New Methods in Data-Based Computational Linguistics

Summer Term 2002

IMSLex and CELEX Two Electronic Syntactic Lexica

Frank Henrik Müller

`fhm@sfs.uni-tuebingen.de`

SfS Tübingen,
10th June 2002

IMSLex and CELEX – Overview of the talk

- IMSLex and CELEX, Two Electronic Syntactic Lexica
  - Theoretical Background of Sub-categorization Frames
  - Anatomy of Syntactic Lexica, Structure of Lexical Entries and Contents of IMSLex and CELEX
  - Acquisition of Syntactic Lexica from Corpora
  - Application of Syntactic Lexica in NLP

What is an electronic syntactic lexicon?

An electronic syntactic lexicon contains the selectional restrictions of the lemmas of a language in a strictly formalized way.

- These selectional restrictions can be given for different word categories: e.g.
  - nouns: Abgrenzung (von . . .)
  - adjectives: auffällig (dass . . .)
  - verbs: beabsichtigen (zu . . .)

Because these selectional restrictions further subclassify a syntactic category, this process can be called subcategorization.

Theoretical Background of SC frames

Sub-categorization frames (SC frames) describe the distributional potential of a lemma by listing the types of constituents which a lemma selects.

- The term 'sub-categorize' is used in two ways:
  - The act of classifying lemmas of one category into a more exact category by virtue of their distributional potential.
  - The act (of a lemma) of selecting certain constituents with certain features. e.g. transitive verbs select a direct object

Theoretical Background of SC frames

Originally the term 'sub-categorization' was restricted to verbal complementation but it was extended to cover nouns, adjectives and adverbs.

- verbal subcategorization is dominated by the phenomena of
  - obligatory complementation
  - optional complementation ($\neq$ adjuncts)

Example: Hans isst jeden Tag Brot.

Hans = obligatory complement

jeden Tag = adjunct

Brot = optional complement

Theoretical Background of SC frames

- complements can be described relative to their
  - syntactic function, e.g. subject, object
  - syntactic category, e.g. NP, PP, subclause

Example: 'essen' demands a subject NP and an object NP

* Dass es schmeckt, isst Hans.

Dass es schmeckt, weiß Hans.

- n.b. verbal complements must not be confused with the arguments of a predicate, which refer to a logical relation

Theoretical Background of SC frames

Unlike basic phrases, which are dominated by syntactic restrictions, SC frames are subject to lexical selection, i.e. they depend on the lemma (e.g. the verb).

Example (tags taken from STTS):

```
APPR ART ADV ADJA NN --> PP
NP VVFIN NP PP --> ???
```

- Consequences for NLP:
  - annotation of verbal complementation impossible without lexical information
  - need for a machine-readable lexicon containing (in the ideal case) all relevant lemmas or:
  - training corpus in which all relevant lemmas are annotated or:
  - some kind of back-up strategy

Anatomy of Syntactic Lexica

Syntactic lexica consist of a list of lemmas which are assigned to one or more SC frames.

Besides syntactic annotation, entries in electronic lexica can contain all kinds of (linguistic) information, e.g. orthographical, phonological, morphological, gender, inflectional paradigm.

The format of a syntactic lexicon can be plain text or a markup language (e.g. XML).

The information can be encrypted or verbose.

Example from CELEX:

```
68\abbuchen\3\4\\\N\N\1\5\00N000000;\I\I\0\I\I\I\I\I\I\\\
```

Every slot contains a linguistic feature. There are tools to make the information more verbose.

Anatomy of Syntactic Lexica

Syntactic lexica can differ as regards the degree of expliciteness with which they present the linguistic information they contain:

The advantage of leaving information implicit is that entries are kept smaller; the disadvantage is that the conversion of entries may cause difficulties and/or errors.

- linguistic information can be left implicit if it is extractable by lexical rules, e.g.
  - diatheses like passives
  - conversions like nominalizations

Anatomy of Syntactic Lexica

SC frames can be simplified if they contain predictable information. Lexical rules restore this information.

<u>Example for passives</u>:

Most verbs which have a direct object (i.e. transitive verbs) can form the passive.

Thus: if exists: ‘A bamboozles B’
then exists: ‘B is bamboozled (by A)’

More formal: if exists: bamboozle [sub_NP,obj_NP]
then exists: bamboozle [sub_NP,(by_PP)]

Problem: This rule does not apply to all transitive verbs. (try ‘Die Nachricht freut ihn.’)

‘On the fly generation’ of SC frames might thus prove to be ‘dangerous’.

Structure of Lexical Entries

The field of conversion by lexical rules can be illustrated by the conversion of the entries for ‘essen’ from the original TSNLP (test suite NLP) format to a format suitable for LFG (Lexical Functional Grammar).

Lemma	Subcategorization frame	Example
essen	(sub(NP_nom),obj(NP_acc)	Hans isst Brot
essen	(sub(NP_nom))	Hans isst

SC frame from IMSLex in TSNLP format

Lemma	Subcategorization frame	Example
essen	{ @(NPnom-NPacc essen) | @(null-NPnom-Pass. essen) | @(NPnom essen) @(DISPR) | @(null-Pass. essen) @(DISPR) }	Hans isst Brot Brot wird gegessen Hans isst Heute wird gegessen

LFG mapping of SC frames*

*Both tables from Lezius (2000)

Contents of IMSLex

<u>Format</u>: IMSLex is available in TSNLP format and in XML

<u>SC frames covered for</u>: approx. 6,500 verbs, 2,000 adjectives, 10,700 nouns

<u>Syntactic features of SC frames</u>:

potential of complementation: is expressed by different SC frames

complement function: e.g. subject complement, object complement, . . . (no adverbial complement)

complement realization: e.g. NP, clause (subordinator specified), infinitive, PP (preposition specified), . . .

Contents of CELEX

Format: CELEX is available in a rather encrypted digit/character format but can be converted into a more verbose version containing feature names which resemble the features they represent (e.g. 'CompSubj').

SC frames covered for: 51,728 (9,400 verbs)

Syntactic features of SC frames:

potential of complementation: possible, impossible, obligatory, undetermined

complement function: e.g. subject complement, adverbial complement, . . .

complement realization: e.g. NP, infinitive, PP (no prep specified), . . .

semantic realization for adverbials: e.g. local, temporal, . . .

Acquisition of Syntactic Lexica from Corpora

Manual construction of syntactic lexica is time and money consuming.

Automatic acquisition of SC frames from corpora is possible by different approaches which can be divided into **stochastic** and **symbolic** approaches.

- all approaches rely on pre-structured data
  - POS tag annotation
  - morphological analysis
  - shallow syntatic annotation (i.e. chunks, topological fields and clauses)
- in reference to the entire system, all approaches use hybrid methods

Acquisition of Syntactic Lexica from Corpora

Stochastic approach (Wauschkuhn 1999)

1. clauses containing the target verb are extracted from corpus
2. words are assigned with morpho-syntactic tags (using a lexicon)
3. clauses are assigned with shallow annotation (using hand-crafted rules)
4. an 'argument structure' is extracted which contains all possible complements
   - deleting 'irrelevant' constituents and highly improbable readings
   - abstracting from constituent order
5. grouping of similar structures and extraction of SC frames by means of frequency

Acquisition of Syntactic Lexica from Corpora

Symbolic approach (Eckle-Kohler 1999)

1. POS tagging (stochastically)
2. morphological information (lexicon)
3. shallow parsing (regular grammar)
4. unambiguous realizations of one SC frame are defined (on morphological and distributional grounds)
5. only these realizations are extracted
6. automatically acquired SC frames are manually checked

In order to apply a syntactic lexicon for parsing, rules have to be translated into a format which can be used by the respective formalism.

The process of translation has to be automated in order to apply it whenever the syntactic lexicon has been updated.

Automatization is also needed because heuristics which assign features like preference marks have to be flexibly integrated into the system

- It is possible to apply a syntactic lexicon for parsing using a finite state approach.
  - lexical rules are non-recursive and thus transferrable into a finite state automaton; thus, an explicit version of the syntactic lexicon can be built
  - the transfer of lexical entries themselves is also possible using a finite state automaton
  - prestructured data allows faster and more accurate annotation of syntactic functions