Corpus Evidence and Compound Structure: The Case of Italian NN Compounds

Marco Baroni[1] Emiliano Guevara[1]
Vito Pirrelli[2] Eros Zanchetta[1]

[1]University of Bologna
Bologna, Italy

[2]Istituto di Linguistica Computazionale
Pisa, Italy

QITL-2, 2/6/2006

Outline

NN compounds in Italian

- They exist
- Obviously, more limited than English/Germanic compounds
- *scimmia bottiglia* can only mean “monkey that has some properties of a bottle, that is of bottle-type (e.g., shaped like a bottle)”
- Cf. *bottle monkey*: monkey shaped like a bottle, monkey that uses bottles to play, monkey sold in bottles, monkey that lives in bottles, monkey that is near a bottle in this moment...
- Stronger constraints might help us uncovering generalizations more easily than from all-purpose Germanic compounding
- Left-headed (syntactic and semantic evidence)

Outline

Candidate NN contexts

Extracted from ∼1.9 billion word Web-derived corpus

```
[pos="DET:.*|ART|NUM|ADJ|PRE|ARTPRE|CHE|CON|WH"]
[pos="NOUN"] [pos="NOUN"]
[pos!="NOUN|NPR|VER.*|AUX:.*|ADJ|PRE|ARTPRE|CLI|CHE|CON|WH"]

[pos="VER.*" & pos!="VER:ppa.*"] [pos="ADV.*|NEG"]{0,3}
[pos="NOUN"] [pos="NOUN"]
[pos!="NOUN|NPR|VER.*|AUX:.*|ADJ|PRE|ARTPRE|CLI|CHE|CON|WH"]

[pos!="NOUN|NPR"]
[pos="NOUN"] [pos="NOUN"]
[pos="ADV.*|NEG"]{0,3}
[pos="VER.*|AUX:.*|ADJ|PRE|ARTPRE|CLI|CHE|CON|WH"]
```


Sampling from 4 frequency ranges

range	types	sample
1	699,659	300
2-5	329,270	300
6-3000	113,147	300
>3000	109	109
all	1,142,185	1,009

Classification of compounds

- Manual filtering leaves us with 252 true compounds (about 1/4 of sample)
- 4 main types emerge (named after function of modifier):
  - Coordinative (COOR): 34 (13.49%)
  - Attributive (ATTR): 41 (16.27%)
  - Argumental (ARGU): 51 (20.24%)
  - Grounding (GROU): 118 (46.82%)
- Residual of 8 compounds (3.17%) that will require further analysis
- Similar to Scalise and Bisetto (2005)

Outline

Coordinatives

- Head and modifier denote similar or compatible entities, and compound has coordinative interpretation (the referent of HM is both H and M)
- E.g., *viaggio spedizione* “trip-expedition” *mago alchimista* “magician-alchemist”, *bambino autista* “child-driver”

Attributives

- Interpretation of M is reduced to some iconic properties of its full semantic representation, and these properties are *attributed* to H
- E.g., *progetto pilota* “pilot project”, *presidente fantoccio* “puppet president”, *brano cardine* “pivot track”
- Often attributive modifiers display adjective-like behavior: *pilota* “pilot” occurs in post-N position about 1/4 of the times (13727/52641); *fantoccio* “puppet” occurs in post-N position about 1/3 of the times (722/2314)

Argumental compounds

- Heads typically deverbal nominalizations, or other nouns able to project “verb-like” arguments; modifier is internal argument of corresponding verb
- I.e., objects of transitives: *protezione persone* “people protection”, *raccolta fondi* “fund collection”, *gestione priorità* “priority management”; subjects of unaccusatives: *arrivo documenti* “document arrival”, *caduta massi* “stone fall”
- Modifier can never be subject/agent of transitive verb – *controllo Senato* is only attested/acceptable as “control exerted *over* the Senate”, not as “control exerted *by* the Senate”

Grounding compounds

- Head does not have verb-like argument structure, but general meaning that needs to be contextualized/specialized by modifier (modifier “grounds” meaning of head)
- Typical grounding heads: containers, aggregators/ions, (information) carriers, pointers, measurable properties, locations
- E.g., *stanza server* “server room”, *associazione ambientalisti* “environmentalists’ association”, *fondo pensioni* “pension fund”, *centro città* “city center”, *altezza righe* “line height”, *posto auto* “car place (parking space)”

Outline

Proportion of heads occurring in more than one compound

- ATTR: 0 (0%)
- COOR: 1 (3%)
- ARGU: 6 (14%)
- GROU 15 (18.5%)

Proportion of modifiers occurring in more than one compound

- ATTR: 4 (12%)
- COOR: 1 (3%)
- ARGU: 2 (4%)
- GROU 10 (9.71%)

Number of modifier

- When head is singular:

cat	s only	p only	both
ATTR	27 (77.14%)	0 (0.00%)	8 (22.86%)
COOR	20 (83.33%)	0 (0.00%)	4 (16.67%)
ARGU	18 (37.50%)	17 (35.42%)	13 (27.08%)
GROU	56 (50.00%)	24 (21.43%)	32 (28.57%)

- When head is plural:

cat	s only	p only	both
ATTR	18 (72.00%)	1 (4.00%)	6 (24.00%)
COOR	0 (0.00%)	18 (72.00%)	7 (28.00%)
ARGU	4 (30.77%)	6 (46.15%)	3 (23.08%)
GROU	26 (50.98%)	13 (25.49%)	12 (23.53%)

Outline

Connector patterns

- E.g., *server room*/*room with servers*
- Not a new idea: Levi (1978), Lauer (1995)...

Automated extraction and selection of connector patterns

- Conjunctions: *e* "and", *o* "or", ...
- Prepositional patterns:
  - *di* "of"
  - Pooled "contentful" prepositions: *a* "∼to", *in* "in", *per* "for", *con* "with", *su* "on/about"
- Past part + prep patterns: *dedicato a* "dedicated to", *basato su* "based on", *destinato a* "intended for/destined to", *finalizzato a* "aimed at"

Attraction

- Frequency of H+M pair with a connector pattern must be weighted by unigram frequency of head and mod (connector patterns and source corpora are kept constant across compared items)
- Attraction score:

$$rank \frac{count(\mathrm{H\ CONN\ M})}{count(\mathrm{H}) + count(\mathrm{M})}$$

- Dice-coefficient-like score; similar results with MI-like score
- Rank computed across classes, over all pairs with $count(\mathrm{H\ CONN\ M}) > 0$

Conjunctions

Distribution of attraction of H-M pairs to pattern

type	min	1st qu.	med	mean	3d qu.	max
COOR	16.00	106.50	134.00	117.40	153.00	162.00
ATTR	5.00	49.25	81.00	79.25	102.30	146.00
ARGU	3.00	32.25	78.00	77.25	118.80	163.00
GROU	1.00	38.50	71.50	74.78	112.30	159.00

di (del) “of”

Distribution of attraction of H-M pairs to pattern

type	min	1st qu.	med	mean	3d qu.	max
ARGU	2.00	66.00	106.00	105.60	154.00	187.00
GROU	8.00	62.50	110.00	104.00	143.50	185.00
ATTR	1.00	21.00	43.50	58.42	78.75	170.00
COOR	5.00	26.00	43.00	45.23	65.00	85.00

Contentful prepositions

Distribution of attraction of H-M pairs to pattern

type	min	1st qu.	med	mean	3d qu.	max
GROU	1.00	51.50	93.50	89.75	130.30	160.00
ATTR	12.00	52.00	80.50	83.36	112.80	147.00
COOR	9.00	58.50	76.50	68.63	81.75	118.00
ARGU	2.00	24.50	51.00	61.56	97.50	161.00

Past-participle+preposition patterns

- *basato su* “based on/upon”: 3 ARGU, 4 GROU
- *dedicato a* “dedicated to”: 1 ARGU, 8 GROU
- *destinato a* “intended for/destined to”: 8 GROU
- *finalizzato a* “aimed to”: 2 ARGU, 4 GROU

Outline

Summary

- Qualitative analysis of corpus-mined Italian NN compounds suggests 4-way classification: ATTR, COO, ARGU, GROU
- Distributional evidence (head and modifier repetition, modifier number, connector patterns) supports 4-way distinction

Future work

- Use (more) cues in supervised/unsupervised machine learning tasks
- Experimental evidence
- Extend analysis to other ways to express relationship between nouns

References

M. Baroni and M. Ueyama (2006). Building general- and special-purpose corpora by Web crawling. *Proc. NIJL Workshop*.
A. Bisetto and S. Scalise (2005). The classification of compounds. *Lingue e linguaggio* **4**, 319-332.
R. Jackendoff (2002). *Foundations of language*. OUP, Oxford.
M. Johnston and F. Busa (1996). Qualia structure and the compositional interpretation of compounds. *Proceedings of ACL SIGLEX 1996*.
M. Lauer (1995). *Designing statistical language learners: Experiments on noun compounds*. PhD thesis, Macquarie University, Sydney.
R. Lees (1960). *The grammar of English nominalizations*. Indiana University Press, Bloomington.
J. Levi (1978). *The syntax and semantics of complex nominals*. New York: Academic Press.
B. Rosario, M. Hearst and Ch. Fillmore (2002). The descent of hierarchy, and selection in relational semantics. *Proceedings of ACL 2002*.
L. Vanderwende (1994). Algorithm for automatic interpretation of noun compounds. *Proceedings of COLING 1994*, 782-788.