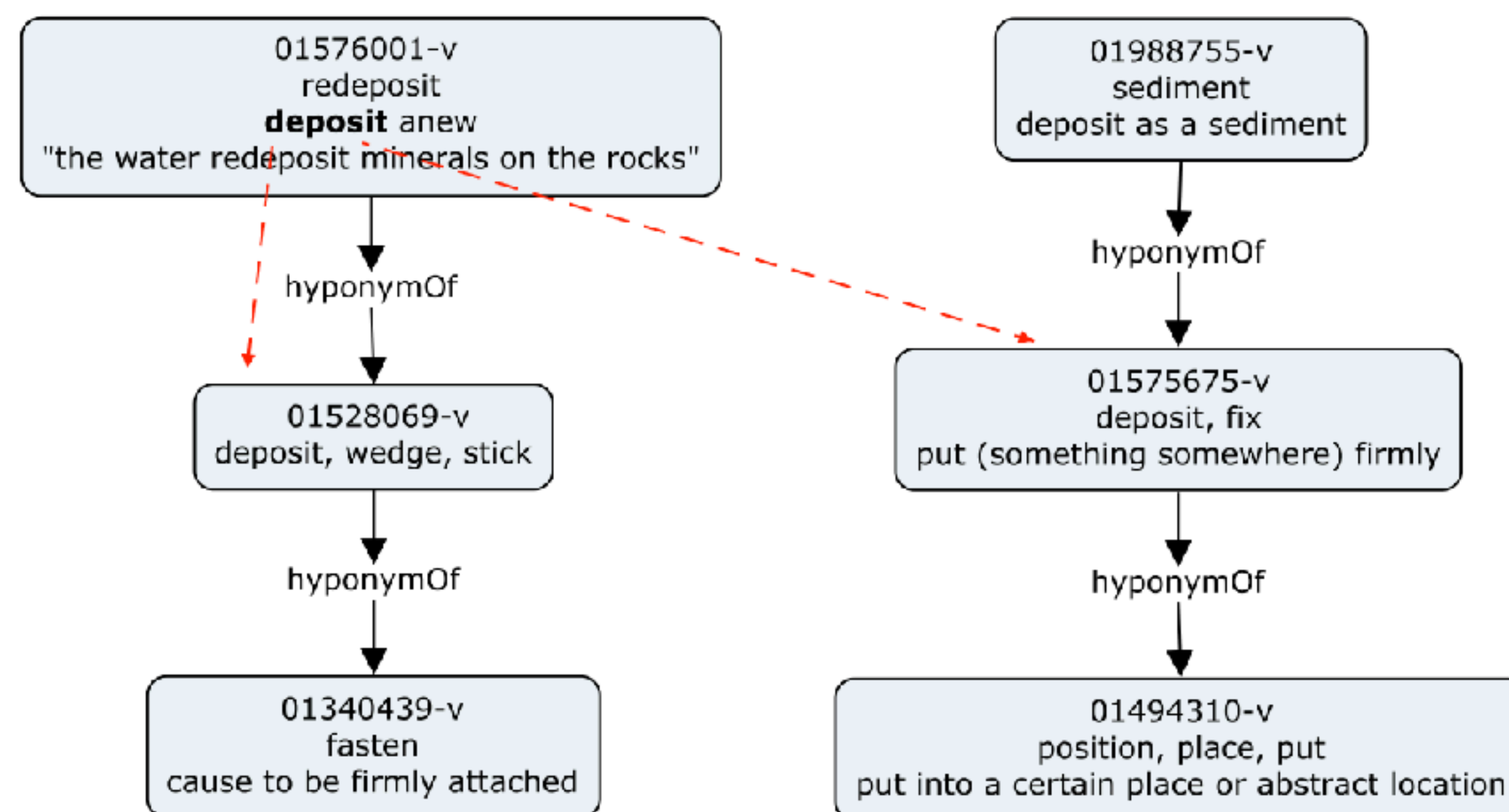
(n) emulsion | ( chemistry ) a colloid in which both phases are liquids ; an oil in water emulsionn1
;

(n) coolant | a fluid agent ( gas or liquid ) that produces cooling ; especially one used to cool a
system by transferring heat away from one part to another ; he added more coolant to the car's
radiator ; the atomic reactor used a gas coolant ; lathe operators use an emulsionn1 of oil and
water as a coolant for the cutting tool ;

(n) potassium iodide | a crystalline salt in organic synthesis and in making photographic
U: %*~ *sensation@emulsion* Top (5,26) (sensation:10/10)
Pick sense for token emulsion with PoS n:

0: No sense in Wordnet
1: + (noun.substance) emulsion | (chemistry) a colloid in which both phases are liquids; "an oil-in-water emulsion"
2: (noun.artifact) photographic_emulsion,emulsion | a light-sensitive coating on paper or film; consists of fine grains of silver bromide sus+
gelatin
Annotating lemma emulsion%1
```

Gloss corpus sense annotation



Gloss corpus sense annotation

```
*sensation@nation*

(n) oil cartel | a cartel of companies or nations formed to control the production and distribution of oil ;
(n) Liberal Party | a political party in Australia , Canada , and other nations , and formerly in Great Britain ;
(n) Economic Commission for Africa | the commission of the Economic and Social Council of the United Nations that is
concerned with economic development of African nations ;
(n) Statistical Commission | the commission of the Economic and Social Council of the United Nations that is concerned with
statistical data from member nations ;
(n) IBRD, World Bank, International Bank for Reconstruction and Development | a United Nations agency created to assist
developing nations by loans guaranteed by member governments ;
(n) nation | a federation of tribes ( especially Native American tribes ) ; the Shawnee nationn3 ;
(n) customs union | an association of nations to promote free trade within the union and set common tariffs for nations that
are not members ;
(n) ally | a friendly nationn2,1 ;
(n) world council | a council with representatives from different nations ;
(n) United States Supreme Court, Supreme Court of the United States, Supreme Court | the highest federal court in the United
States ; has final appellate jurisdiction and has jurisdiction over all other courts in the nation ;
(n) Mossad | the Israeli foreign intelligence agency ; the primary focus of the Mossad is on Arab nations ;
(n) hegemony | the dominance or leadership of one social group or nation over others ; the hegemony of a single member state is
not incompatible with a genuine confederation ; to say they have priority is not to say they have complete hegemony ; the
consolidation of the United States' hegemony over a new international economic system ;
(n) mercantile system, mercantilism | an economic system ( Europe in 18th century ) to increase a nation's wealth by government
regulation of all of the nation's commercial interests ;
(n) civil law, the body of laws established by a state or nation for its own regulation ;

U:***- *sensation@nation* 59% (123,24) (sensation:26/205)
Pick sense for token nation with PoS n:
----->
0: No sense in Wordnet
1: + (noun.group) body_politic, res_publica, commonwealth, land, country, nation, state | a politically organized body of people under a
government; "the state has elected a new president"; "African nations"; "students who had come to the nation's capitol"; "the
country's largest manufacturer"; "an industrialized land"
2: + (noun.group) country, land, nation | the people who live in a nation or country; "a statement that sums up the nation's mood"; →
announced to the nation"; "the whole country worshipped him"
3: (noun.group) nation | a federation of tribes (especially Native American tribes); "the Shawnee nation"
Annotating lemma nation%1
```


Data format, tokenization etc

```
</terms>
<keys>
  <sk>hit%1:04:03::</sk>
</keys>
<gloss desc="orig">
  <orig>(baseball) a successful stroke in an athletic contest (especially in baseball); "he c
</gloss>
<gloss desc="text">
  <text>( baseball ) a successful stroke in an athletic contest ( especially in baseball ) ;
</gloss>
<gloss desc="wsd">
  <classif type="cat">
    <wf id="n00043902_tok1" sep="" tag="ignore" type="punc"></wf>
    <wf id="n00043902_tok2" lemma="baseball%1" sep="" tag="un">baseball</wf>
    <wf id="n00043902_tok3" tag="ignore" type="punc"></wf>
  </classif>
  <def id="n00043902_d">
    <wf id="n00043902_wf1" lemma="a" pos="DT" tag="ignore">a</wf>
    <wf id="n00043902_wf2" lemma="successful%3" pos="JJ" tag="auto">
      <id id="n00043902_id.1" lemma="successful" sk="successful%3:00:00::"/>successful</wf>
    <wf id="n00043902_wf3" lemma="stroke%1|stroke%2" pos="NN" tag="un">stroke</wf>
    <wf id="n00043902_wf4" lemma="in" pos="IN" tag="ignore">in</wf>
    <wf id="n00043902_wf5" lemma="an" pos="DT" tag="ignore">an</wf>
    <cf coll="a" id="n00043902_wf6" lemma="athletic%3" pos="JJ" tag="un">
      <glob coll="a" glob="auto" id="n00043902_coll.a" lemma="athletic_contest%1" tag="auto">
        <id coll="a" id="n00043902_id.3" lemma="athletic contest" sk="athletic_contest%1:11:00::
      </glob>athletic</cf>
    <cf coll="a" id="n00043902_wf7" lemma="contest%1|contest%2" pos="NN" tag="un">contest</cf>
    <wf id="n00043902_wf8" pos="(" sep="" tag="ignore" type="punc"></wf>
    <wf id="n00043902_wf9" lemma="especially%4" pos="RB" tag="ignore">especially</wf>
    <wf id="n00043902_wf10" lemma="in" pos="IN" tag="ignore">in</wf>
    <wf id="n00043902_wf11" lemma="baseball%1" pos="NN" sep="" tag="man">
      <id id="n00043902_id.4" lemma="baseball" sk="baseball%1:04:00::"/>baseball</wf>
    <wf id="n00043902_wf12" pos=")" sep="" tag="ignore" type="punc"></wf>
    <wf id="n00043902_wf13" pos=":" tag="ignore" type="punc"></wf>
  </def>
  <ex id="n00043902_ex1">
    <qf rend="dq">
      <wf id="n00043902_wf14" lemma="he" tag="ignore">he</wf>
      <wf id="n00043902_wf15" lemma="come%2" tag="un">came</wf>
      <wf id="n00043902_wf16" lemma="all%3|all%4" tag="un">all</wf>
      <wf id="n00043902_wf17" lemma="the" tag="ignore">the</wf>
      <wf id="n00043902_wf18" lemma="way%1|way%4" tag="un">way</wf>
      <wf id="n00043902_wf19" lemma="around%4" tag="un">around</wf>
      <wf id="n00043902_wf20" lemma="on" tag="ignore">on</wf>
      <wf id="n00043902_wf21" lemma="Williams%1" tag="un">Williams'</wf>
      <wf id="n00043902_wf22" lemma="hit%1|hit%2" sep="" tag="auto">
        <id id="n00043902_id.2" lemma="hit" sk="hit%1:04:03::"/>hit</wf>
    </qf>
  </ex>
</gloss>
```

U:--- noun.xml 1% (3939,0) (nXML Validated:0)

```
<term>spectacle</term>
</terms>
<keys>
  <sk>spectacle%1:04:00::</sk>
</keys>
<gloss desc="orig">
  <orig>a blunder that makes you look ridiculous; used in the phrase 'make a spectacle of' yourself</orig>
</gloss>
<gloss desc="text">
  <text>a blunder that makes you look ridiculous ; used in the phrase ' make a spectacle of ' yourself</text>
</gloss>
<gloss desc="wsd">
  <def id="n00075471_d">
    <wf id="n00075471_wf1" lemma="a" pos="DT" tag="ignore">a</wf>
    <wf id="n00075471_wf2" lemma="blunder%1|blunder%2" pos="NN" tag="man">
      <id id="n00075471_id.2" lemma="blunder" sk="blunder%1:04:00::"/>blunder</wf>
    <wf id="n00075471_wf3" lemma="that" pos="WDT" tag="ignore">that</wf>
    <wf id="n00075471_wf4" lemma="make%1|make%2" pos="VBZ" tag="man">
      <id id="n00075471_id.1" lemma="make" sk="make%2:30:00::"/>makes</wf>
    <wf id="n00075471_wf5" lemma="you" pos="PRP" tag="ignore">you</wf>
    <wf id="n00075471_wf6" lemma="look%1|look%2" pos="VB" tag="man">
      <id id="n00075471_id.3" lemma="look" sk="look%2:39:01::"/>look</wf>
    <wf id="n00075471_wf7" lemma="ridiculous%3" pos="JJ" sep="" tag="un">ridiculous</wf>
    <wf id="n00075471_wf8" pos=":" tag="ignore" type="punc"></wf>
  </def>
  <aux tag="ignore">
    <wf id="n00075471_tok1" lemma="use%2|used%3" tag="un">used</wf>
    <wf id="n00075471_tok2" lemma="in" tag="ignore">in</wf>
    <wf id="n00075471_tok3" lemma="the" tag="ignore">the</wf>
    <wf id="n00075471_tok4" lemma="phrase%1|phrase%2" tag="un">phrase</wf>
    <qf rend="sq">
      <wf id="n00075471_tok5" lemma="make%1|make%2" tag="un">make</wf>
      <wf id="n00075471_tok6" lemma="a" tag="ignore">a</wf>
      <wf id="n00075471_tok7" lemma="spectacle%1" tag="un">spectacle</wf>
      <wf id="n00075471_tok8" lemma="of" sep="" tag="ignore">of</wf>
    </qf>
    <wf id="n00075471_tok9" lemma="yourself" sep="" tag="ignore">yourself</wf>
    <wf id="n00075471_tok10" tag="ignore" type="punc">;</wf>
  </aux>
</gloss>
</synset>
<synset id="n00075618" ofs="00075618" pos="n">
  <terms>
    <term>ballup</term>
    <term>balls-up</term>
    <term>cockup</term>
    <term>mess-up</term>
  </terms>
  <keys>
    <sk>ballup%1:04:00::</sk>
  </keys>
```

U:--- noun.xml 1% (10807,0) (nXML Validated:0)

Mark saved where search started

Processing glosses with ERG

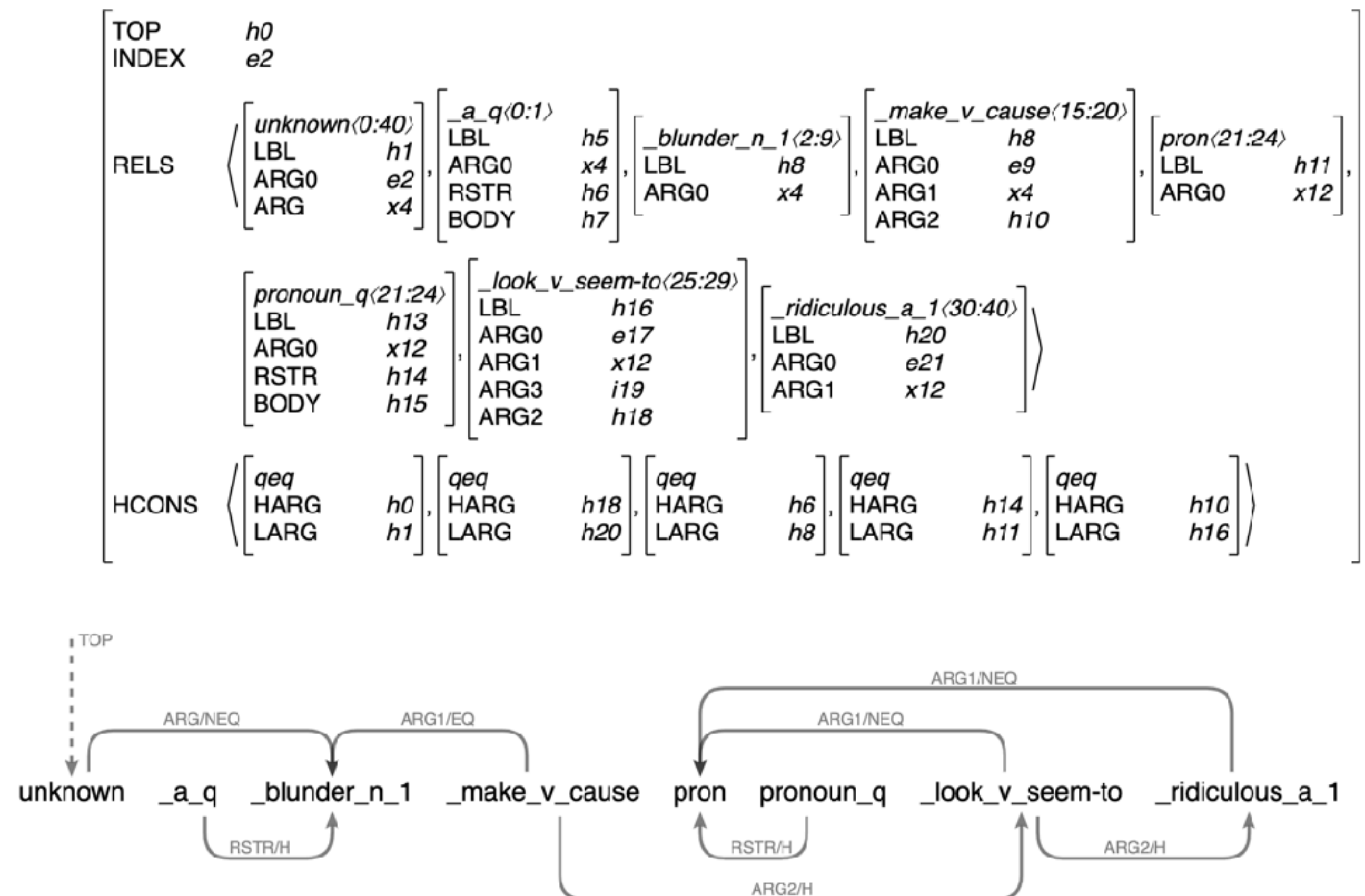
- Golden profile at <http://svn.delph-in.net/erg/trunk/tsdb/gold/omw/>
- ace/config-dict.tdl
- no maxent model training (?)

tsdb(1) 'omw' Performance Profile

| Aggregate | items # | etasks Ø | filter % | edges Ø | first Ø (s) | total Ø (s) | tcpu Ø (s) | tgc Ø (s) | space Ø (kb) |
|------------------------|-------------|-----------|----------|------------|--------------|-------------|-------------|--------------|--------------|
| i-length in [25 .. 30] | 5 | -1 | - | 7114 | -1.00 | 2.73 | 2.73 | -1.00 | 467376 |
| i-length in [20 .. 25] | 30 | -1 | - | 4359 | -1.00 | 1.13 | 1.13 | -1.00 | 173109 |
| i-length in [15 .. 20] | 101 | -1 | - | 1954 | -1.00 | 0.40 | 0.40 | -1.00 | 71645 |
| i-length in [10 .. 15] | 310 | -1 | - | 868 | -1.00 | 0.15 | 0.15 | -1.00 | 29418 |
| i-length in [5 .. 10] | 664 | -1 | - | 302 | -1.00 | 0.04 | 0.04 | -1.00 | 10204 |
| i-length in [0 .. 5] | 435 | -1 | - | 92 | -1.00 | 0.01 | 0.01 | -1.00 | 3510 |
| Total | 1545 | -1 | - | 585 | -1.00 | 0.11 | 0.11 | -1.00 | 20834 |

(generated by [incr tsdb()] at 22-Jul-21 (19:55))

Close LaTeX PostScript



The OMW profile

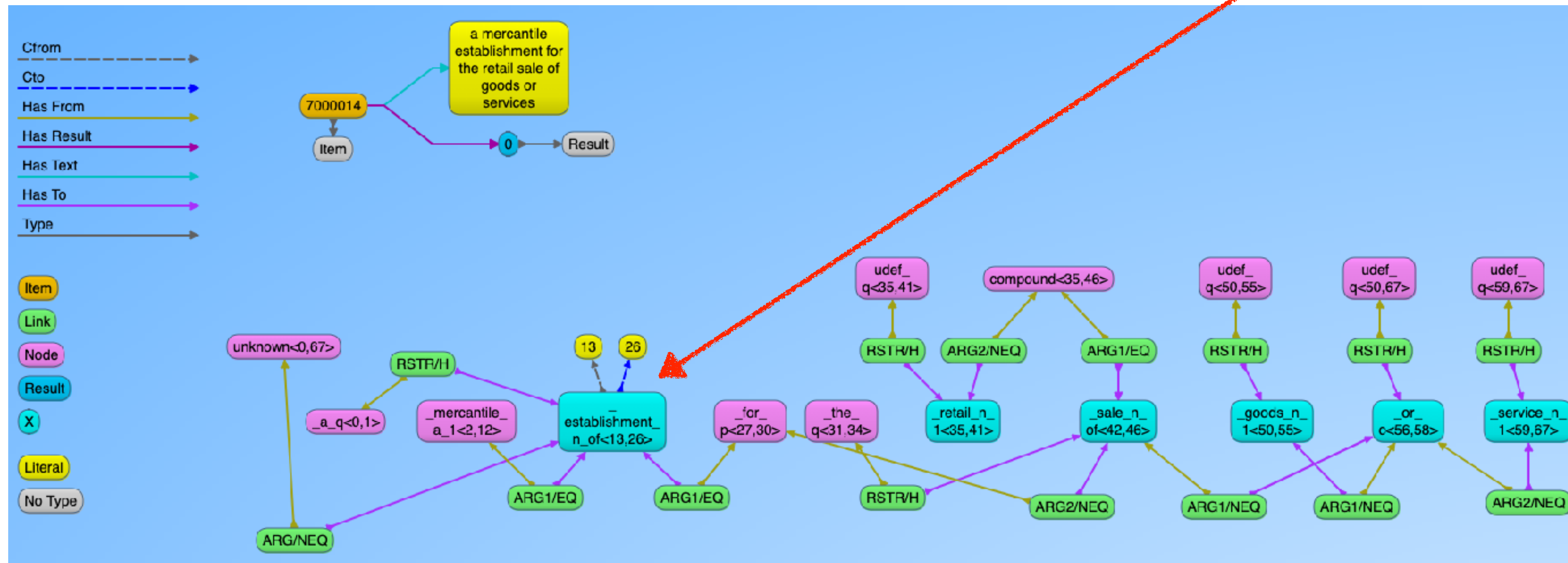
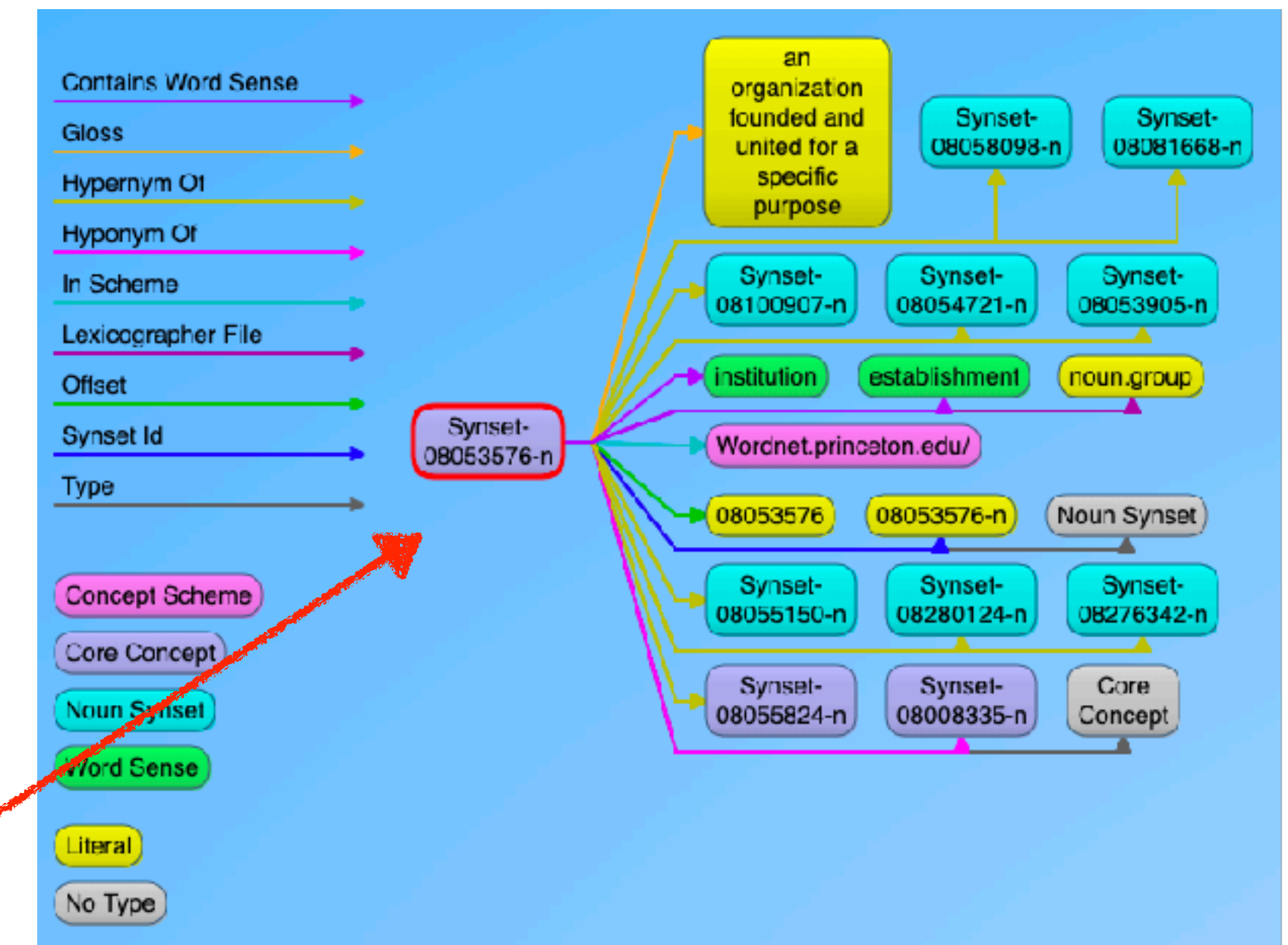
<https://pydelphin.readthedocs.io/en/latest/>

<https://github.com/own-pt/delphin-rdf>

```
(venv) ar@tenis wn % delphin profile-to-rdf --to dmrs -v -p http://example.org -o omw.nq -f nquads ../terg/tsdb/gold/omw
INFO:delphin.cli.profile_to_rdf:Getting parsers for representation: dmrs
INFO:delphin.cli.profile_to_rdf:Converting 1489 analysis of 1545 sentences from ../terg/tsdb/gold/omw
INFO:delphin.cli.profile_to_rdf:Loading the profile
INFO:delphin.cli.profile_to_rdf:Converting the profile
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000043 is not well formed
/Users/ar/venv/lib/python3.9/site-packages/delphin/dmrs/_operations.py:81: DMRSWarning: unusable TOP: h0
  warnings.warn(f'unusable TOP: {top_var}', dmrs.DMRSWarning)
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000059 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000084 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000086 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000111 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000254 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000261 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000305 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000353 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000404 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000462 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000472 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000520 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000541 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000558 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000608 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000657 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000696 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000806 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7000964 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7001051 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7001112 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7001135 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7001147 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7001175 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7001186 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7001197 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7001239 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7001247 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7001266 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7001288 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7002120 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7002167 is not well formed
WARNING:delphin.cli.profile_to_rdf:Result 0 of sentence 7002199 is not well formed
INFO:delphin.cli.profile_to_rdf:Serializing results to omw.nq
INFO:delphin.cli.profile_to_rdf:DONE
(venv) ar@tenis wn %
```


A DMRS in RDF

- RDF graphs can be easily integrated
- Combination of two layers (senses and semantic representation) into a single graph



OWN-PT and OWN-EN

- Portuguese and PWN wordnets in RDF
- RDF/OWL Representation of WordNet, <https://www.w3.org/TR/wordnet-rdf/>
- <https://github.com/own-pt/openWordnet-PT> (first release after +10 years!)
- LMF in <https://github.com/goodmami/wn>, Michael again! 😊

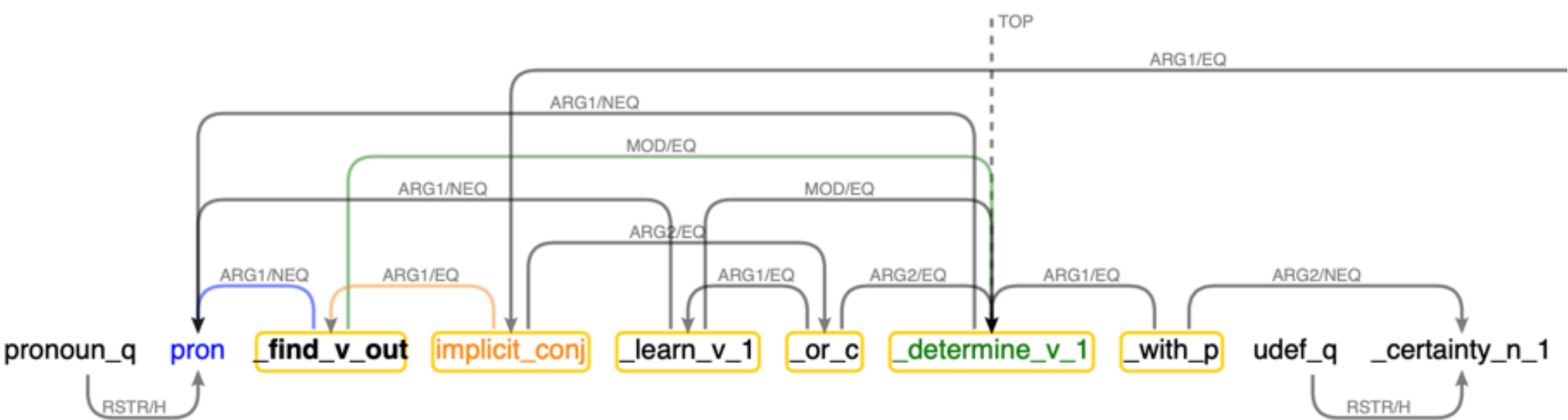
Mapping Tokens to Predicates

- N-N
- abstract vs surface predicates
- MWE detected by ERG
- annotated MWE

Mapping Tokens to Predicates

find out, learn, or determine with certainty, usually by making an inquiry or other effort

[http://wn.mybluemix.net/search?search field=all&term=find out](http://wn.mybluemix.net/search?search%20field=all&term=find%20out)

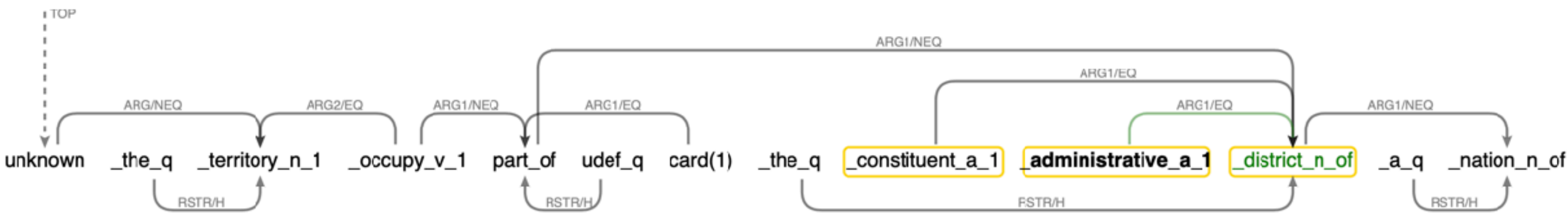


```
((kind "glob" "b")
 (lemmas "find_out%2")
 (tag . "un")
 (glob . "man"))
((kind "cf" "b")
 (form . "find")
 (lemmas "find%1" "find%2")
 (tag . "un")
 (meta
  (pos . "VB"))))
((kind "cf" "b")
 (form . "out")
 (lemmas "out%1" "out%2" "out%3" "out%4")
 (tag . "un")
 (meta
  (pos . "RB")
  (sep . "")))
```

Mapping Tokens to Predicates

the territory occupied by one of the constituent **administrative districts** of a nation

<http://wn.mybluemix.net/synset?id=08491826-n>



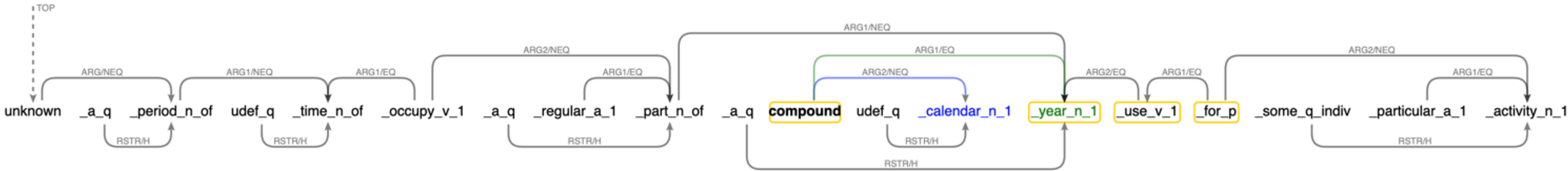
```
((kind "glob" "a")
 (lemmas "administrative_district%1")
 (tag . "auto")
 (senses "administrative_district%1:15:00::")
 (glob . "auto"))
((kind "cf" "a")
 (form . "administrative")
 (lemmas "administrative%3")
 (tag . "un")
 (meta
  (pos . "JJ"))))
((kind "cf" "a")
 (form . "districts")
 (lemmas "district%1" "district%2")
 (tag . "un")
 (meta
  (pos . "NNS"))))
```


Mapping Tokens to Predicates

a period of time occupying a regular part of a **calendar year** that is used for some particular activity

<http://wn.mybluemix.net/synset?id=15202634-n>

```
((kind "glob" "b")
 (lemmas "calendar_year%1")
 (tag . "auto")
 (senses "calendar_year%1:28:00::")
 (glob . "auto"))
(kind "cf" "b")
 (form . "calendar")
 (lemmas "calendar%1" "calendar%2")
 (tag . "un")
 (meta
  (pos . "NN"))))
(kind "cf" "b")
 (form . "year")
 (lemmas "year%1")
 (tag . "un")
 (meta
  (pos . "NN"))))
```



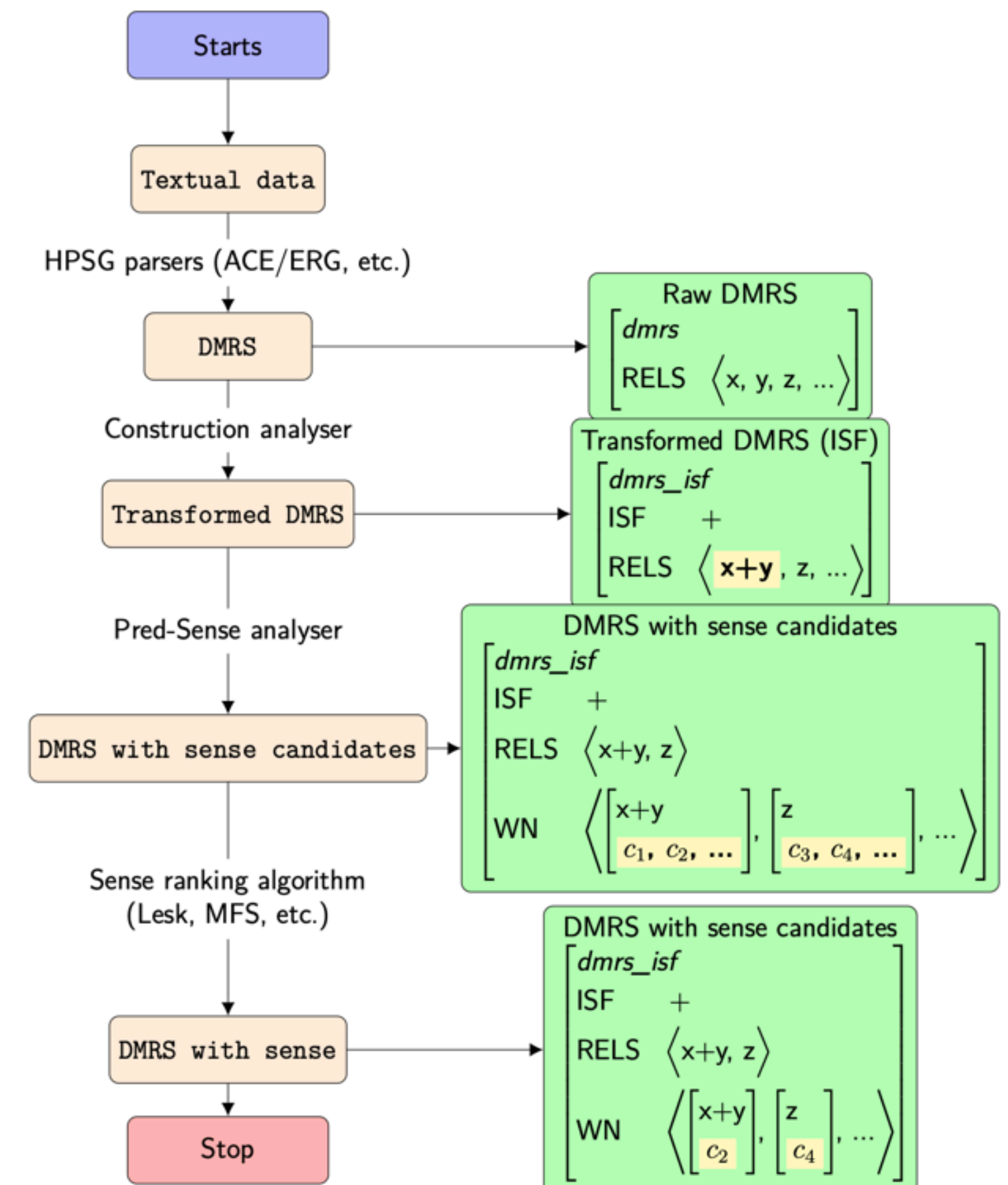
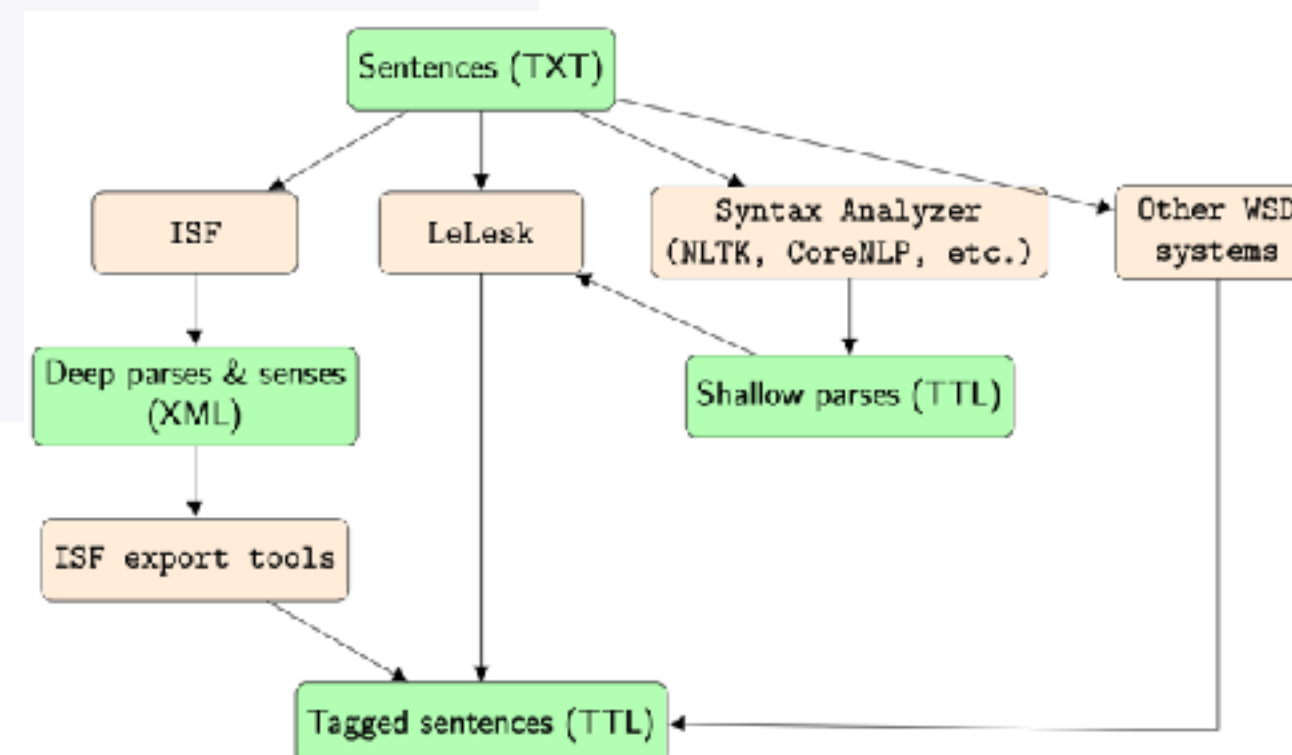
The ISF Framework (Le, Tuan Anh)

- <https://dr.ntu.edu.sg/handle/10220/49370>
- <https://osf.io/9udjk>

```
python -m coolisf text "I drink green tea." -f dmrs
```

```
:`I drink green tea.` (len=5)
```

```
-----  
dmrs {  
  10000 [pron<0:1> x ind=+ num=sg pers=1 pt=std];  
  10001 [pronoun_q<0:1> x ind=+ num=sg pers=1 pt=std];  
  10002 [_drink_v_1_rel<2:7> e mood=indicative perf=- prog=- sf=prop tense=pres];  
  10003 [udef_q<8:18> x num=sg pers=3];  
  10004 [_green+tea_n_1_rel<8:18> x num=sg pers=3];  
  0:/H -> 10002;  
  10001:RSTR/H -> 10000;  
  10002:ARG1/NEQ -> 10000;  
  10002:ARG2/NEQ -> 10004;  
  10003:RSTR/H -> 10004;  
}  
# 10002 -> 01170052-v[drink/lelesk]  
# 10004 -> 07935152-n[green tea/lelesk]  
...
```



Questions

- Limitations from a non-native English speaker
- Dynamically Annotated Treebank ... sync layers?
 - sense annotation under constant review
 - treebanking (using FFTB)
- The @dan announcement about lexical entries in ERG?!
- Data formats and Representations
 - MRS/DMRS serializations (and string representations) with extra layer?
 - JSON to RDF of the expanded glosstag
 - The final data format? How to release?
 - WQL/WSI (earlier talk and the senses)
- From English to Portuguese
- Other datasets (propbank frames <http://verbs.colorado.edu/propbank/framesets-english-aliases/find.html>)
 - ERG, Propbank, Wordnet... Unified Verb Index