The Lexicon

Miriam Butt
November 2002

What’s in a Lexicon

- What kind of Information should a Lexicon contain?

Semantic: information about lexical meaning and relations (thematic roles, selectional restrictions, hyponomy).

Syntactic: POS, subcategorization frames, co-occurrence restrictions.

Morphological: information about tense/aspect, case, agreement or other syntactically/semantically relevant morphology, information about the morphological form for well-formedness checking.

Phonological: Pronuniciation, perhaps input representation for postlexical phonology.

What’s in a Lexicon

- Not all computational lexicons contain all of this information.
- Most of them concentrate on getting a subset right:

Application Driven

- **ParGram Lexicons:** syntactic and morphological information, no attempt at semantics or phonology.

Lexicon vs. Dictionary

Much work has been done in making dictionaries be *machine-readable* in order to extract computationally useful information from them.

But Dictionaries do not contain enough useful information: further information must be built by hand or by information extraction from corpora or other databases.

How does a Computational Lexicon Differ from a Traditional Dictionary?

Semantic Information

Selectional Restrictions: very difficult to code within a lexical entry. One needs an extra module which encodes world knowledge or something like WordNet.

Is-A Relations (Hyponomy): HPSG lexicons can make use of these via default-inheritance hierarchies, which is part of the standard equipment. But still --- very costly to implement, better to have something like WordNet as an external source of knowledge.

Semantic Information

Precise Lexical Meaning: nobody has yet figured out how to do that --- there are some attempts with lexical decomposition or qualia structures (Pustejovsky, Jackendoff --- see extra handout), but none of this seems really satisfactory.

Thematic Roles: Difficult. Everybody wants them, nobody really knows how to define them well (recent attempt: FrameNet).

Thematic Roles

Typical Thematic Roles: agent, patient/theme, beneficiary/goal/experiencer, instrument, location

They allow a level of abstraction which can potentially make use of nice linguistic generalizations:

- in many languages agents usually end up as subjects, themes usually as objects.
- other languages exactly reverse this pattern: themes usually end up as subjects, agents as objects.

If one is doing MT using thematic roles, that's one problem less to worry about.

Thematic Roles

They allow a level of abstraction which can potentially make use of nice linguistic generalizations:

- in many languages, case marking seems to be sensitive to thematic roles (datives go on goals/experiencers, instrumental case on instrumentals, accusatives on patients, ergatives on agents, etc.)

Example: Urdu Grammar

(see discussion in J&M, Ch. 16, my chapter on Grammatical Relations)

Syntactic Information

POS: minimally, the lexical entry needs to say something about the POS of the word/lemma: N, V, Adj, D, etc.

Subcategorization Frames: the syntactically required arguments of a predicate --- this is related to, but distinct from the thematic roles of a predicate.

Linking Theory

Often the mapping from argument structure (thematic roles) to grammatical relations is one-to-one.

Sometimes it is not.

kill (agent, theme)
Active: kill <SUBJ, OBJ>
Passive: kill <SUBJ>

give(agent, goal, theme)
To-Goal: give <SUBJ, OBJ, OBL>
Dative Shift: give<SUBJ, OBJ, OBJ2>

The farmer killed the duckling.
The duckling was killed.

Sandy gave the book to Kim.
Sandy gave Kim the book.

Linking Theory

The determination of the mapping between thematic roles and grammatical relations is known as *Linking* or *Mapping Theory*.

It would be nice to be able to exploit this mapping computationally, however, the generalizations have proven too fragile (not well understood enough) to be viable in a large-scale implementation.

ParGram Lexicons: only syntactic subcategorization information: SUBJ, OBJ, OBL. No use of thematic roles (exception, the tiny Urdu Grammar).

Syntactic Co-occurrence Restrictions

Not all phrasal co-occurrence restrictions can easily be captured by phrase structure rules alone.

Example: English adverbs

alternatively (etc.)	can occur sentence initially, before a comma (not all can do that)
right (etc.)	can modify a PP (*right after the light*) (not all can do that)
approximately (etc.)	can modify a number (*approximately six*) (not all can do that)

This kind of information must be encoded **lexically.**

Phonological Information

In some languages, you have **focus clitics** which contribute not only semantic information, but also a high tone (e.g., Bengali).

This should arguably be encoded lexically.

o CL TONE = HIGH

Phonetic/Phonological Information: each lexical entry should contain information about the pronunciation of the item (like a dictionary). Most NLP applications are text-oriented and their lexicons not contain such information.

Morphological Information

As much as possible, morphological information should be provided via a seperate morphological analyzer so that the lexicon can consist almost entirely of lemmas.

Output of a typical morphological analyzer:

```
walked:     walk+PastPart
            walk+Past+123SP
```

The information about tense (Past) and agreement (123SP) has a straightforward place in the lexicon and is needed for syntactic analysis. But how about information about the morphological type of the word (PastPart)?

Morphological Wellformedness Checking

English Auxiliaries are wellknown for providing constraints on what kind of a form can follow them.

She has eaten the apple.

She will have eaten the apple.
She may have been eating the apple.

Morphological Wellformedness Checking

In English, this could be done phrase-structurally, but for other languages like German, this is more difficult because the verbal elements may be *scrambled* even though the same kinds of wellformedness restrictions as in English apply.

Sie hat den Apfel gegessen.
Sie wird den Apfel **gegessen haben**.
Gegessen haben wird sie den Apfel.
Gegessen wird sie den Apfel **haben**.

Again, this is something that must be encoded **lexically.**