Multiword expression representation in DELPH-IN: successes and problems

Ann Copestake

Natural Language and Information Processing Group
Computer Laboratory
University of Cambridge

September 2013

Outline.

Outline.

DELPH-IN: Deep Linguistic Processing using HPSG

- Informal collaboration on tools and grammars, 17 groups: see `http://www.delph-in.net/`
- Large grammar for English (ERG), moderately large for German, Japanese, Spanish, Norwegian, Portuguese. Many small grammars.
- Shared technology for parsing etc, common semantic framework.
- Grammar Matrix: framework/starter kit for the development of grammars for languages from all families.
- Multiword expression (MWE) project (Stanford, NTT, Cambridge): funded by NSF and NTT, 2001–2004.

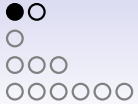
The DELPH-IN English Resource Grammar (ERG) (Flickinger et al)

- Broad-coverage, precise, bidirectional grammar for English, used in a number of projects.
- Approximately 80% - 90% coverage for most domains/genres tried: others tools exist for robustness.
- Parse/realization ranking / extensive treebanks (Redwoods).
- Variety of strategies for adding lexicon automatically.
- Applications with end users: currently English teaching.
- Grammars for other languages developed partially on basis of the ERG (Japan, German, Matrix grammars).

Some ERG design decisions

- ERG development is primarily empirically driven: based on producing parses from corpora of interest (hence good coverage of MWEs that affect parsing results, others less good).
- Fairly detailed compositional semantics (MRS), shallow lexical semantics: only distinguish word senses when they affect analyses.
- Deeper lexical semantics being developed via links to other resources (e.g., WordNet, distributional semantics).

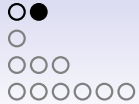
Outline.



Classes of MWE

From Sag et al (2002):

- Fixed expressions: e.g., *by and large*
- Semi-fixed expressions: including complex proper names (e.g., names of sports teams), compound nominals and non-decomposable idioms.
- Syntactically flexible expressions: including verb particle constructions, decomposable idioms and light verb constructions.
- Institutionalized phrases: i.e., phrases which are compositional but statistically idiosyncratic. Idioms of encoding but not idioms of decoding.



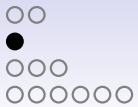
Representation techniques in DELPH-IN

Representation of MWEs in typed feature structures

- Words with spaces: for fixed and (some) semi-fixed expressions.
- Selection for specific lexemes.
- Idioms.

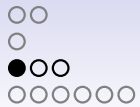
Also:

- Paraphrases via semantic transfer rules.
- Constructions for productive classes: e.g., *by* transport phrases (*by car* etc).
- Fluency ranking for generation: captures some aspects of institutionalized phrases.



Fixed expressions / words with spaces

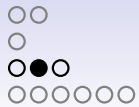
- used when no internal modification: **by and very large* (external modification possible: e.g., *very ad hoc*) and (ideally) individual parts don't relate to other lexemes
- lexically specified so orthography is a list of strings
- parts are combined in parser after tokenization splits them
- mechanism allows for morphological variation
- about 3500 cases in current ERG lexicon



Selection in the lexicon

HPSG and related theories assume lexical selection for classes: e.g., simple transitive selects for NP (via COMPS in ERG): coded in lexicon via types

```
abbreviate_v1 := v_np_le &
[ORTH < "abbreviate" >,
 SYNSEM[LKEYS.KEYREL.PRED "_abbreviate_v_1_rel"]].
```

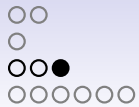


Semantic selection in the lexicon

```
by_temp_p := p_np_i-tmp-vm_le &
  [ ORTH < "by" >,
    SYNSEM [ LKEYS [ -COMPKEY temp_abstr_rel,
                     KEYREL.PRED _by_p_temp_rel ]]].
```

Selection for words via relation:

```
audition_v1 := v_pp*_le &
  [ ORTH < "audition" >,
    SYNSEM [ LKEYS [ -COMPKEY _for_p_rel,
                     KEYREL.PRED "_audition_v_1_rel" ]]].
```

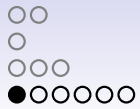


Verb particle entries

Non-compositional verb-particle: particle has ‘dummy’ relation which doesn’t appear in compositional semantics.

```
call_up_v1 := v_p-np_le &
 [ ORTH < "call" >,
   SYNSEM [ LKEYS [ -COMPKEY _up_p_sel_rel,
                    KEYREL.PRED "_call_v_up_rel" ]]].
```

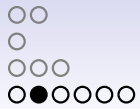
Considerable work on verb particle acquisition: about 1600 verb particle entries in ERG.



Classes of idiom, from Sag et al

1. words not found in other contexts
   by dint of, *tit for tat*
2. syntactically ill-formed
   by and large
   NB: *to lose face* is syntactically regular
3. not decomposable
   kick the bucket, *red herring*
4. decomposable once idiom meaning is known
   let the cat out of the bag, *spill the beans*, *curry favour*
5. transparent (conventional metaphor?)
   cast light on (seeing as understanding), *grease the wheels*

Nunberg, Sag and Wasow (1994), Riehemann (2001)

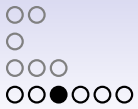


Idioms as compositional

Hypothesis: some speakers attribute meaning to the individual words in decomposable and transparent idioms:

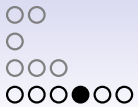
- *spill the beans* corresponds to *reveal the secrets*
- *cat out of the bag* corresponds to *secret out of the hiding place*
- *light at the end of the tunnel* corresponds to *good outcome at the end of the difficult circumstances*
- *curry favour* corresponds to *obsequiously seek support*

That is a cat which has been a very long time coming out of its bag.



Idiomatic lexical signs

- lexical variation: *cast/throw/shed light on*
- recurring uses
  shed light on (help understanding of)
  see the light (come to understanding)
  light dawns (understanding happens)
- mixing idioms
  drop a bombshell (utter something startling)
  drop a brick (utter something stupid)
  These idioms can be mixed:
  Kim is unpredictable: she'll either drop a bombshell or a brick
  conjunction implies the same 'drop' in both idioms.



Intuitive idea of formalisation

- Decomposable idioms are compositional, given the idiomatic meaning-form correspondance
- Idiomatic lexical signs, constrained by idiomatic phrase types to co-occur (possibly with normal signs)
- Specify semantics on idiomatic phrase to get the right idiom pattern:
  curry_v_i(e,u,y), favor_n_i(y)

Idioms in the ERG

Lexical signs:

```
curry_v1_i := v_nb_idm_le &
 [ ORTH < "curry" >,
  SYNSEM [ LKEYS.KEYREL.PRED "_curry_v_i_rel" ]].

favor_n1_i  := n_-_c-brno-ibm_le &
 [ ORTH < "favor" >,
  SYNSEM [ LKEYS.KEYREL.PRED "_favor_n_i_rel" ]].
```

Phrasal constraint:

```
curry+favor := v_nbar_idiom_mtr &
[ INPUT.RELS.LIST <[PRED "_curry_v_i_rel" ],
                   [PRED "_favor_n_i_rel"], ... >].
```

More details

Phrasal constraints:

- Ensure that all the required parts of the idioms are there: need to block e.g., idiomatic curry without (idiomatic) favor.
- Lexical selection not adequate: non-idiomatic words in idioms, non-headed idioms (*cat out of bag*).
- Constraint implemented as a root symbol/start symbol in grammar: all bits of idiom must appear in same sentence.

Paraphrase:

- if idiom decomposable, allows internal modification: *curry Establishment favour* paraphrased as *seek Establishment support*

Also used for determinerless phrases (e.g., *in sequence*) where no idiomatic words.

Outline.

MWEs in DELPH-IN

- Most extensive investigation in ERG, some MWEs in other languages.
- Explicit and implicit MWE representation: some classes of classic MWE aren't MWEs in the grammar (e.g., light verbs). Possibility of identification at semantic level.
- Ambiguity! Some MWE entries removed from ERG because duplicating analyses, incompatible with shallower processing (including many named entities).
- End use is crucial: e.g., idioms of encoding only needed when generating.

Notion of an MWE is (to some extent) context-dependent: irregularity at different levels, productivity is a cline.

Reuse of resources

All DELPH-IN resources are Open Source: available via `www.delph-in.net`

- Adapting techniques to other approaches/representations
- Lexicons (especially ERG) and lexical databases.
- Databases of MWEs
- Redwoods corpora: MWEs extractable.
- MWE bibliography, papers etc:
- ERG demo

Implications for PARSEME

- DELPH-IN representations could be adapted for other frameworks: not typed feature structure dependent.
- DELPH-IN resources available: good coverage for English MWEs with syntactic irregularity.
- Post-processing semantic representations (MRS/DMRS) found to be most plausible approach for idioms and other MWEs that have no syntactic irregularity.