PLN

Introducción. Técnicas clásicas

1. CONTEXTO HISTÓRICO
2. CADENAS DE PROCESAMIENTO
3. PREPROCESAMIENTO DE TEXTOS
4. TOKENIZACIÓN
5. SEGMENTACIÓN DE FRASES
6. ANÁLISIS MORFOLÓGICO
7. ANÁLISIS SINTÁCTICO
8. ANÁLISIS SEMÁNTICO

1. CONTEXTO HISTÓRICO

- 50’s Machine Translation en universidades de U.K. y U.S.A.
- 1963 1ª reunión Association for Computational Linguistics

- Técnicas simbólicas
- 80’s Aproximaciones estadísticas o empíricas: basadas en corpus

- Revolución estadística: aproximaciones simbólicas vs. estadísticas
- Técnicas híbridas: incorporación de conocimiento lingüístico en procesamientos estadísticos
- Evolución histórica: gramáticas sobreviven vs métodos estadísticos

2. CADENAS DE PROCESAMIENTO CLÁSICAS

Teoría:

- A nivel de frase:
  · Sintáxis (estructura)
  · Semantica (significado)
- A nivel de discurso
  · pragmática (discurso, contexto)

Práctica:

- No es tan fácil separar
- Cuanto más profundizamos en la cadena, más difícil es el procesamiento, y peores resultados. cada vez más abstracto, menos lingüístico

2. CADENAS DE PROCESAMIENTO CLÁSICAS

Análisis del lenguaje Natural vs. Generación del Lenguaje Natural:

- Mucho más trabajo publicado en análisis que en generación. Análisis empieza por algo concreto (palabras), pero por dónde empezar con Generación? con ideas?

- NLG: para cuando enlatar respuestas no es suficiente

2. CADENAS DE PROCESAMIENTO CLÁSICAS

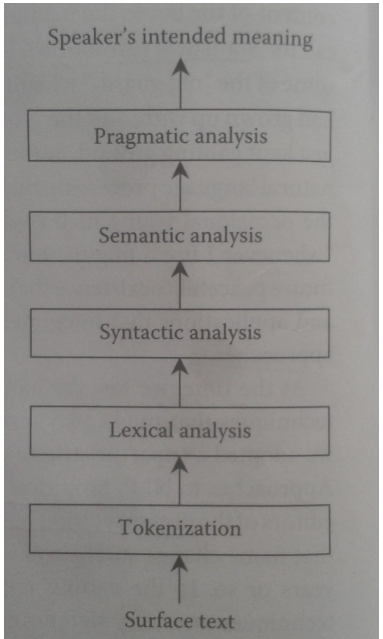


FIGURE 1.1 The stages of analysis in processing natural language.

3. PREPROCESAMIENTO DE TEXTOS

- Document triage

- Construcción de corpus limpios
- Internet Download
- Identificación automática de encoding de caracteres: 7 bits=128 (ASCII), 8 bits=256 (ISO-8859), 16bits=65.536 (chino: Big-5, GB),1-4 bytes (Unicode 5.0, UTF-8)
- Detección de idioma:
    - · Fácil: caracteres que solo hay en un cierto idioma, distancia vectorial de caracteres + frecuentes
    - · Fácil? documentos multilingües, fragmentos cortos
- Filtrado de textos: suprimimos html, imágenes, publicidad y otras basuras

- Segmentación de textos

- Tokenización: segmentación de palabras
- Análisis léxico o morfológico

4. TOKENIZACIÓN

- Tokenización: segmentación de palabras

  · Fácil
    - Idiomas delimitados por espacios

  · Fácil?:
    - Idiomas no segmentados: chino, japonés, tailandés..
    - textos reales: que es una palabra
    - sistemas de escritura: logográfico, silábico, alfabético
    - palabras compuestas
    - ambigüedad
    - morfologías: aislante, aglutinante, inflexional, polisintética

5. SEGMENTACIÓN DE FRASES

· Detección de límites de frases (puntos...)

· Fácil:
    Idiomas con puntuación

· Fácil?
    Qué es una frase donde empieza y acaba
    Textos de internet
    Idiomas sin puntuación

6. ANÁLISIS LÉXICO O MORFOLÓGICO

· Descomposición de tokens en partes:
· Normalización de textos: formas canónicas (1,uno)
· Lematizar
· Asignar etiquetas POS
· Diccionarios vs. reglas de combinación
· Palabras desconocidas (word guessing)
· Identificación de abreviaturas (Sr. U.S.A.) importante para siguiente fase de segmentación de frases
· Ambigüedad
· Lematización<->generación morfológica
usos: MT, IR, stemming, nuevas palabras, correctores, POS tagging, tokenizacion en idiomas sin puntuación
· palabra = morfemas = lexema+morfemas derivativos (pre/in/su - fijo)

6. ANÁLISIS LÉXICO O MORFOLÓGICO

· **Transductores de Estados Finitos (FSTs)**
- son un tipo de Maquina/Automata de Estados Finitos (FSM) con 2 cintas: entrada y salida
- facil: raíz+afijos
- facil? idealizado

· **3 Modelos de análisis léxico:**
- I&A (Item and Arrangement) Finite State Morphology
- I&P (Item and Process) Finite State Morphonology
- W&P (Word and Paradigm)

6. ANÁLISIS LÉXICO O MORFOLÓGICO

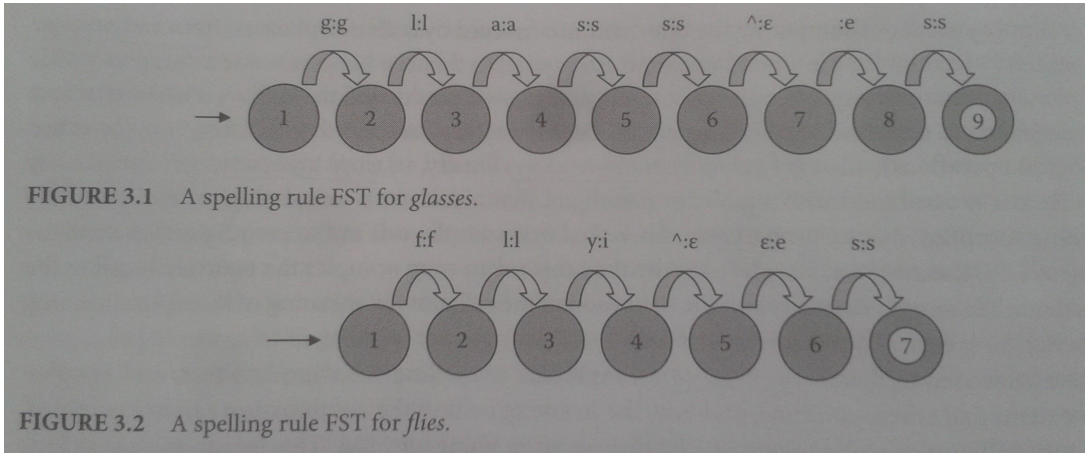
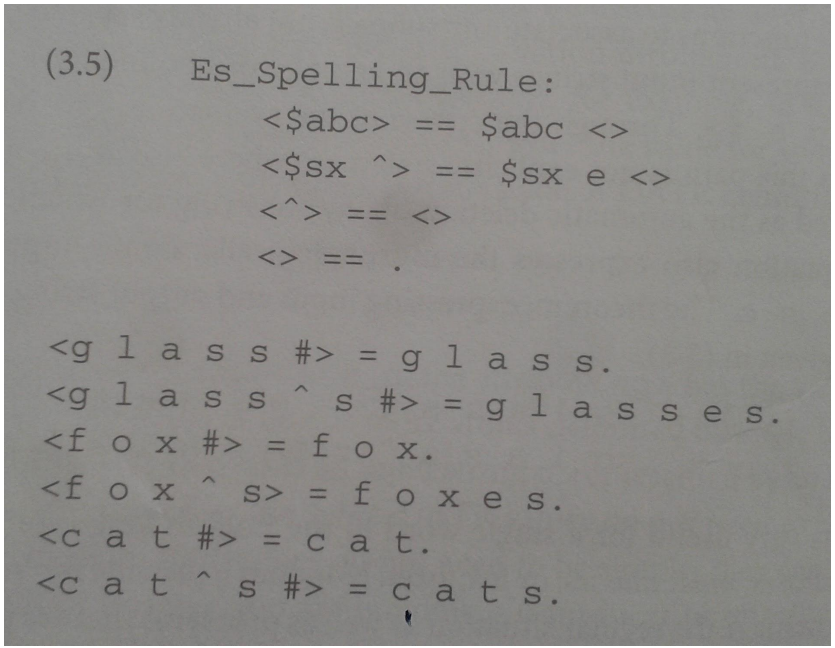


FIGURE 3.1 A spelling rule FST for *glasses*.

FIGURE 3.2 A spelling rule FST for *flies*.

6. ANÁLISIS LÉXICO O MORFOLÓGICO



(3.5)

```
Es_Spelling_Rule:
    <$abc> == $abc <>
    <$sx ^> == $sx e <>
    <^> == <>
    <> == .

<g l a s s #> = g l a s s.
<g l a s s ^ s #> = g l a s s e s.
<f o x #> = f o x.
<f o x ^ s> = f o x e s.
<c a t #> = c a t.
<c a t ^ s #> = c a t s.
```

7. ANÁLISIS SINTÁCTICO

- Árboles sintácticos. Representación: árboles o paréntesis

- Tipos de analizadores:
    - · Recognizer vs. parser vs robust parser
    - · top-down vs bottom-up
    - · deterministico vs no determinístico
    - · de izquierda a derecha vs. otros órdenes

- Estructura de frases.

- Basado en teorías de Lingüística Generativa.

- Facilita el siguiente paso: semántica.

- Parsers NLP != Parsers de programas informáticos:
    - · Distinta capacidad generativa. En NLP mejor SGC > WGC
    - · ambigüedad del lenguaje natural
    - · Ruido: errores y nuevas palabras

7. ANÁLISIS SINTÁCTICO

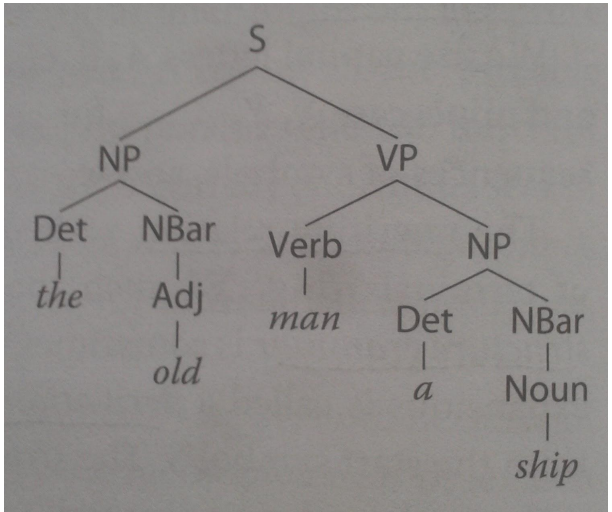


FIGURE 4.2 Syntax tree of the sentence "the old man a ship."

[S [NP [Det *the*] [NBar [Adj *old*]]] [VP [Verb *man*] [NP [Det *a*] [NBar [Noun *ship*]]]]]

7. ANÁLISIS SINTÁCTICO

Who did you sell the car to __?

Who do you think that you sold the car to __?

Who do you think that he suspects that you sold the car to __?

Put the block in the box on the table (4.4)

Assuming that "put" subcategorizes for two objects, there are two possible analyses of (4.4):

Put the block [in the box on the table] (4.5)

Put [the block in the box] on the table (4.6)

7. ANÁLISIS SINTÁCTICO

- Gramáticas Libres de Contexto (CFGs)

- G={T,N,S,R}
- Formas normales:
  - FN de Chomsky (CNF) -> algoritmo CKY
    - A->w
    - A->BC
  - CNF relajada I
    - A->w
    - A->BC
    - A->B
  - CNF relajada II
    - A->w
    - A->BC…
  - Reglas no vacías

7. ANÁLISIS SINTÁCTICO

S	→	NP	VP	Det	→	*a* | *an* | *the*
NP	→	Det	NBar	Adj	→	*old*
NBar	→	Adj	Noun	Noun	→	*man* | *men* | *ship* | *ships*
NBar	→	Noun		Verb	→	*man* | *mans*
NBar	→	Adj				
VP	→	Verb				
VP	→	Verb	NP			

FIGURE 4.1 Example grammar.

7. ANÁLISIS SINTÁCTICO

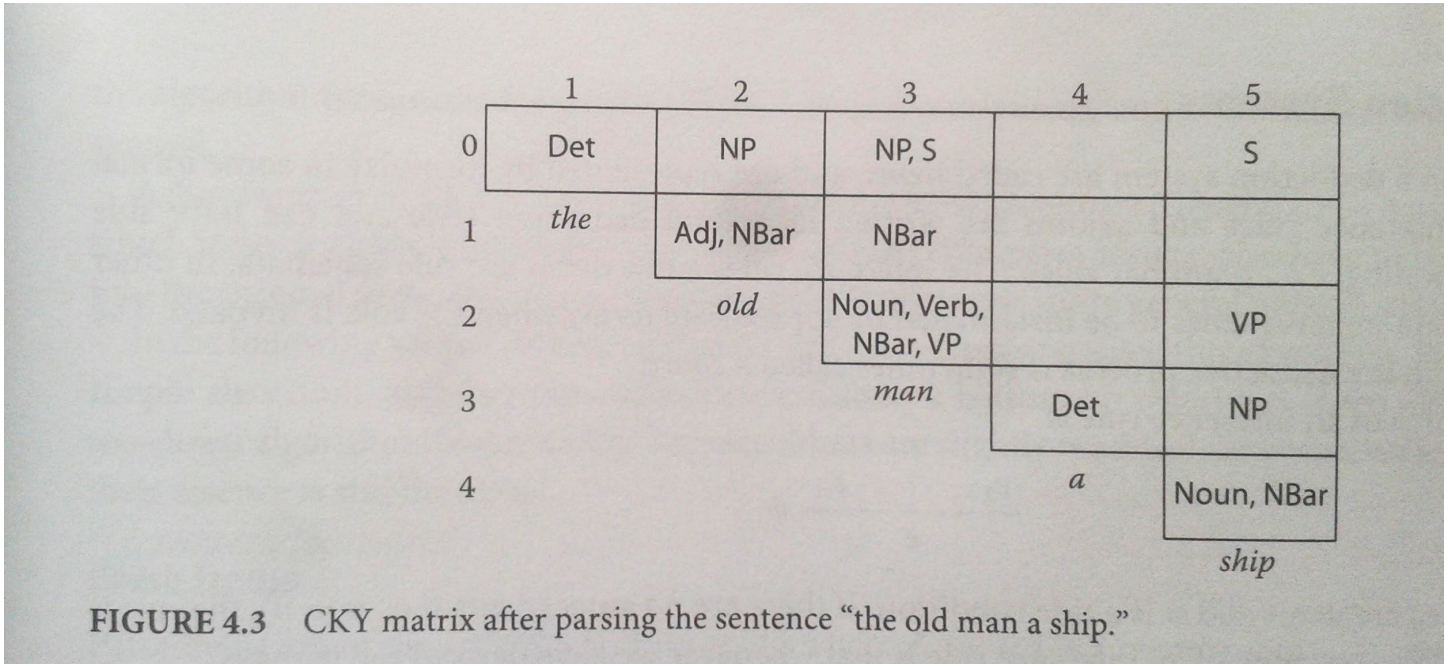


FIGURE 4.3 CKY matrix after parsing the sentence "the old man a ship."

7. ANÁLISIS SINTÁCTICO

- Gramáticas basadas en Restricciones o de Unificación y Rasgos

$$X_0 \rightarrow X_1\ X_2$$
$$\langle X_0\ category\rangle = \mathsf{NP}$$
$$\langle X_1\ category\rangle = \mathsf{Det}$$
$$\langle X_2\ category\rangle = \mathsf{Noun}$$
$$\langle X_1\ agreement\rangle = \langle X_2\ agreement\rangle$$

Any such rule description can be represented as a phrase-structure rule where the symbols consist of feature terms. Below is a feature term rule corresponding to the previous rule (where $\boxed{1}$ indicates identity between the associated elements):

$$\left[category : \mathsf{NP}\right] \rightarrow \begin{bmatrix} category : & \mathsf{Det} \\ agreement : & \boxed{1} \end{bmatrix} \begin{bmatrix} category : & \mathsf{Noun} \\ agreement : & \boxed{1} \end{bmatrix}$$

The basic operation on feature terms is *unification*, which determines if two terms are compatible by merging them to the most general term compatible with both. As an example, the unification $A \sqcup B$ of the terms $A = \left[agreement : \left[number : \mathsf{plural}\right]\right]$ and $B = \left[agreement : \left[gender : \mathsf{neutr}\right]\right]$ succeeds with the result:

$$A \sqcup B = \left[agreement : \begin{bmatrix} gender : & \mathsf{neutr} \\ number : & \mathsf{plural} \end{bmatrix}\right]$$

However, neither A nor $A \sqcup B$ can be unified with

$$C = \left[agreement : \left[number : \mathsf{singular}\right]\right]$$

7. ANÁLISIS SINTÁCTICO

- Características de un buen parser:

- Robustez
- Poder de desambiguación
  - undergeneration vs. overgeneration
  - gramáticas especializadas
  - ranking estadístico
- Eficiencia
  - Eficiencia teórica > Eficiencia práctica

8. ANÁLISIS SEMÁNTICO

- Menos estudiado que los anteriores pasos de la cadena
- Menos consenso
- Entender
- Metalenguaje semántico
- aplicaciones: IR, IE, Resumen textos, data mining, MT, user queries, ontologias, KR
- Semántica léxica vs. Semántica supra léxica
- 3 ambigüedades semánticas:
    - Léxica (homonimia, polisemia)
    - De alcance
    - Referencial
- Teorías semánticas:
    - Formales vs. cognitivas
    - composicionales vs léxicas (descomposicionales y relacionales)

8. ANÁLISIS SEMÁNTICO

- Lógica. Lógica de Predicados.

· variables, términos, predicados, conectivas, cuantificadores existenciales

(1) a. *Some politicians are mortal.*

b. $\exists x$ (**politician**(x) $\wedge$ **mortal**(x))

[There is an x (at least one) so that x is a politician and x is mortal.]

(2) a. *All Australian students like Kevin Rudd.*

b. $\forall x$ ((**student**(x) $\wedge$ **Australian**(x)) $\rightarrow$ **like**(x, k))

[For all x with x being a student and Australian, x likes Kevin Rudd.]

8. ANÁLISIS SEMÁNTICO

(3) a. Modus ponens:
(i) P (premise)
(ii) P → Q (premise)
(iii) Q (conclusion)
b. (i) *Conrad is tired* (P: **tired**(c))
(ii) *Whenever Conrad is tired, he sleeps* (P: **tired**(c), Q: **sleep**(c), P → Q)
(iii) *Conrad sleeps* (Q: **sleep**(c))

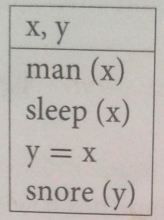
8. ANÁLISIS SEMÁNTICO

- Teoría de Representación del discurso (DRT)

- · Estructuras de Representación del Discurso (DRS)
  - - dinámicas
  - - recursivas
- · DRS = referentes + condiciones
- · Ventajas: capaz de soportar frases nominales indefinidas, cuantificación y resolución de anáfora

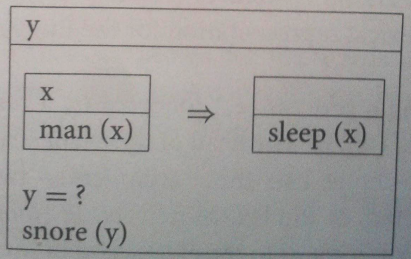
8. ANÁLISIS SEMÁNTICO

(4) *A man sleeps. He snores.*



Recursiveness is an important feature. DRSs ca
example is given in (5). Notice that according to
though on the face of it *every man* is a singular no

(5) *Every man sleeps. He snores.*

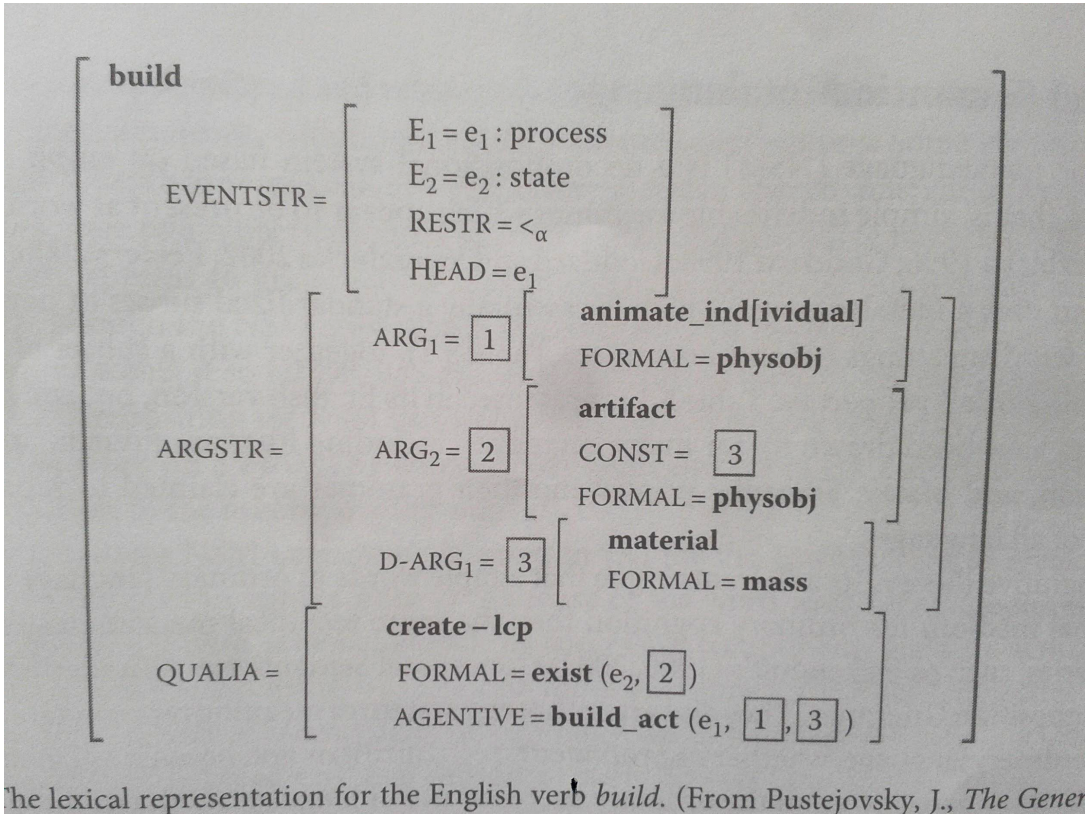


8. ANÁLISIS SEMÁNTICO

- Teoría del Lexicón Generativo de Pustejovsky

- · Dinámica
- · Basada en lexical items
- · Representación en 4 niveles:
  - Argumentos
  - Eventos
  - Qualia
  - Herencia Léxica
- · No orientado a cuantificación, anáfora ni presuposición sino descomposicional

8. ANÁLISIS SEMÁNTICO



The lexical representation for the English verb *build*. (From Pustejovsky, J., *The Gener*

8. ANÁLISIS SEMÁNTICO

- Metalenguaje Semántico Natural (NSM)

- Descompositivo, cognitivo
- Universales semánticos (semantic primes)
- Moléculas semánticas (semantic molecules)

8. ANÁLISIS SEMÁNTICO

TABLE 5.1 Semantic Primes, Grouped into Related Categories

I, YOU, SOMEONE, SOMETHING/THING, PEOPLE, BODY	Substantives
KIND, PART	Relational substantives
THIS, THE SAME, OTHER/ELSE	Determiners
ONE, TWO, SOME, ALL, MUCH/MANY	Quantifiers
GOOD, BAD	Evaluators
BIG, SMALL	Descriptors
KNOW, THINK, WANT, FEEL, SEE, HEAR	Mental predicates
SAY, WORDS, TRUE	Speech
DO, HAPPEN, MOVE, TOUCH	Actions, events, movement, contact
BE (SOMEWHERE), THERE IS, HAVE, BE (SOMEONE/SOMETHING)	Location, existence, possession, specification
LIVE, DIE	Life and death
WHEN/TIME, NOW, BEFORE, AFTER, A LONG TIME, A SHORT TIME, FOR SOME TIME, MOMENT	Time
WHERE/PLACE, HERE, ABOVE, BELOW, FAR, NEAR, SIDE, INSIDE	Space
NOT, MAYBE, CAN, BECAUSE, IF	Logical concepts
VERY, MORE	Intensifier, augmentor
LIKE/WAY	Similarity

Notes: Primes exist as the meanings of lexical units (not at the level of lexemes). Exponents of primes may be words, bound morphemes, or phrasemes. They can be formally complex. They can have combinatorial variants (allolexes). Each prime has well-specified syntactic (combinatorial) properties.

8. ANÁLISIS SEMÁNTICO

TABLE 5.2 Semantic Roles and Their Conventional Definitions

Role	Description
agent	a wilful, purposeful instigator of an action or event
effector	the doer of an action, which may or may not be wilful or purposeful
experiencer	a sentient being that experiences internal states, such as perceivers, cognizers, and emoters
instrument	a normally inanimate entity manipulated by an agent in carrying out an action
force	somewhat like an instrument, but it cannot be manipulated
patient	a thing that is in a state or condition, or undergoes a change of state or condition
theme	a thing which is located or is undergoing a change of location (motion)
benefactive	the participant for whose benefit some action is performed
recipient	someone who gets something (recipients are always animate or some kind of quasi-animate entity)
goal	destination, which is similar to recipient, except that it is often inanimate
source	the point of origin of a state of affairs
location	a place or a spatial locus of a state of affairs
path	a route

Source: Van Valin, R.D. and LaPolla, R.J., *Syntax: Structure, Meaning and Function*, Cambridge University Press, Cambridge, U.K., 1997, 85–89.

8. ANÁLISIS SEMÁNTICO

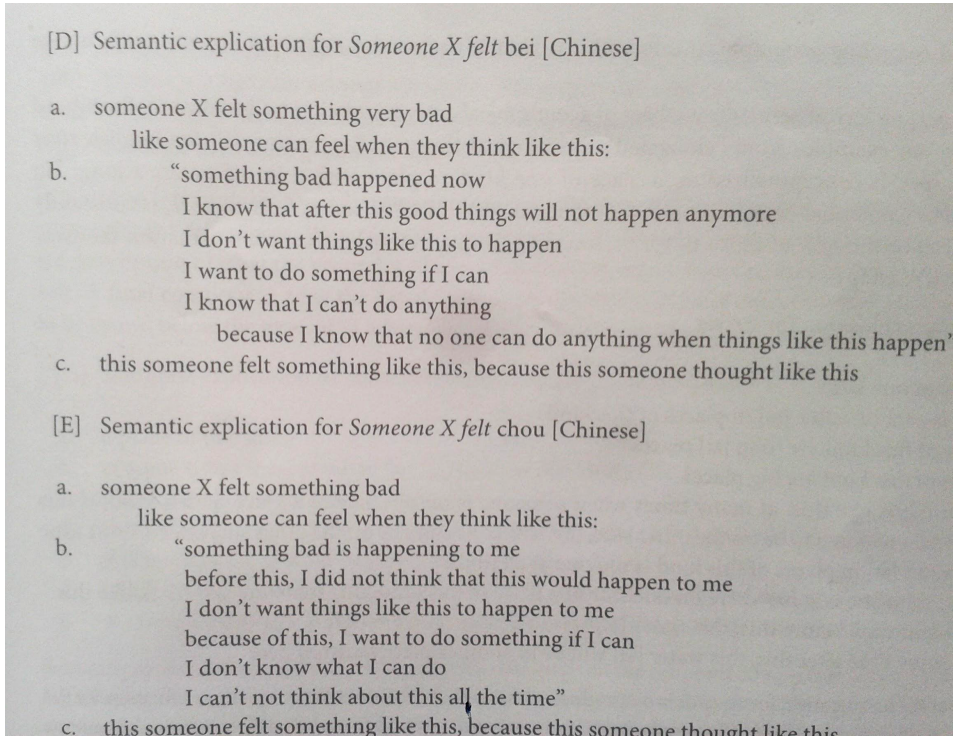
[B] Semantic explication for *Someone X felt sad*

a. someone X felt something bad
like someone can feel when they think like this:
b. "I know that something bad happened
I don't want things like this to happen
I can't think like this: I will do something because of it now
I know that I can't do anything"

[C] Semantic explication for *Someone X felt unhappy*

a. someone X felt something bad
like someone can feel when they think like this:
b. "some bad things happened to me
I wanted things like this not to happen to me
I can't not think about it"
c. this someone felt something like this, because this someone thought like this

8. ANÁLISIS SEMÁNTICO



[D] Semantic explication for *Someone X felt* bei [Chinese]

a. someone X felt something very bad
like someone can feel when they think like this:
b. "something bad happened now
I know that after this good things will not happen anymore
I don't want things like this to happen
I want to do something if I can
I know that I can't do anything
because I know that no one can do anything when things like this happen"
c. this someone felt something like this, because this someone thought like this

[E] Semantic explication for *Someone X felt* chou [Chinese]

a. someone X felt something bad
like someone can feel when they think like this:
b. "something bad is happening to me
before this, I did not think that this would happen to me
I don't want things like this to happen to me
because of this, I want to do something if I can
I don't know what I can do
I can't not think about this all the time"
c. this someone felt something like this, because this someone thought like this

8. ANÁLISIS SEMÁNTICO

TABLE 5.3 Three Different Types of Classificatory Relationships in English

True Taxonomic Category Words	"Plural-Mostly" Functional Macro-Category Words	Singular-Only Functional-Collective Macro-Category Words
birds—sparrow, wren, eagle, . . .	*vegetables—carrots, peas, celery, . . .*	*furniture—table, chair, bed, . . .*
fish—trout, tuna, bream, . . .	*herbs—basil, oregano, rosemary, . . .*	*cutlery—knife, fork, spoon, . . .*
animal—dog, cat, horse, . . .	*cosmetics—lipstick, powder, mascara, . . .*	*jewelry—ring, earring, necklace, . . .*

8. ANÁLISIS SEMÁNTICO

- Semántica Orientada a Objetos

- · Centrada en el verbo
- · Representación UER (Unified Eventity Representation)

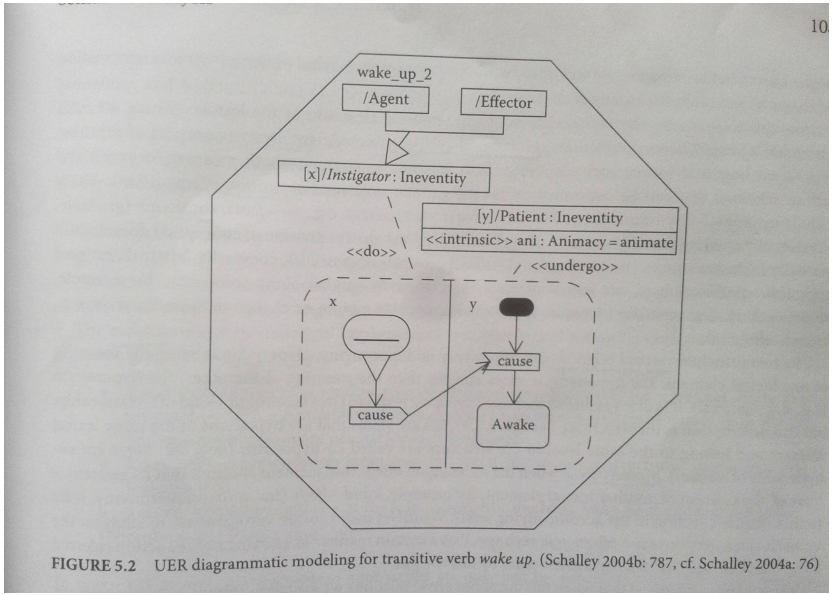


FIGURE 5.2 UER diagrammatic modeling for transitive verb *wake up*. (Schalley 2004b: 787, cf. Schalley 2004a: 76)

8. ANÁLISIS SEMÁNTICO

- Relaciones semánticas y Ontologías

· 1. Relaciones semánticas.

- Wordnet
- Relaciones horizontales:
  - · Sinonimia
  - · Oposición
    - Incompatibilidad
    - Antonimia
    - Complementaridad
    - Conversidad
    - Reversibilidad
- Relaciones verticales
  - · Hiponimia
  - · Meronimia
  - · Troponimia

8. ANÁLISIS SEMÁNTICO

- **2. Relaciones ontológicas**

- **3. Semántica ontológica**
    - Arquitectura:
        - Fuentes de Conocimiento
        - Lenguajes de Representación
        - Módulos de procesamiento

8. ANÁLISIS SEMÁNTICO

- Roles semánticos

- Centrada en el verbo
- FrameNet

TABLE 5.2 Semantic Roles and Their Conventional Definitions

Role	Description
agent	a wilful, purposeful instigator of an action or event
effector	the doer of an action, which may or may not be wilful or purposeful
experiencer	a sentient being that experiences internal states, such as perceivers, cognizers, and emoters
instrument	a normally inanimate entity manipulated by an agent in carrying out an action
force	somewhat like an instrument, but it cannot be manipulated
patient	a thing that is in a state or condition, or undergoes a change of state or condition
theme	a thing which is located or is undergoing a change of location (motion)
benefactive	the participant for whose benefit some action is performed
recipient	someone who gets something (recipients are always animate or some kind of quasi-animate entity)
goal	destination, which is similar to recipient, except that it is often inanimate
source	the point of origin of a state of affairs
location	a place or a spatial locus of a state of affairs
path	a route

Source: Van Valin, R.D. and LaPolla, R.J., *Syntax: Structure, Meaning and Function*, Cambridge University Press, Cambridge, U.K., 1997, 85–89.