

Projecte de tesi

Anàlisi i representació semàntiques de les unitats verbals en el discurs especialitzat: primeres experimentacions

Anna Joan-Casademont
Directora: Dra. Mercè Lorente Casafont
Doctorat en Ciències del Llenguatge i Lingüística Aplicada
Bienni 2002-2004
Institut Universitari de Lingüística Aplicada (IULA), UPF

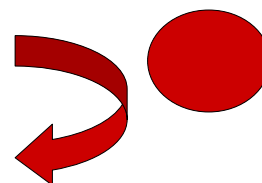
Guió

- Objecte d'estudi i antecedents
- Premisses i hipòtesis
- Objectius
- Marc teòric
- Corpus
- Metodologia
- Primeres experimentacions
- Pla de treball futur

Objecte d'estudi i antecedents

Objecte d'estudi

- UCE
- Semàntica
- Verbs



- Nucli recerca:
Posar en relació propostes de la semàntica lèxica amb el discurs especialitzat.

Antecedents

- Recerca en Museologia
 - Treball acadèmic (vocabulari i estudi semàntic)
 - Comunicacions (variació semàntica; casos concrets)
- Lèxic i Sintaxi (FrameNet per a representació d'UCE)
- Treball de Línia (*Anàlisi de tipologies de descripció semàntica per a una futura aplicació en terminologia*)

Altres elements de l'estudi

- Llengua i àmbits del corpus
- TCT: punt de partida i punt final
- Vessants aplicades:
 - Directes als usuaris: Lexicografia i Terminografia
 - Suport a sistemes: Lingüística computacional
- Interdisciplinarietat

Premisses i hipòtesis

Premisses 1 (TCT)

- **Premissa 1:** el llenguatge especialitzat és llenguatge natural i, per tant, les unitats de coneixement especialitzat (UCE) pertanyen al llenguatge natural produït pels parlants i no són un ítem artificial com preconitzen les perspectives tradicionals de la disciplina terminològica.
- **Premissa 2:** el llenguatge especialitzat, com a llenguatge natural, presenta fenòmens de variació semàntica i pragmàtica en els usos lèxics de les unitats, fenomen especialment rellevant en produccions comunicatives de disciplines especialitzades d'àmbits que situaríem en l'àrea de les ciències humanes i socials.
- **Premissa 3:** una unitat de coneixement especialitzat (UCE) activa aquestes característiques d'especialitat depenent de l'ús que se'n fa i del context, de la situació comunicativa.

Premisses 2 (TCT)

- **Premissa 4:** les unitats predicatives verbals presenten la seva informació semàntica de forma "oberta" en comparació amb les unitats de tipus nominal, ja que sovint despleguen significat complet en la seva predicació, és a dir, en la seva realització de forma conjunta amb els arguments que l'acompanyen.
- **Premissa 5:** des d'una perspectiva comunicativa d'aproximació al llenguatge, hi hauria casos en què allò que volen expressar els parlants a través de la llengua seria detectable a través de la forma amb la qual es comuniquen (relació entre semàntica i sintaxi).
- **Premissa 6:** per poder descriure i representar el llenguatge de forma exhaustiva i fidedigna, cal analitzar les produccions comunicatives reals dels parlants i no partir, per exemple, dels models que estableix la correcció.

Hipòtesis 1

- **Hipòtesi 1:** Les característiques que presenten els verbs en context (forma), juntament amb els trets dels arguments que els acompanyen, són rellevants a l'hora de caracteritzar els diversos sentits d'aquests.
 - Trets sintàctics
 - Trets semàntics
 - Combinació de trets sintàctics i semàntics
 - Trets distintius en diferents nivells i combinació d'aquests

Hipòtesis 2

- **Hipòtesi 2:** De la mateixa manera que amb qualsevol altre tret, el factor "valor terminològic" que activa un verb es manifesta en les ocurrències d'aquest verb a través de la modificació d'un o més trets que el caracteritzaven i/o caracteritzaven els seus arguments, o de la modificació del seu comportament.
 - Tendència canvis d'activació valor terminològic en mateixos tipus unitats i mateixos valors descriptius
 - Tendència canvis d'activació valor terminològic diferents per a cada àmbit d'especialitzat

Hipòtesis 3

- **Hipòtesi 3:** És possible descriure i establir trets i/o patrons de comportament de les unitats lèxiques en ús real a partir d'una anàlisi *in vivo* d'ocurrències de verbs.
 - Comparació aparició i valor trets → correlacions

Hipòtesis 4

- **Hipòtesi 4:** La classificació de les ocurrences dels verbs en patrons iguals de comportament observats, permet relacionar aquests patrons amb sentits concrets d'aquests verbs, ja siguin generals o especialitzats.
 - Tipologia sentits verb amb trets prototípics → predicció nova ocurrencia

Hipòtesis 5

- **Hipòtesi 5:** En les diverses aplicacions de representació semàntica, establir uns criteris sistemàtics que defineixin semànticament els verbs ajuda a (a) la selecció adequada de la informació pertinent, (b) la millora en la forma de representarlos, (c) l'establiment de les relacions adequades entre unitats, i (d) l'especificació de relacions rellevants dins d'una mateixa entrada.
 - Model / protocol informacions verbs → creació aplicacions més consistents, ordenades i sistemàtiques (en relació necessitats específiques de cada cas)

Objectius

Objectius generals

- Objectiu 1:** Anàlisi i descripció exhaustiva de la semàntica de les unitats lèxiques objecte d'estudi (verbs terminològics o que formen part d'UCE) tenint en compte les seves ocurrences reals en el corpus.
- Objectiu 2:** Representació, de forma adequada a les aplicacions, dels trets de la semàntica de les unitats predicatives verbals detectats mitjançant la compleció del primer objectiu.

Objectius específics 1

- objectiu 1.1:** trets que activen especialtat + si presenten sistematicitat
- objectiu 1.2:** variació semàntica, sobretot en ús àmbits pròxims llenguatge general
- objectiu 1.3:** frontera entre sintaxi i semàntica
- objectiu 1.4:** comportament formal dependent del sentit
- objectiu 1.5:** comportament dependent activació valor terminològic
- objectiu 1.6:** rellevància diverses informacions semàntiques obtingudes
- objectiu 1.7:** classes semàntiques dels verbs
- objectiu 1.8:** relacions de comportament entre diversos tipus d'informacions semàntiques
- objectiu 1.9:** relacions semàntiques amb altres verbs (sinonímia, hiperonímia...)
- objectiu 1.10:** intuïcions de comportament fins ara no validades emprícament

Objectius específics 2

- objectiu 2.1:** Informacions semàntiques especialitzades pertinents en aplicacions lexicogràfiques d'àmbit general
- objectiu 2.2:** Informacions semàntiques especialitzades pertinents en aplicacions terminogràfiques
- objectiu 2.3:** Informacions pertinents de variació semàntica especialitzada en aplicacions lexicogràfiques d'àmbit general i terminogràfiques
- objectiu 2.4:** Informacions semàntiques especialitzades per millorar aplicacions computacionals en eficiència i sistematicitat (ex. diccionaris computacionals)
- objectiu 2.5:** Informacions semàntiques especialitzades per millorar aplicacions computacionals en equilibri granularitat i economia (ex. etiquetatge de corpus)
- objectiu 2.6:** informacions semàntiques especialitzades com a element útil per a expansió de consultes en RI amb control terminològic

Objectius: resum

- Dos aspectes d'innovació bàsics com a objectius de la recerca:
 - estudiar unitats lingüístiques mitjançant selecció i adaptació de diverses propostes de descripció semàntica per al llenguatge general (**metodologia analítica eclèctica**)
 - aplicar la metodologia analítica establerta a **produccions comunicatives especialitzades** (necessitem marc teòric com TCT: terminologia = llenguatge natural)
(→ *anàlisi híbrida*)

Marc teòric

Marc teòric: fonaments

- Plantejar-se el llenguatge, aproximacions combinades:
 - **Terminologia** (fonaments de base)
 - Semàntica lèxica
 - Relació entre sintaxi i semàntica
 - Lexicografia
 - Lingüística computacional

Marc teòric: terminologia

- TCT (Cabré): lingüística, comunicativa i cognitiva
 - Llenguatge *natural*
 - Recerca *in vivo*
 - **Verbs** i fraseologia (Lorente)
 - **Variació** (Freixa)
(→ *conseqüències en la pràctica*)

Marc teòric: propostes diverses

TCT → propostes llenguatge natural

- Representatives de les diferents aproximacions
- **Per al llenguatge general**
- Fonts de decisions sobre representació

Marc teòric: propostes diverses

- Jackendoff (semàntica cognitiva, lexicó generatiu)
- Pustejovsky (lexicó generatiu, aplicació computacional)
- Subirats (sintaxi lèxica)
- Levin (relació sintaxi i semàntica, verbs)
- Wierzbicka (semàntica i lexicografia)

Marc teòric: propostes diverses

- **FrameNet** (marcs comunicatius, relació semàntica i arguments)
- **Grimshaw** (estructura argumental)
- **Dowty** (protopapers i papers temàtics)
- **Futures revisions**
 - Aproximacions **recents**
 - Aspectes de **variació**
 - Aplicacions: **lexicografia i terminografia**, i computacionals

Corpus

Corpus: informació bàsica

- **Font:** Corpus Tècnic IULA
- **Àmbit:** Economia (i genòmica)
- **Llengua:** Català (castellà + ...)
- **Elements extrets per a l'anàlisi:** ocurrencies (corpus textual)
- **Elements objecte d'anàlisi:** unitats verbals (base de dades lèxica)

Corpus: evolució

- **Treball de línia:** 51 ocurrencies x 10 verbs
- **Projecte de tesi:** 25 ocurrencies x 2 àmbits x 2 verbs
- **Tesi:** (*provisional*)
 - genòmica i economia
 - català
 - UT i UFE
 - forma personal
 - freqüència (+ de X ocurrencies)
 - ...

```
<m00416> <s>En un inici , tots els dits estan
junts i formen una sola estructura ; per
independitzar - se , les cèl·lules que hi ha
entre els dits moren per un procés de mort
cel·lular programada </s>
<e00099> <s>De resultes d' una
racionalització deshumanitzadora en els
centres de treball , un obrer de 21 anys va
morir a la fàbrica de Murayama </s>
<doc_codi e00037> : </s> <item>1'
especialització en el treball , de manera que
cada treballador produceix només un mateix
producte o part d' un producte
<doc_codi m00523> : neural </s> <s>La
disrupció de l gen xxx en ratolins produceix
mutants amb un ampli espectre d' anomalies
en els teixits
```

Corpus: filtres 1

- **Motiu:** no personals i en perífrasis ≠ comportament
- **Procés:**
 - Observació construccions perífrasis en català
 - Propostes específiques per a estructures
 - Proposta global (soroll vs. silenci); ex.

→

Corpus: filtres 2

Fórmula proposada	Què vull?	Soroll? Silenci?
[pos="V.*(1 2 3)(S P)."] [pos="A B C D E I J L N R T X."]	Vull totes les formes que tinguin un verb conjugat + determinant / identificador / conjunció / adverbi / especificador / interjecció / adjectiu / locució / nom / pronom / data / o xifra,	Recupera les formes personals però no recull les compostes "haver + participi", les formes en pretèrit perfet perífràstic, les construccions "verb + preposició", etc.

Corpus: filtres 3

- Resultat:

```

((pos="V.(1|2|3)(S|P).") [pos="A|B|C|D|E|I|J|L|N|P|R|T|X|."] | ((lemma="haber" &
pos="V.(1|2|3)(S|P).") [pos="VC-SM"] | ((lemma="estar|quedar|restar|deixar|tenir." &
pos="V.(1|2|3)(S|P).") [pos="H"] | ((lemma="haber" & pos="V.(1|2|3)(S|P).") [word="estar"
& pos!="N"] [pos="H|VC-...."] | ((lemma="soler|deure|gostar|poder|podar." &
pos="V.(1|2|3)(S|P).") [pos="VI-...."] | ((pos="V.(1|2|3)(S|P).") [lemma="a|de|d." &
pos="P"] | ((pos="V.(1|2|3)(S|P).") [lemma="a|de|d." &
pos="A|B|C|D|E|H|I|J|L|N|P|R|T|X|VC-....|VG-...."]
o
((pos="V.(1|2|3)(S|P).") [pos="A|B|C|D|E|I|J|L|N|P|R|T|X|."] | ((lemma="haber" &
pos="V.(1|2|3)(S|P).") [pos="VC-SM"] | ((lemma="estar|quedar|restar|deixar|tenir." &
pos="V.(1|2|3)(S|P).") [pos="H"] | ((lemma="haber" & pos="V.(1|2|3)(S|P).") [word="estar"
& pos!="N"] [pos="H|VC-...."] | ((lemma="anar" & pos="V.(1|2|3)(S|P).") [pos="VI-...."] |
((pos="V.(1|2|3)(S|P).") [lemma="a|de|d." & pos="P"] | ((pos="V.(1|2|3)(S|P).")
[lemma="a|de|d." & pos="A|B|C|D|E|H|I|J|L|N|P|R|T|X|VC-....|VG-...."]
    
```

Projecte de tesi



Metodologia

Projecte de tesi



Metodologia: general

- Idea bàsica:
 - Ocurrences en context (sintaxi-semàntica)
 - Informacions:
 - En diferents unitats
 - A diferents nivells de descripció
 - A diferents nivells d'explicitació formal
 - De diferent granularitat informativa
 - De diferent rellevància / representativitat

Projecte de tesi



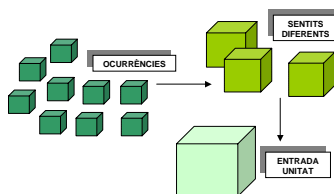
Metodologia: projecte de tesi

- Fitxes manuals
- Conversió en variables 1
- Conversió en variables 2
- i Conversió en variables 3
- Metodologia cap a la tesi

Projecte de tesi



Metodologia: fitxes manuals



Fitxes inicials proposades (esquema)

Projecte de tesi



Metodologia: variables 1

- Primera generalització resultats → **variables i llista numèrica dels valors**
- Problemes:
 - Variables massa lligades a la teoria
 - Intervenció massa subjectiva en l'anàlisi
 - No separació variables verbs i arguments
 - ...

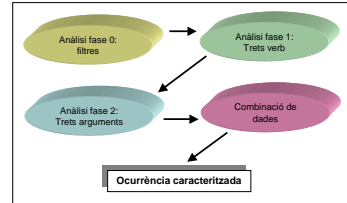
Projecte de tesi



Metodologia: variables 2

- Segona generalització resultats →
- **Canvis:**
 - Supressió variables massa lligades teoria
 - Separació variables verbs i arguments
 - Separació veu i tipus passiva: 2 variables
 - Pas de protopapers a papers temàtics
 - Creació subcapa més específica sintaxi

Metodologia: variables 2 i 3

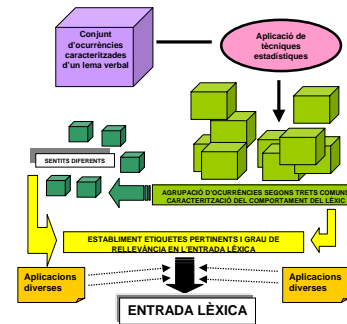


Procés caracterització ocurrències (esquema)

Metodologia: variables 3

- Proposta definitiva projecte de tesi:
- **Canvis:**
 - Eliminació *aktionsart* i telicitat inherent
 - Addició marcs semàntics
 - Canvi en proposta desenvolupament etiquetes semàntiques
- **Futur:** comptable o no, jerarquia de tipus complexa, informació variacional...

Metodologia: cap a la tesi



Primeres experimentacions

Primeres experimentacions 1

Ex. *els gens produeixen el sistema nerviós*

Tipus informació	Dades recollides	V. Núm.
Lema i conjugació	<i>produir</i> , reg., 3a conj. amb -ix., 3a pl.	Vproduir1432
Subcorpus	Genòmica	1
Ocurrència	<doc_codi m00285>: - lo en dues parts : primer , com els gens produeixen el sistema nerviós i segon , com el sistema nerviós	----
Núm. Arguments	2	2
Veu	Activa	1
Pronom reflexiu	No	1
Verb terminològic	Si	2
Marc semàntic	Creació física	45

Primeres experimentacions 2

Ex. *els gens produeixen el sistema nerviós*

Argument 1 (A1)	<i>gens</i>	----
Posició A1	Anterior al verb	1
Terminològic A1?	Sí	2
Sintaxi A1	Grup nominal	1
Tipus Sintaxi A1 (GP...)	El grup nominal no té especificacions	----
Paper temàtic A1	Agent	1
Etiqueta semàntica A1	Física individual animada, organisme	15
Argument 2 (A2)	<i>sistema nerviós</i>	----
Posició A2	Posterior al verb	2
Terminològic A2?	Sí	2
Sintaxi A2	Grup nominal	1
Tipus Sintaxi A2	El grup nominal no té especificacions	----
Paper temàtic A2	Pacient	6
Etiqueta semàntica A2	Física matèria natural	9

Projecte de tesi



43

Primeres experimentacions 3

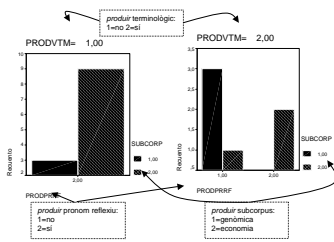
- 2 eixos bàsics anàlisi:
 - Eix 1: establiment variables, relació entre variables, rellevància variables
 - Eix 2: classificació, rellevància variables en classificació
- Diverses classes segons informació:
 - Verbs i formes de realització
 - Verbs, arguments i terminològic
 - Informacions sintàctiques (bàsiques / específiques) i/o informacions semàntiques (bàsiques / específiques)

Projecte de tesi



44

Primeres experimentacions: ex. 1



En aquest país es produeix molt blat

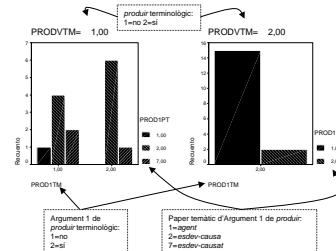
Aquella situació es produïa cada vegada que hi anaven

Projecte de tesi



45

Primeres experimentacions: ex. 2



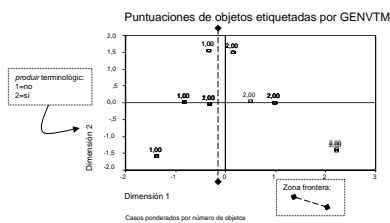
Projecte de tesi



El pas del feudalisme al capitalisme no es produeix per una transformació ràpida
L'empresa produeix articles

46

Primeres experimentacions: ex. 3a



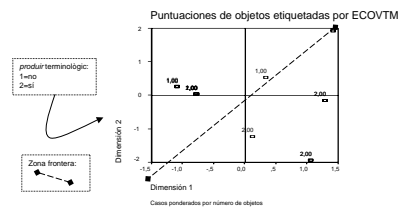
El gen produeix el sistema nerviós

Projecte de tesi



47

Primeres experimentacions: ex. 3b



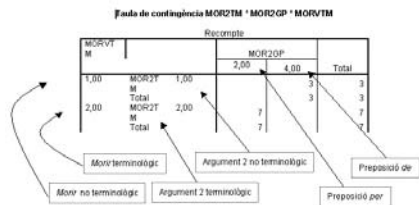
El treballador produeix màquines

Projecte de tesi



48

Primeres experimentacions: ex. 4



Morir de vell / Morir d'un atac de cor
Morir per apoptosi

Pla de treball futur

Pla de treball futur

OBJECTIUS	DATES
Conformació del corpus per a la tesi doctoral	09/05 – 12/05
Prospecció de propostes sobre semàntica i aplicacions concretes	09/05 – 12/05 10/06 – 12/06
Establiment de les variables definitives a analitzar	01/06 – 03/06
Preparació de la interfície d'entrada de dades	01/06 – 03/06
Entrada de dades	04/06 – 06/06
Establiment dels mètodes estadístics per a l'anàlisi	07/06 – 09/06
Anàlisi de comportament	10/06 – 12/06
Redacció de tesi doctoral	01/07 – 04/07
Defensa de tesi doctoral	06/07 – 07/07

MOLTES GRÀCIES

Anàlisi i representació semàntiques
de les unitats verbals en el discurs
especialitzat: primeres
experimentacions

Anna Joan-Casademont
Institut Universitari de Lingüística Aplicada (IULA)
Universitat Pompeu Fabra