

# Building Pattern Dictionaries with Corpus Analysis

*Patrick Hanks*

Faculty of Informatics, Masaryk University, Brno

Riga: day 3

**28 November, 2007**

# A Truism

- Meaning is determined by context.
- “You shall know a word by the company it keeps” -- J. R. Firth
  - But what counts as (relevant) context?
  - How to distinguish context from noise?
  - How to infer what is not even explicitly present in the context?
  - How to understand what influences what?
  - How to tell apart senses of a polysemous word?

# A modest proposal

- Words have meaning potentials, not meanings
- Meaning is constructed compositionally
- Sense definitions need to be attached to **patterns**
  - not words in isolation
- Sense definitions need to be data driven (i.e. induced from **corpus evidence**)

# What is a Pattern? (1)

A pattern is an **argument structure** with semantic values for the arguments – i.e. **semantic types** – populated by **lexical sets**.

- **[[Human]] attend [[Event]]**
  - Lexical set [[Event]] = {meeting, conference, funeral, ceremony, course, **school**, seminar, lecture, session, class, rally, dinner, hearing, briefing, reception, workshop, wedding, inquest, **summit**, concert, event, premiere}

# What is a Pattern? (2)

- [[Persona]] partecipare a [[Evento]]
  - Lexical set [[Evento]] = {gara, riunione, **selezione**, manifestazione, seduta, cerimonia, conferenza, votazione, elezione, celebrazione, esequia, competizione, **maratona**, discussione, messa, festa, marcia, fiaccolata, trattativa, missione, commemorazione, incontro, concorso, convegno, raduno, iniziativa, stage, evento, seminario, torneo, attività, corso, asta, raggruppamento, dibattito, progetto, festival}

Semantic types and lexical sets capture different semantic distinctions.

# Patterns are contrastive

- [[Human]] launch [[Boat]]
- [[Human]] launch [[Projectile]]
- [[Human]] launch [[Activity | Plan]]
- [[Human | Institution]] launch [[Artifact = Product]]

# What is a Pattern Dictionary?

- a inventory of **all normal** patterns of verb use.
  - **not** all possible uses.
- an inventory of **semantically motivated syntagmatic distinctions**

# Why is a Pattern Dictionary Necessary?

- **Standard dictionaries:** Dictionaries do not provide the contexts that distinguish one sense of a word from another.
  - very poor syntagmatic information
  - give equal prominence to normal and merely possible senses
  - don't say what needs to be said for NLP
- **WordNet:** synset  $\neq$  word sense!
- **FrameNet** (English only): frame  $\neq$  word syntagmatic behaviour!

# Tools needed to build Pattern Dictionaries

- A balanced corpus of the language (i.e. general language)
- A theory
  - An initial lexical architecture that guides clustering (GL)
  - A lexical model that distinguishes norms from exploitations (TNE)
- A methodology: Corpus Pattern Analysis technique
  - Hanks 2004, Hanks and Pustejovsky 2005
  - *Including statistical corpus analysis*
    - Church and Hanks 1989, Kilgarriff et al. 2004, 2005
- A shallow ontology
  - A hierarchical organization of semantic types
- A suite of corpus tools: Manatee, Bonito, Word Sketch Engine
  - Kilgarriff, Rychlý ???

# What is Corpus Pattern Analysis (CPA)?

Corpus Pattern Analysis (CPA) is a technique that:

1. identifies the typical syntagmatic patterns for each word and determines discriminant context features.
2. catalogs semantic types of arguments that are relevant for distinguishing between different senses.
3. creates an inventory of syntactic and lexical realizations for relevant semantic types.

# CPA procedure

- Create a sample concordance (KWIC index) for a word:
  - 250 examples of actual uses of the word
- Identify the **typical syntagmatic patterns**.
- Assign **each** line of the sample to one of the patterns.
- Take further samples if necessary.
  - Introspection is used to interpret data, but not to create data.
- **Store the pattern in the entry manager.**

# In CPA, every line in the sample must be classified

The choices are:

- Norms
- Exploitations
- Alternations
- Names (*Midnight Storm*: name of a horse, not a storm)
- Mentions (to **mention** a word or phrase is not to **use** it)
- Errors (e.g. *learned* mistyped as *leaned*)
- Unassignables
  - See *Proceedings of the Eleventh EURALEX International Congress*, pages 105–116, Lorient, France, 2004.

# Lexical sets don't map neatly to semantic types

- **[[Human]] divora [[Food]]**
  - Lexical set = {cibo, barretta di cioccolato, fetta di torta, salame, hotdog, colazione, pesce, biscotti, dolci, maiale, pasto, cena, agnello, panino, bistecca}
- **[[Human]] divora [[Document]]**
  - Lexical set = {libro, romanzo, diario, biografia, autore, volume, fumetto, la Deledda}

# Semantic types and Coercion

[[Human]] divorare [[Food | {Animal = Food}]]

with direct objects as follows:

**bistecca, cena**

canonical examples of the semantic type

**pesce, agnello, maiale**

coercions

# Semantic types and Coercion

- [[Human]] divora [[Document ]] ??????????

with direct objects as follows:

**libro, romanzo**

canonical examples of the semantic type

**autore, la Deledda**

coercions

# Lexical sets shimmer

- *leggere (read):* *opinione, commento, libro, avvertenza, recensione, giornale, cronaca, intervista, articolo, blog, messaggio, poesia, notizia, resoconto, racconto, fumetto, romanzo.*
- *pubblicare (publish):* *articolo, intervista, nota, avviso, bando, studio, libro, sondaggio, notizia, volume, saggio, testo, giornale, comunicato, monografia, rivista, documento, inesattezza, annuncio.*
- *spedire (send):* *cartolina, pacco, e-mail, fax, messaggio, lettera, raccomandata, foto, telegramma, copia, sms, merce curriculum, pacchetto, invito, vaglia, posta, file, libro, coupon, modulo.*
- *tradurre (translate):* *testo, frase, bibbia, parola, brano, libro, poesia, vocabolo, concetto, opera, versetto, termine, idea, nome, spiegazione, romanzo, canzone, pensiero, vangelo, espressione.*

# Conclusions

- Creation of selection context dictionaries for NLP applications
- Development of a corpus-driven type system
- Basis for linguistic investigation of mechanisms of coercion and exploitation

# References

- Hanks 2004
- Church - Hanks 1989
- Kilgarriff et al. 2004, 2005
- Hanks - Pustejovsky 2005
- Pustejovsky 2007 (in press)
- Jezek - Lenci 2007 (GL Paris)
- Jezek 2006 Euralex
- Kilgarriff, Richly (2004)
- others?