Syntactic patterns and their lexical fillers: Extraction from parsed data and representation in a pre-dictionary data collection

Ulrich Heid, Università di Hildesheim,
Linguistica Computazionale

Modena, Workshop PhrasLex, 19/20-9-2019

Overview

- Towards combining syntax and collocation in a dictionary
  - Examples from specialized dictionaries
  - Examples from learners' dictionaries
- Combining syntax and collocation as a task for corpus-based lexicography
  - Sample data from Italian
  - Existing approaches
  - Two mini-experiments
- Consequences for lexicography
  - Pre-dictionary data collection
  - Lexicographic presentation
- Conclusions

Towards combining syntax and collocation in a dictionary

- Objective:
  - Cover syntactic (valency) properties of (verbal) predicates
  - Include relevant lexical preferences
- Syntactic dictionaries:
  - Mainly focused on valency patterns
  - Examples sometimes used to show typical combinatorics
- Collocation dictionaries:
  - Mainly focused on word pairs
  - Syntax often represented in an abstract way

$\Rightarrow$ Need to bring the two strands together

$\Rightarrow$ "Ideal" solution for learners (?)

Valency dictionaries

Blumenthal/Rovere 1998 as an example

- s.v. **trovare**
  18. N-si V-Agg pred
  sein
  - ◇ Bsp. **1.** Il magistrato si **trova** coinvolto in tutta una serie di attività [...] (Sole) **2** {...} **3.** Vi siete chiesti a questo punto, perché qualcuno si **trova** costretto ad abbandonare il cane? (Sole) ... sich gezwungen sieht ...

{...}: my omissions

Collocation dictionaries

Tiberii 2012 as an example

- Mainly follows the OCDSE model:
  Syntactic constructions indicated by formulae:
  - AGGETTIVI
  - VERBO + COMPLEMENTO
  - AVVERBI
- Example:
  s.v. **ridurre**
  ridurre v *portare a dimensioni minori*
  AVVERBI al minimo, a zero, considerevolmente, costantemente, drasticamente, {...}, sensibilmente, significativamente, {...}

{...}: my omissions

Combining valency and collocation

Lexicographic approaches

- Learner Lexicography:
  - Aware of interaction: Bartsch 2003
    Collocations as preferred lexical "fillers" of patterns
  - But few recommendations so far
    for lexicographic presentation
    * Traditional view: binary word combinations, Hausmann 2004
      base and collocate Mel'čuk 2003, ...
    * Collocation combinations: Zinsmeister/Heid 2003
      X übt {heftige, scharfe, massive, harsche} Kritik
  - Most recently:
    Collocations at the centre of the syntax-lexicon continuum
    ⇒ Longer chains: collo-constructions Herbst 2018

Combining valency and collocation

An example of collo-constructions

data discussed by Th. Herbst

- Valency patterns of EN [*to*] *earn*
  1. Y earns **sth.**
  2. X earns Y **sth.**

  $\Rightarrow$ Focus on lexical realization of direct object

- Pattern (1):
  - sbdy earns n pounds
  - sbdy earns {money, profits, interest, ...}
  - sbdy earns {salary, wages, revenue, ...}
  - sbdy earns {a, his, her, ...} living

- Pattern (2):
  - sth earns Y respect
  - sth earns Y {reputation, fame, recognition, award, ...}
  - sth earns Y the {title, nickname, sobriquet, epithet, ...} NOUN

Combining valency and collocation

A relevant task for learner lexicography

- Data suggest a need for detailed lexicographic description and presentation
- Likely not enough space in small learners' dictionaries:
  - Cornelsen 2013, s.v. *trovare*
    6 trovarsi ≈ *sein* sich befinden:
    Marco si trova all'estero al momento.
    Marco ist zur Zeit im Ausland. {...}
  - Hueber 2006, *Italienisch ganz leicht – Wörterbuch,*
    s.v. *trovare*
    {...} **II.** *vr* **-rsi 1.** (*essere*) sich befinden **2.** (sentirsi) sich fühlen {...}

Combining valency and collocation

An issue for computational lexicography

- Data provision:
  - Large corpora, e.g. Web as Corpus data, news archives, ...
  - Our mini-experiments:
    based on the PAISÀ corpus (223M words)
- Data pre-analysis: standard techniques
  - Tokenizing — words – sentences – ...
  - POS-Tagging — word class labels
  - Lemmatization — lemmas for word forms
  - Parsing — syntactic analysis: grammatical functions

PAISÁ corpus data: an example

word	lemma	POS	dependecy relation
Bisogna	bisognare	V/verb	ROOT
ridurre	ridurre	V/verb	arg
drasticamente	drastico	B/adverb	mod
l'	il	R/determiner	det
impiego	impiego	S/noun	obj
dei	di	E/preposition	comp
pesticidi	pesticida	S/noun	prep
,	,	F/punctuation	con
e	e	C/conjunction	con
cercare	cercare	V/verb	conj
le	il	R/determiner	det
sostanze	sostanza	S/noun	obj
meno	meno	B/adverb	mod
dannose	dannoso	A/adjective	mod
.	.	F/punctuation	

Extracting data for valency and collocations

Elements of current methods: Valency

- Basis: Theoretically inspired concept of valency frames:
  Subject + Object, P-Objects, Predicatives, etc.
- Assumptions about possibly relevant adjuncts,
  e.g. adverb(ial)s ridurre drasticamente
- Types of approaches
  - Interactive small-scale experiments
  - Verb frame induction: search for frames of a given verb,
    possibly extended by the identification
    of semantic features of complements, or semantic roles

    Schulte im Walde 2009

Extracting data for valency and collocation

Elements of current methods: Collocations (1/2)

- Tasks:
  - Identifying significant word pairs
  - Identifying collo-constructions:
    How many and which components belong together?
- Basis:
  - So far mostly association measures,
    applied to word pairs within dependency relations
    Evert 2004
  - Concepts for extending the approach
    to larger word combinations

Extracting data for valency and collocation

Elements of current methods: Collocations and Collo-Constructions (2/2)

- Approaches:
  - Incremental extension of n-grams — Da Silva et al. 1999
  - Generalizing expected frequencies and association measures — Lin 1999; Zinsmeister/Heid 2003
  - Hypothesis tests in n-dimensional contingency tables — Blaheta/Johnson 2001
  - Various heuristics, e.g. c-value and NC-value — Frantzi et al. 2000

Extracting data for valency and collocation

Simplistic extraction patterns (1/3)

- ridurre + INTENSIFIER
  ⇒ are all of these exchangeable?

drasticamente	287
notevolmente	259
...	
sensibilmente	110
significativamente	71
considerevolmente	53
...	

- Observation

PAISÀ, 1466 instances of *ridurre + -mente*

  - ridurre {drasticamente, sensibilmente, notevolmente} il tempo di ...
  - ridurre {drasticamente, sensibilmente, notevolmente} il costo di ...
  - ridurre {notevolmente} le possibilità di ...
    - *possibilità:* very few examples with the other adverbs

Extracting data for valency and collocation

Simplistic extraction patterns (2/3)

- trovarsi + ADJ + {a|ad|di} + INFINITIVE
- Simple query:

```
[pos="P"]                    | pronoun: si, ci, mi, ...
[lemma = "essere"]?          | tense auxiliary: optional
[lemma="trovare"]            | 'trovare': litteral
[pos!="F|S"]{0,5}            | up to 5 words, no nouns
@[deprel= "pred"]            | element with PREDicative role
"a|ad"                       | preposition: litteral
[pos = "V"];                 | verb(al infinitive)
```

⇒ Approximation of the typical contexts

Extracting data for valency and collocation

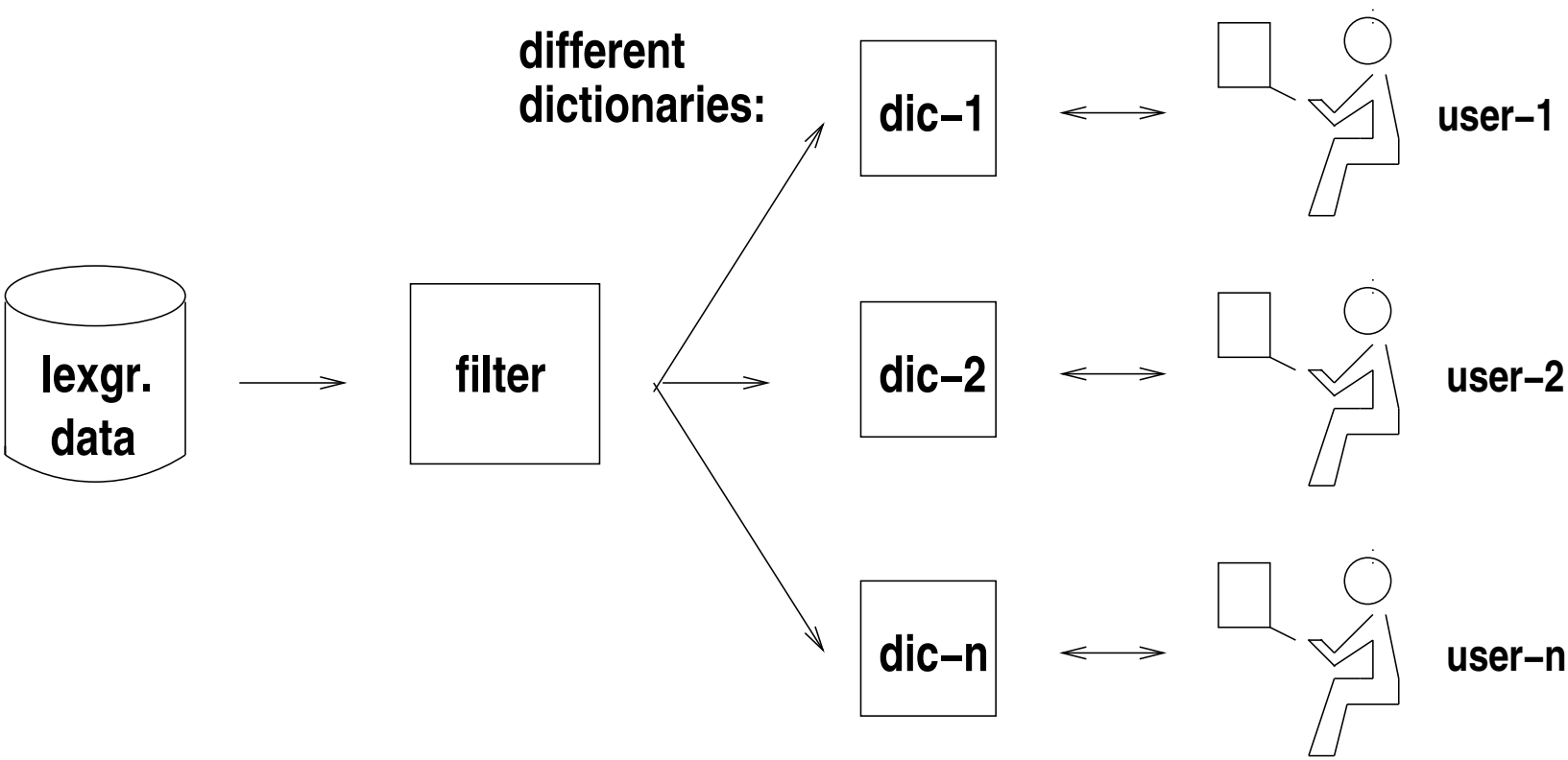
Simplistic extraction patterns (3/3)

- Result data for *trovarsi* + *ADJ* + {*a*|*ad*|*di*} + *INFINITIVE*
- Typical adjectives:
  - costretto a 55
  - pronto a 5
  - (in)capace di 9
  - obbligato a 4
  - libero di 3
  - less than 10 others, with very few occurrences: denoting possibility/impossibility to act
    ⇒ Collocation? Collo-Construction?
    ⇒ Artifact of the kind of corpus analyzed?

Consequences for lexicography

Pre-dictionary data collection

- Collecting examples extracted from the corpus,
  as raw material for subsequent lexicographic description



Consequences for lexicography

The shape of a pre-dictionary data collection

- Should be table-like:
  - Based on dependency relations
  - Should contain occurrence numbers
    for combinations of different length:
    pairs, but also longer combinations
  - Should keep track of relevant properties of the items,
    e.g.:
    * Surface order (where relevant)
    * Morphosyntactic properties, e.g. singular $\leftrightarrow$ plural,
      use of determiners, etc.

Consequences for lexicography

Presenting the data in a(n online) dictionary to the user

- Lexicographic Function Theory: Tarp 2008
  - Linguistic classification of data: less relevant for users
  - Corpus sources: text types, genres, ...
  - Source (and reliability) of data:
    Users must be informed when indications
    are the result of computational analysis
    without inspection by the lexicographer
- Quantitative data may contribute
  to the lexicographic classification as collo-constructions

Conclusions

- There is a need for dictionaries
  that combine valency and collocational description
- Data extraction from parsed corpora
  can provide useful raw material for lexicographic description
  - In many cases, existing valency descriptions
    can serve as a starting-point for corpus analysis
  - We need methods for the statistical analysis
    of longer combinations
- Pre-dictionary data collections should
  - be table-like, based on dependency relations
  - contain detailed metadata on sources and discovery processes