UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

PHRASALEX

Phraseological approaches to learner's lexicography
www.sensepatterndictionaries.net

ACTIVE Sense Pattern Dictionaries

Joining the quest for John Sinclair's ultimate dictionary

Laura Giacomini

Paolo V. DiMuccio-Failla

Adriana Orlandi

Universität Hildesheim

Inst. für Informationswissenschaft und Sprachtechnologie

INSTITUT FÜR ÜBERSETZEN UND DOLMETSCHEN

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Modena, 19-20 settembre 2019

Overview of the first part of the presentation

I. Theoretical background
II. Usage pattern theory with focus on learner's lexicography
III. A model of a *sense-disambiguating* dictionary based on word sense patterns

I. Theoretical background

§I.1 Sinclair's Hypothesis (about lexical units)

- In our work, we presuppose the validity of (a slightly weakened version of) *Sinclair's Hypothesis* (SH), stating that, in general, ***lexical meaning*** is not a feature of single words in isolation, but of words in their various distinct ***patterns of* (*normal*) *usage*** (Sinclair 1991), determined by their *colligation*, *collocation*, *semantic preference* (and *semantic prosody*).

1. »When you **put** something in a particular place or position, you move it into that place or position«
2. »If you **put** someone ... [in a particular place or position], you cause them to go there and to stay there for a period of time«
3. »To **put** someone or something in a particular state or situation means to cause them to be in that state or situation«.

- COBUILD dictionary

(cf. Also Moon 1987: 91)

§I.2 Hanks's *Theory of Norms and Exploitations*

- P. Hanks simplified and formalized Sinclair-patterns for applications in NLP (and also in language teaching). For example, the sense patterns proposed by Hanks in his *Pattern Dictionary of English Verbs* (PDEV) are syntagmatic patterns consisting of an **argument structure** assigned together with the most general **semantic types** (and possibly semantic **roles**) to which the arguments of a verb ***normally*** refer (cf. Hanks 2013).

- Let us look for example at the syntagmatic patterns of the verb *lead* according to Hanks's PDEV (simplified):

1. Pattern: [[Eventuality]]$_1$ **leads** to [[Eventuality]]$_2$
→ [[Eventuality]]$_1$ is the cause of [Eventuality]]$_2$

2. Pattern*:* [[Eventuality]]$_1$ **leads** up to [[Eventuality]]$_2$
→ [[Eventuality]]$_1$ precedes [[Eventuality]]$_2$

3. Pattern: [[Eventuality]] **leads** [[Human]]/[[Institution]] to…
→ [[Eventuality]] causes or triggers [[Human]]/[[Institution]] to…

4. Pattern*:* [[Human]]/[[Institution]]$_1$ **leads** [[Human group]]/[[Institution]]$_2$
→ [[Human]]/[[Institution]]$_1$ organizes or directs activity of [[Human group]]/[[Institution]]$_2$

- A semantic role is used in the following pattern for *abdicate*:

[[Person=Monarch]]$_1$ **abdicate** (in favor of [[Person=Monarch]]$_2$)

_____

(According to CPA conventions (cf. Hanks, 2004: 93), double square brackets indicate semantic types and curly brackets (braces) indicate sets of specific lexical items. The keyword is written in bold letters.)

- Identifying the right semantic types as selectional preferences, in particular not leaving out normal usage on one side and not generalizing into abnormal usage on the other side, requires **linguistic and ontological expertise**.

- P. Hanks and E. Jezek (among others) notice in fact that semantic types in general do not map neatly onto empirically well-founded semantic preferences.

- However, the question whether a better ontology can be conceived for similar purposes remains open.

(§I.3) Sense pattern dictionaries as *active* dictionaries

- Traditional dictionaries prioritize **completeness** over **normality**, giving, for every word, all its meanings in an imagined ideal corpus. Learner's dictionaries are no exception in this respect.
- PRO: A learner is therefore given the means to **understand** all possible meanings of a word (in normal daily usage) when hearing/reading it.
- CON: On the flip side, a learner cannot acquire the ability to **produce** that same word in all contexts and situations a first-language speaker would.

- The COBUILD dictionary is one of the few exceptions to this rule. Sinclair-patterns do indeed give learners the means to master active word usage, while the macrostructure is the usual one.

- So why has the COBUILD dictionary not become the gold standard for modern (learner's *and* general) lexicography? T. Herbst and other scholars have pointed out that Sinclair-patterns tend to be long-winded and repetitious (cf. Heuberger 2016; Herbst 1991: 1382), while others have criticized the sense ordering criteria (Lew 2013: 7) of the dictionary.

II. Usage pattern theory with focus on learner's lexicography

§II.1 Our present inquiry

- We are currently investigating the possibility of:

  1) devising word sense patterns which are easily readable and yet formalizable, for linguistic rigor and possible applications to NLP;

  2) finding **semantic types** better suited for our purposes;

  3) adding ***semantic properties and conditions*** to the semantic types and roles of Hanks's patterns, in an attempt to pin down the exact semantic restrictions of word meanings;

  4) extending SH to ***word sense clusters*** (see further down).

§II.2 Using natural language ontologies

- Every natural language is committed to a naive ontology (cf. Moltmann 2016). Its entities are not just the semantic values of its referential terms (mainly nouns and noun phrases), but also the implicit arguments of its predicates (semantic restrictions).
- Notice that it is only presuppositions, not assertions, that reflect the ontology implicit in a natural language.

- Only **WordNet** and **EuroWordNet**, as formal ontologies, are linguistic in nature, but, as noticed by P. Hanks and E. Jezek (Jezek & Hanks 2010), they cannot be considered "truly" linguistic, since, while describing a hierarchy of concepts, they **do not account for combinatorial constraints** on lexical items.

- We currently hypothesize that a **true linguistic ontology** should indeed be constructed the other way around: the right semantic types, roles, properties, and conditions should be **found studying the semantic preferences of words**. Let us see how this could be even possible.

§II.3 Cognitivistic account

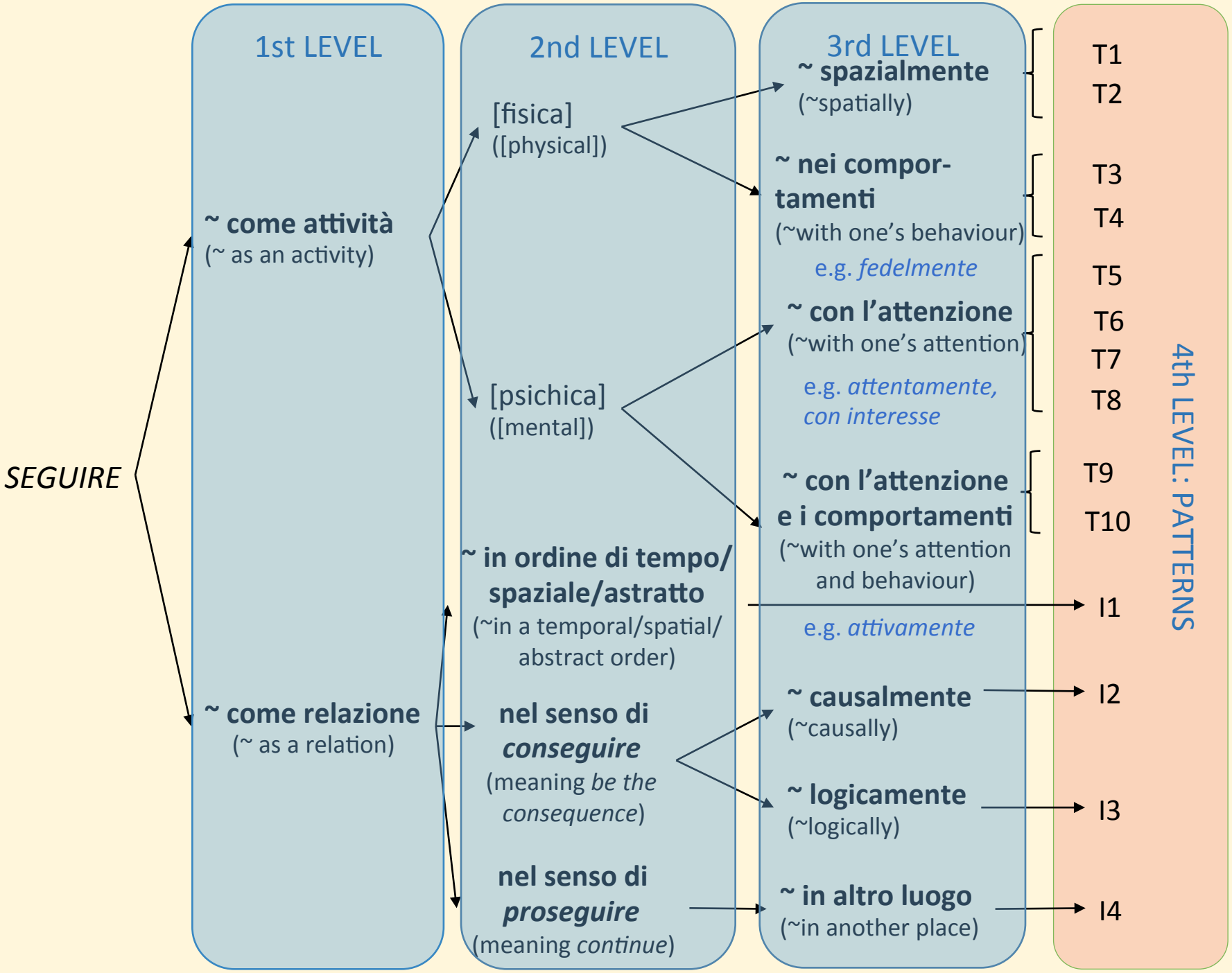
- From a psycholinguistic perspective, we found the main senses of many verbs are related by **cognitive metonymy** and **metaphor**.

- Their **conceptual network** verifies the **cognitivist account** of (complementary) polysemy given by Brugman and Lakoff (cf. Brugman 1988 , Lakoff 1987, Brugman & Lakoff 1988), which postulates that the **related senses** of a word are organized in a **radial set** around usually one **prototypical concept**, just like each individual sense is (in most cases) a conceptual class organized around prototypical members.

§II.4 Extending Sinclair's Hypothesis to sense clusters

- We can often organize the senses of a highly polysemous word in a topology of an ontological nature, grouping them together into ***sense clusters*** according to their semantic similarities, by means of what we may call a (conceptual) ***disambiguating tree***.

- We are currently trying to empirically attest fundamental clusters by the collocations they share (semantically closer senses should share a greater number of collocates), thus testing the hypothesis that fundamental sense clusters, just like individual senses (the true lexical units of language), are **identifiable by phraseology**.


1st LEVEL
2nd LEVEL
3rd LEVEL
4th LEVEL: PATTERNS
SEGUIRE
~ come attività
(~ as an activity)
~ come relazione
(~ as a relation)
[fisica]
([physical])
[psichica]
([mental])
~ in ordine di tempo/ spaziale/astratto
(~in a temporal/spatial/ abstract order)
nel senso di conseguire
(meaning be the consequence)
nel senso di proseguire
(meaning continue)
~ spazialmente
(~spatially)
~ nei compor-tamenti
(~with one's behaviour)
e.g. fedelmente
~ con l'attenzione
(~with one's attention)
e.g. attentamente, con interesse
~ con l'attenzione e i comportamenti
(~with one's attention and behaviour)
e.g. attivamente
~ causalmente
(~causally)
~ logicamente
(~logically)
~ in altro luogo
(~in another place)
T1
T2
T3
T4
T5
T6
T7
T8
T9
T10
I1
I2
I3
I4

§II.3 Finding the right semantic types

- Sinclair's hypothesis is perfectly in line with *construction grammar*[1] and with Tomasello's *Usage-Based Theory of Language Acquisition* (UBTLA), which states (among other things), that:
  1. the **primary psycholinguistic unit** of linguistic communication and in particular of child language acquisition is the **utterance**, not the word. In general, children learn the meaning of utterances before they learn the meaning of the words composing them;
  2. children's **earliest utterances** are almost totally **concrete**: they are collocations;
  3. **new patterns** result from children generalizing across the **semantic variation** they observe **at particular "slots"** in otherwise very similar collocations (tokens of the same utterance).

---

[1] Any linguistic pattern is considered to be a construction as long as some aspect of its form or its meaning cannot be predicted from its component parts, or from other constructions that are recognized to exist. One of the most distinctive features of construction grammar is its emphasis on the importance of multi-word phrases and idioms as fundamental building blocks of language.

- So we ask ouselves: when and how are conventions about word usage stipulated?

III. A model of a sense-disambiguating dictionary based on word sense patterns

§III.1 The shortcomings of sense enumeration

- The shortcomings of the traditional enumerative approach to the representation of word senses are well known. The most common criticism is that enumerations do not make reference to the relations between the senses, and do not adequately describe the kind of knowledge at play in the disambiguation process (Brugman & Lakoff 1988; Norvig 1989; Pustejovsky 1995).

- Entries show a high degree of **meaning fragmentation**.
- Cognitive associations such as those between the prototypical sense of a word and its metonymical and metaphorical “descendants” are difficult to reconstruct.
- This is likely to cause some problems for language learners, who naturally rely on cognitive relations in their *mental lexicon* to comprehend, store and actively access words.

§III.2 From sense enumeration to sense disambiguation

- Our signposts are ***phraseological disambiguators*** at a more general contextual level than sense patterns and mostly correspond to sense clusters. They can be used not only to find the desired senses, but also to learn about the Sinclairian ***extended canonical forms*** of lexical units[2].
- At the highest level of generality are the ***categorical disambiguators*** in square brackets, which correspond to different *morphosyntactic variants* of lemmas, exploiting the relatively tight correspondence, in many Indo-European languages, between semantic and syntactic categories.
- At the lowest disambiguating level are of course **sense patterns**.

_____

[2] Sinclair used the term '(extended) canonical form' to refer to the most explicit presentation of a lexical unit (Sinclair 2004: 298). The shortest unambiguous presentation of the lexical unit he called 'short canonical form' (Sinclair et al. 2004: xxiv).

- As to the claim made by Pustejovsky (1995: 48) that lexica should express the **logical relations between the senses** of a polysemous word, we do not think that this applies to learner's dictionaries, since most of the time cognitive metonymies and cognitive metaphors are only subconsciously perceived by speakers: consciously noticing them can indeed be confusing at first.
- Humans activate the right sense of a word by phraseological disambiguation. As Sinclair realized 35 years ago, phraseology is the true key to solving the polysemy paradox. This is why we are convinced that the fundamental sense clusters of words are very important for disambiguation.

§III.3 Other features

- Our version of word sense patterns is more compact than the one in the COBUILD, in order to speed up the process of disambiguation.
- Their identification numbers appear after them, so that the most impatient and unsystematic readers can quickly skim through all definitions.
- Notice also that we introduced minor senses, such as trivial domain- or situation-specific generalizations and specializations.
- We have listed idioms under the minor senses to make them more easily accessible and comprehensible.

§III.4 Concluding remarks

- The **fourth level** of phrasal information is constituted by the **prototypical instances** of *seguire,* which are collocations.
- John Sinclair, a few years ago, envisioned what he called the "**ultimate dictionary**", containing the most "explicit, full, and unambiguous presentation" of word sense patterns (Sinclair et al. 2004: xxiv).
- Our wish is to be able to make a little but significant step toward its design.

References

Brugman, Claudia M. (1988): The Story of Over. Polysemy, semantics, and the structure of the lexicon. New York, London: Garland (Outstanding dissertations in linguistics).

Brugman, Claudia M.; Lakoff, George (1988): Cognitive topology and lexical networks. In Steven Lawrence Small, Garrison Weeks Cottrell, Michael K. Tanenhaus (Eds.): Lexical ambiguity resolution. Perspectives from psycholinguistics, neuropsychology, and artificial intelligence. San Mateo, Calif.: Morgan Kaufmann Publishers, pp. 477-507.

Hanks, Patrick (2004): Corpus Pattern Analysis. In Geoffrey William, Sandra Vessier (Eds.): Proceedings of the XI EURALEX International Congress, vol. 1. EuraLex XI. Lorient, France, 6.-10.07.2004, pp. 87-98.

Hanks, Patrick (2013): Lexical analysis. Norms and exploitations. Cambridge, Mass.: MIT Press.

Herbst, Thomas (1991): Wörterbücher der Fremdsprachendidaktik: Englisch Dictionaries for Foreign Language Teaching: English Les dictionnaires pour l'enseignement de la langue étrangàre: anglais. In Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand, Ladislav Zgusta (Eds.): Wörterbücher Tome second. Ein internationales Handbuch zur Lexikographie = Dictionaries : an international encyclopedia of lexicography. Berlin: De Gruyter Mouton (Handbücher zur Sprach- und Kommunikationswissenschaft /HSK], 5.2), pp. 1379-1385.

Heuberger, Reinhard (2016): Learners' Dictionaries. History and Development; Current Issues. In Philip Durkin (Ed.): The Oxford Handbook of Lexicography. First edition. Oxford: Oxford University Press (Oxford handbooks in linguistics), pp. 25-43.

Jezek, Elisabetta; Hanks, Patrick (2010): What lexical sets tell us about conceptual categories. In *Lexis - Journal in English Lexicology* 4.

Lakoff, George (1987): Women, fire, and dangerous things. What categories reveal about the mind. Chicago: University of Chicago Press.

Moltmann, Friederike (2016): Natural Language Ontology. In Mark Aronoff (Ed.): Oxford Research Encyclopedia in Linguistics. Oxford: Oxford University Press.

Moon, Rosamund (1987): The analysis of meaning. In John Sinclair (Ed.): Looking Up. An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English language dictionary. London: Collins ELT, pp. 86-103.

Norvig, Peter (1989): Building a large lexicon with lexical network theory. In N. S. Shridaran (Ed.): 11th International Joint Conferences on Artificial Intelligence. Workshop on Lexical Acquisition. Detroit, Michigan, USA. American Association for Artificial Intelligence.

Pustejovsky, James (1995): The generative lexicon. Cambridge, Mass, London: MIT Press.

Sinclair, John (1991): Corpus, Concordance, Collocation. Oxford: Oxford University Press (Describing English language).

Sinclair, John (2004): New evidence, new priorities, new attitudes. In John Sinclair (Ed.): How to Use Corpora in Language Teaching. Philadelphia: J. Benjamins (Studies in corpus linguistics, 1388-0373, v. 12), pp. 271-299.

Sinclair, John; Jones, Susan; Daley, Robert (2004): English Collocation Studies. The Osti Report. Univ of Birmingham.