Deverbal Subcategorization Frames
Corpus Evidence Using Human Language Technologies

Tommaso Caselli

Dept. of Chinese and Bilingual Studies
Hong Kong Polytechnic University

Dip. di Linguistica “T. Bolelli”
Università degli Studi di Pisa

NOMICO, September 7th 2012, Nancy
Outline

1. Deverbal Nouns, Argument Structure and Subcategorization Frames
2. Automatic Acquisition of Deverbal Subcategorization Frames
3. Evaluation and Data Analysis
4. Conclusion Future Work
The purpose of this work is two-folded:

- automatically acquire and evaluate subcategorization frames of deverbal nouns in the perspective of developing a (computational) lexicon of nominalizations for Italian similar to the NOMLEX (English);
- collect data in order to provide empirical (quantitative and qualitative) analysis on open issues in the research on nominalizations and related phenomena.
Deverbal Nouns in Italian

We concentrated on deverbal nouns realized by the transpositional suffixes -zione, -mento, -tura and -aggio.

These type of nominals are particularly interesting as they give rise to well known lexical ambiguity phenomenon between eventive and non-eventive reading which has a relevant impact in Theoretical Linguistics and for the development of Natural Language Processing systems.
Deverbal Nouns in Italian

The literature on this topic (Grimshaw 1990, Pustejovsky, 1995, Alexadiou, 2001 among others) points out the selection of specific cues for the identification of the two readings.

<table>
<thead>
<tr>
<th>Features/cues</th>
<th>Event reading</th>
<th>Non-eventive reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obligatory realization of verb argument</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>structure by means of a PPs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pluralization</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Telicity of the verb</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Verb grammatical class</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Type of determiner</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Aspectual modifiers</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Agent-oriented modifiers</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Co-occurrence with eventive predicate</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Complement clause at the infinitive</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td><em>by</em>-phrases, relational adjectives and possessive determiners as realizations of the subject of the deverbal noun</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>
Deverbal Nouns in Italian - Realization of Argument Structure

Grimshaw (1990): compulsory for complex event deverbal nouns, absent in simple event nouns and result nouns;

Pustejovsky (1995); Alexadiou (2001): all event nouns can have arguments and its overt, superficial realization, is not necessary in order to instatiate the event reading;

1. La **fucilazione**\textsubscript{EVENT} della prigioniera\textsubscript{Arg1} (da parte dei soldati\textsubscript{Arg0}).
   
   *The shooting of the prisoner (by the soldiers).*

2. La **fucilazione**\textsubscript{EVENT} ha avuto luogo nella piazza.
   
   *The shooting took place in the square.*
Deverbal Nouns in Italian - Realization of Argument Structure

Corpus study (Russo et al. 2011): 581 deverbal nouns realized by transpositional suffixes, manually annotated from the It-TimeBank-ILC Corpus; 440 event deverbal nouns, 141 non-eventive deverbal nouns.

<table>
<thead>
<tr>
<th>Noun type</th>
<th>Possessive modifiers</th>
<th>PPs</th>
<th>Di / Del</th>
</tr>
</thead>
<tbody>
<tr>
<td>eventive deverbal nouns</td>
<td>0.8%</td>
<td>47%</td>
<td>40%</td>
</tr>
<tr>
<td>non eventive deverbal nouns</td>
<td>2.5%</td>
<td>28%</td>
<td>22%</td>
</tr>
</tbody>
</table>

Argument structure is a relevant cue for the identification event reading of deverbal nouns. However, for the automatic acquisition of the argument structure of deverbal nouns, the identification of the argument status of a phrase is strictly related to the sense disambiguation of the deverbal noun.
As a solution, we have to change perspective and investigate the argument structure in terms of subcategorization frames.

Verb subcategorization:

"The potential of verbs to choose their complements is referred to as ‘verb subcategorisation’ or ‘verb valency’, and a combination of functional complements that are evoked by a verb is often called a ‘verb subcategorisation frame’, or simply a ‘verb frame’. " [Shulte im Walde, 2009]

complement = it subsumes both arguments and adjuncts.
Key role is played by the notion of complement.

Following Bierwisch (2009), we consider that all lexical entries exhibit a categorization, and that they can be associated with a pertinent subcategorization.

Noun subcategorization:

*The potential of nouns to choose their complements is referred to as ‘noun subcategorisation’ or ‘noun valency’, and a combination of functional complements that are evoked by a noun is called a ‘noun subcategorisation frame’, or simply a ‘noun frame’.*
Key role is played by the notion of complement.

Following Bierwisch (2009), we consider that all lexical entries exhibit a categorization, and that they can be associated with a pertinent subcategorization.

The distinction of the grammatical roles of the complements is accomplished at a further stage by taking into account semantic restrictions (selectional preferences) on the lexical fillers of the subcategorization frames.
Acquisition of Subcategorization Frames

a. Corpus selection and preparation;
b. Frame types;
c. Acquisition method
d. Filtering
e. Evaluation
Corpus selection and preparation:

La Repubblica Corpus (Baroni et al., 2004); 300 million tokens of newspaper articles.

Corpus has been automatically parsed with a stochastic dependency parser: DeSR (Attardi and Dell’Orletta, 2009). Accuracy: 84.56 on Italian data (CoNLL 2007).
Frame types:

NO predefined set of subcategorization frames is assumed. The Dev-SCF Extractor "collects" all phrases which have one of the following dependency relations with the noun(s) in analysis:

- \( \text{arg} \): argument;
- \( \text{comp} \): complement (it subsumes both arguments and adjuncts)
Acquisition of Subcategorization Frames

Acquisition method:

The Dev-SCF Extractor is an unsupervised system. It takes in input parsed data, and it is composed by three modules:

- a pattern extractor, which identifies possible SCF patterns for deverbal nouns;
- an SCF builder which assigns a list of eligible SCFs to the noun in analysis; and
- a filter which removes SCFs considered incorrect.
Acquisition of Subcategorization Frames

Filtering Evaluation:

Filtering: statistical measure - the maximum likelihood estimate (MLE) (relative frequency);

Evaluation: we developed an ad hoc lexical suite of 30 event nominalizations extracted from the It-TimeBank ILC corpus. The lexical suite has been obtained from: a.) a manually created lexicon, the PAROLE Lexicon; b.) an automatically created lexical resource, LexIt, and c.) manual exploration of corpus data (150 random contexts per deverbal noun from the La Repubblica corpus).

Noun selection: frequency and presence/absence of the event/non-event reading.

225 couples of nominalization-SCFs, for a total of 32 unique SCF types.
SCF: $COMP-DI

$COMP-DI[MOTIVAZIONE=2, ROTTAMAZIONE=1,
DIFFERENZA=3, VALIDITÀ=2, POSSIBILITÀ=2,
DEFICIENZA=1, DELITTO=1, VOLUME=1, INDAGINE=1,
COMUNIONE=1, MINISTERO=4, PARTICOLARE=1,
MORTE=49, MAGISTRATO=2, ATTITUDINE=1, 1988=1,
CALO=1, ...]

MOST FREQUENT FILLERS:
VERITÀ=350, RESPONSABILITÀ=177, FATTO=109,
MORTE=49, REDDITO=41, VALORE=24, VIOLAZIONE=23,
INVALIDITÀ=21, CONDIZIONE=20, REATO=18, IMPOSTA=18,
UFFICIO=17, REQUISITO=15,
The identification of the best MLE threshold has been conducted empirically. The results have been computed in terms of (type-)precision, (type-)recall and f-measure. A total of 209 couples of nominalization-SCF frame has been extracted.

<table>
<thead>
<tr>
<th>MLE threshold</th>
<th>Dev-Extractor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
</tr>
<tr>
<td>unfiltered</td>
<td>.094</td>
</tr>
<tr>
<td>MLE 0.008</td>
<td>.921</td>
</tr>
<tr>
<td>MLE 0.005</td>
<td>.80</td>
</tr>
<tr>
<td>MLE 0.003</td>
<td>.655</td>
</tr>
<tr>
<td>MLE 0.001</td>
<td>.406</td>
</tr>
</tbody>
</table>
Data Analysis

The results for the 5 most frequent and the 5 less frequent deverbal nouns in the gold standard show that there is a little (if not null) correlation between the size of the input data and the acquisition, contrarily to what happens for verb frames.
## Data Analysis

<table>
<thead>
<tr>
<th>Lemma</th>
<th>Freq</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPERAZIONE  [operation]</td>
<td>71343</td>
<td>.533</td>
</tr>
<tr>
<td>MOVIMENTO  [movement]</td>
<td>40498</td>
<td>1</td>
</tr>
<tr>
<td>DIREZIONE  [direction]</td>
<td>38190</td>
<td>.6</td>
</tr>
<tr>
<td>PROVVEDIMENTO  [action]</td>
<td>31978</td>
<td>.75</td>
</tr>
<tr>
<td>RIFERIMENTO  [reference]</td>
<td>31599</td>
<td>.545</td>
</tr>
<tr>
<td>ACCERATAMENTO  [assessment]</td>
<td>3634</td>
<td>.705</td>
</tr>
<tr>
<td>STABILIMENTO  [establishment]</td>
<td>8191</td>
<td>9.41</td>
</tr>
<tr>
<td>ACQUISIZIONE  [acquisition]</td>
<td>8712</td>
<td>.666</td>
</tr>
<tr>
<td>COLLOCAMENTO  [collocation]</td>
<td>9164</td>
<td>.666</td>
</tr>
<tr>
<td>INTRODUZIONE  [introduction]</td>
<td>9195</td>
<td>.842</td>
</tr>
</tbody>
</table>
The highest source of errors is represented by SCFs containing $\text{COMP-DAPARTEDI}$ (realization of the SUBJ role; by-phrases in English).

26/56 false negatives

$\text{COMP-DAPARTEDI}$ has low frequency - excluded by the filter
Subcategorization frame inheritance (Lapishinova-Koltunski, 2009; Schierholz, 2001)

$\textit{COMP-DI}$ have been excluded from the analysis.

On those nouns which present a logical polysemy between eventive vs. non-eventive reading, 61 SCF tokens are directly inherited from the corresponding verbs (29.18%) on "indirect complementation" (e.g. $\textit{COMP-IN}; \textit{COMP-CON}$).
13 deverbal SCF frames do not occur in the extracted verb frames (low frequency), but can be promoted to verb frames.

17 SCFs realized by prepositional phrases out of 42 false positives which have the same surface realization and same complement type as the frames of the corresponding verbs: TRUE POSITIVE

All inherited frames belong to verbs with a passive diathesis or inaccusative-ergative.
Data Analysis - $\text{COMP-DI (of-phrases)}$

Preliminary analysis (transitive verbs).
$\text{COMP-DI (of-phrases)}$ = it may realize arguments (i.e. subject/direct object) or be a noun modifier.
The different role of the $\text{COMP-DI}$ facilitates the identification of the eventive vs. non eventive reading of deverbal nouns.

Hypothesis: if $\text{COMP-DI}$ shares the same semantic type of the $\text{OBJ}$, then the noun is likely to be an argument and have an event reading.

Manual exploration on semantic type assignment to the most frequent noun fillers of the verb and corresponding deverbal noun.
11 deverbal nouns:

7 cases of full inheritance - $COMP-DI is an argument
4 cases of partial inheritance - $COMP-DI is an argument only with a subset of the semantic types

<table>
<thead>
<tr>
<th>Verb</th>
<th>Sem. Type SOBJ filler</th>
<th>Deverbal Noun</th>
<th>Sem. Type $COMP-DI</th>
<th>Inheritance</th>
</tr>
</thead>
<tbody>
<tr>
<td>APRIRE</td>
<td>artifact; body; location; event; state</td>
<td>APERTURA</td>
<td>artifact; event; state; location</td>
<td>full</td>
</tr>
<tr>
<td>INTRODURRE</td>
<td>artifact; person; communication</td>
<td>INTRODUZIONE</td>
<td>artifact; person communication;</td>
<td>artifact</td>
</tr>
<tr>
<td>COPRIRE</td>
<td>artifact; body; object; event; possession; location;</td>
<td>COPERTURA</td>
<td>artifact; event, possession; object</td>
<td>possession; object; event</td>
</tr>
</tbody>
</table>
Development of a system for the automatic acquisition of deverbal SCFs;

Low frequent frames has to be validated by manual post-processing;

SCF inheritance: we have identified a preference of inheritance for verbal SCF realized by PP phrases;

Deverbal SCF seems to have a Zipfian distribution (see also Preiss et al., 2007);
Future Work

- Further analysis of the correlation between $OBJ$ (and $SUBJ$) and $COMP-DI$ ($of$-phrases) complement to facilitate the identification of event vs. non-event reading;
- Identification and assignment of the thematic roles to the complement to facilitate the distinction between argument and adjuncts;
- Refinement of the extractor on the basis of the post-processing;
- Distribution of the lexicon.
Acknowledgments

Thanks to EU Erasmus Mundus Action 2 program MULTI (grant agreement n. 2009-5259-5)
Thank You!