

PROJECTE X-TRACT  
PD98-126  
WORKING PAPER N°02/01

Estado del arte sobre treebanks (I)

ref.:WP-XTRACT02/01

Montserrat Civit      Núria Bufí

12 de abril de 2002

# Índice General

<b>1</b>	<b>NEGRA+TIGER</b>	<b>4</b>
1.1	Lectura 1 ((Brants et al., 2001)) . . . . .	4
1.2	Lectura 2 ((Brants et al., 1999)) . . . . .	4
1.3	Lectura 3 ((Brants, 2000)) . . . . .	4
1.4	Otras ((Brants and Plaehn, 2000)) . . . . .	4
<b>2</b>	<b>Prague Dependency Treebank</b>	<b>5</b>
2.1	Lectura 1 ((Hajic and Hladká, 1998)) . . . . .	5
2.2	Lectura 2 ((Bemova et al., 1999)) . . . . .	5
2.3	Lectura 3 ((Böhmova and Hajicova, 1999)) . . . . .	6
2.4	Lectura 4 ((Hajic, 1998)) . . . . .	7
2.5	Lectura 6 y 7 ((Böhmova et al., 1999),(Böhmova and Sgall, ))	7
<b>3</b>	<b>French treebank</b>	<b>8</b>
3.1	Lectura 1 ((Abeillé et al., 2001)) . . . . .	8
3.2	Lectura 2 ((Abeillé et al., 2000)) . . . . .	8
<b>4</b>	<b>TUT</b>	<b>9</b>
4.1	Lectura 1 ((Bosco et al., 2000)) . . . . .	9
<b>5</b>	<b>Spanish Treebank</b>	<b>10</b>
5.1	Lectura 1 ((Moreno et al., 2000)) . . . . .	10
5.2	Lectura 2 ((Moreno and López, 1999)) . . . . .	10
5.3	Lectura 3 ((Moreno et al., 2001)) . . . . .	10
<b>6</b>	<b>ISST</b>	<b>12</b>
6.1	Lectura 1 ((Montemagni et al., 2001)) . . . . .	12
<b>7</b>	<b>MINIPAR</b>	<b>13</b>
7.1	Lectura 1 ((Lin, 2001)) . . . . .	13
<b>8</b>	<b>PENN Treebank</b>	<b>14</b>
8.1	Lecturas 1 ((Taylor et al., 2001)) y 2 ((Marcus et al., 1993)) .	14
<b>9</b>	<b>Susanne corpus</b>	<b>16</b>
9.1	Lectura 1 ((Sampson, 1995)) . . . . .	16

<b>10 Parser Evaluation</b>	<b>17</b>
10.1 Lectura 1 ((Carroll et al., 1998)) . . . . .	17
10.2 Lectura 2 ((Carroll et al., 1999a)) . . . . .	18
10.3 Lecturas 3 ((Carroll et al., 2001)) y 4 ((Carroll et al., 1999b))	18
<b>11 Completing parsed corpora</b>	<b>19</b>
11.1 Lectura 1 ((Wallis, 1999)) . . . . .	19
11.2 Lectura 2 ((Wallis, 2001)) . . . . .	19
<b>Apéndices</b>	<b>21</b>
<b>A Anotación del corpus francés</b>	<b>21</b>
A.1 Principios generales . . . . .	21
A.2 Los diferentes tipos de sintagmas . . . . .	22
<b>Bibliografía</b>	<b>23</b>

# 1 NEGRA+TIGER

## 1.1 Lectura 1 ((Brants et al., 2001))

Anotación sintáctica del corpus a tres niveles:

- (i) estructura sintáctica
- (ii) categorías sintácticas
- (iii) funciones gramaticales

Problemas que presenta la anotación:

- (a) anotación context-free  $\implies$  problemas con los elementos discontinuos
- (b) anotación de dependencias  $\implies$  problemas con los elementos sin núcleo

Proponen un método híbrido de:

- nodos sintácticos
- no se añaden categorías vacías
- estructuras planas (sólo  $X^0$  y  $X'$ , pero no  $X'$ )

Las estructuras planas presentan menor potencial de ambigüedad en el attachment; los árboles están más cercanos a la semántica. Además, si hay ambigüedades remanentes el attach se hace por defecto al nodo más alto posible. Por último, sólo se distinguen las funciones gramaticales más claras.

Método de trabajo: el sistema (un interfaz gráfico) propone de entre los posibles el árbol con la puntuación más alta a partir de la anotación morfológica que debe ser aceptado / rechazado / corregido por el anotador.

## 1.2 Lectura 2 ((Brants et al., 1999))

Las ramas del árbol pueden cruzarse (y pueden convertirse automáticamente en estructuras context-free con huellas).

Categorías sintagmáticas = tipo de sintagmas

funciones gramaticales = funciones sintácticas

Los árboles se construyen bottom-up de modo incremental: el sistema propone hipótesis a partir de lo que ya está analizado para que el anotador lo confirme, ...

## 1.3 Lectura 3 ((Brants, 2000))

Reflexiones sobre la consistencia de la anotación. Comparación de resultados entre distintos anotadores. Comentarios sobre algunos de los casos que causan más problemas.

## 1.4 Otras ((Brants and Plaehn, 2000))

## 2 Prague Dependency Treebank

### 2.1 Lectura 1 ((Hajic and Hladká, 1998))

Anotación morfológica:

3030 etiquetas

378 clases de ambigüedad

Data training: 130.000 tokens desambiguados manualmente dos veces y corregidos por un único *juez*. En general, hay poca ambigüedad en las categorías pero mucha en los rasgos morfológicos.

### 2.2 Lectura 2 ((Bemova et al., 1999))

Tres niveles de anotación:

(i): morfológica

(ii): sintáctica ==> nivel analítico (ATS = *analytic tree level*)

(iii): *linguistic meaning* (TGTS = *tectogrammatical tree structure*)

#### El nivel analítico (ATS)

Representación de las relaciones sintácticas en el interior de una frase: **estructura de dependencias**

1. cada palabra / marca de puntuación se representa en un único nodo
2. no se añaden nodos; excepto el nodo ROOT, de modo que:  
 $\#nodos = \#palabras/signos + 1$
3. no se permite el cruce de ramas
4. RESULTADO: árbol de dependencias donde cada *edge* (= link) está etiquetado
5. TAGs de cada nodo:
  - (a) parte léxica = word form
  - (b) etiqueta morfológica
  - (c) etiqueta sintáctica: nombre del link de dependencia
6. TAGSET:  
60 etiquetas básicas multiplicadas por 3 (coordinación, aposición, paréntesis)  
25 funciones analíticas + funciones para nodos auxiliares (Ejemplos de funciones analíticas: Pred, Obj, Adv, ...)

Algunas observaciones:

1. estructura de dependencias: *head*  $\Leftrightarrow$  *modificador*; PERO esto es problemático para representar ciertas relaciones, como la coordinación
2. Se admiten dobles funciones cuando el anotador no *puede* decidir (p.ej.: ObjAtrib)
3. en la coordinación no hay head

### 2.3 Lectura 3 ((Böhmová and Hajicová, 1999))

ATS: relaciones de dependencia superficiales

TGTS: representaciones subyacentes de la oración

Características de las TGTSs:

1. sólo las palabras autosemánticas tienen nodo (las otras se *añaden* a las palabras autosemánticas)
2. se añaden nodos si claramente no aparecen en la estructura superficial
3. no hay cruce de ramas
4. las funciones analíticas se sustituyen por funciones tectogramaticales (Actor, patient, ...)
5. se añaden rasgos básicos de la estructura informativa de las oraciones: tópico–foco

TAGs del nivel TGTSs:

- lema

- gramatemas morfológicos (= significado de las categorías morfológicas)

- funtores: funciones sintácticas (unas 40, aprox) + algunas subespecificaciones (como por ejemplo: tiempo = {antes, durante, después})

Ejemplos de funtores: actor/bearer, addressee, patient, ...

Proceso de **conversión ATS > TGTS**: dos fases: automática + manual

**Proceso automático secuencial:**

1. Attribute assignment
  - (a) resolución de la modalidad oracional
  - (b) el sujeto de los verbos activos pasa a ACTOR (ssi se sumplen ciertas condiciones)

- (c) conversión de símbolos gráficos a valores + borrado de dichos símbolos
- (d) algunos atributos reciben determinados valores según la información morfológica

## 2. Cambios en la estructura del árbol

- (a) fusión del verbo con los auxiliares verbales
- (b) fusión del verbo con los auxiliares modales
- (c) fusión de los nodos de las preposiciones complejas
- (d) borrado de los nodos de las preposiciones (su valor léxico se incorpora al nombre como parte de la etiqueta) y de las conjunciones de coordinación
- (e) cambio en la dirección de algunas dependencias (p.ej.: en ATS el numeral es el head y el nombre el modifier; en TGTS es al revés)
- (f) cambio en el head del complemento predicativo que pasa a depender del verbo
- (g) borrado de los nodos auxiliares (con función analítica AuxX)
- (h) marcaje con ??? de los casos no resueltos

### **Proceso manual (para ciertos subárboles)**

1. fusión de los nodos para números expresados por más de una palabra
2. esconder subárboles
3. añadir nodos ACTOR para los sujetos elípticos (checo = prodrop)

### **2.4 Lectura 4 ((Hajic, 1998))**

Descripción del tagset de nivel morfológico y analítico

### **2.5 Lectura 6 y 7 ((Böhmova et al., 1999),(Böhmova and Sgall, ))**

Más sobre la conversión automática y manual de ATS en TGTS

## 3 French treebank

### 3.1 Lectura 1 ((Abeillé et al., 2001))

#### POS-tagging

tagset = 212

los compuestos reciben doble etiquetación: el conjunto y cada una de las partes

- tagging pipeline (cada fase trabaja con un subconjunto (no necesariamente disjunto) del conjunto total de etiquetas):

1. 103 etiquetas para el tagger automático
2. 122 etiquetas para los anotadores
3. el conjunto final de etiquetas parece ser de 212 (aunque en ocasiones hablan de 250)

#### Parsing

se anotan constituyentes y funciones sintácticas a nivel superficial

la anotación es de tipo XML:

```
<NP>Paul</NP> <VN>va</VN> <NP>le lundi</NP>...
```

marcan también las valencias verbales

tres fases: chunker, parser, tagger funcional

El **chunker** se utiliza para el clustering léxico.

El **parser** para establecer los límites de sintagmas = marcar constituyentes no recursivos. Funciona con reglas (unas 50) escritas manualmente. Cada palabra funcional es inicio (y final) de un sintagma

El **tagger funcional** asigna las funciones sintácticas a los sintagmas y las valencias verbales

### 3.2 Lectura 2 ((Abeillé et al., 2000))

Más de lo mismo.

**Ahora están desarrollando la fase manual de marcaje de constituyentes** ((Abeillé et al., 2001)).

## 4 TUT

### 4.1 Lectura 1 ((Bosco et al., 2000))

Esquema de anotación basado en dependencias predicado – argumento  
elementos discontinuos representados con huellas

Referencias a (comparación con) otros proyectos:

**PENN**: estructura sintagmática + representación de elementos discontinuos

**PDT**: representación basada en dependencias y orientada a la estructura predicado – argumento

**NEGRA**: *mixed framework* con cruce de ramas

Comparación de los dos modelos básicos:

**estructura sintagmática**: se etiquetan dos tipos de nodos: nodos terminales (palabras) y nodos no terminales

**estructura de dependencias**: sólo se etiquetan nodos terminales y la estructura sintáctica se representa en términos de relaciones binarias entre pares de palabras. En cierta forma está más cerca de la semántica

Proyecto **TUT**: dependencias: se consideran mejores para lenguas no-configuracionales (= de orden libre). También se representan los constituyentes discontinuos y fenómenos como el sujeto vacío y la coordinación (con huellas)

Se ha establecido una jerarquía de relaciones: la parte alta de la jerarquía representa las dependencias básicas, mientras que la parte baja representa dependencias más específicas. El sistema es flexible porque se permite la subespecificación (= el marcaje de una relación con un nodo de la parte alta de la jerarquía). Una parte de esta jerarquía puede observarse en la figura 1.

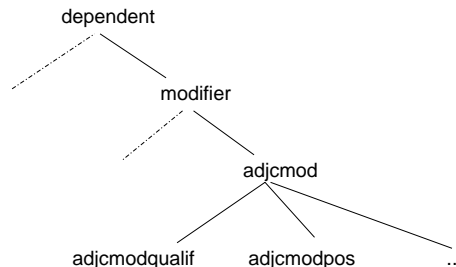


Figura 1: Ejemplo parcial de la jerarquía TUT

Corrección: se sigue el modelo interactivo incremental de NEGRA.

## 5 SpanishTreebank

### 5.1 Lectura 1 ((Moreno et al., 2000))

Se siguen criterios de anotación basados en PennTreebank, Eagles y Negra corpus.

Los árboles son estructuras parentizadas anidadas donde en cada nivel se incluye:

- (a) categoría (*pos* para las palabras o *tipo* para los sintagmas)
- (b) rasgos: sintácticos y semánticos
- (c) nodos constituyentes

El nivel de análisis es de sintaxis superficial, aunque se tratan los sujetos nulos y la elipsis de conjunciones.

#### **TOOLS:**

para **anotación**: un tagger-POS estadístico + desambiguación con una gramática reduccionista y un chunker que reconoce SN, SA, SA<sub>v</sub> y SP.

para **depuración**: un tree-drawer gráfico, un checker de rasgos y un generador de reglas de estructura sintagmática.

### 5.2 Lectura 2 ((Moreno and López, 1999))

Se utilizan etiquetas (*pos*) y rasgos que especifican la información morfosintáctica de cada elemento (terminal y no terminal).

Niveles de anotación:

- *pos* (N, A, ...)
- funciones sintácticas (Suj, Obj1, Obj2, ...)
- rasgos sintácticos (género, número, tiempo, ...)
- algunos rasgos semánticos (human, time, ...)
- lema (hablan de lexema)

La información sólo se especifica una vez, en el nivel correspondiente

Si el anotador no está seguro sobre qué poner, no se anota nada

El nivel es de sintaxis superficial, aunque también se tratan sujeto nulos y otro material elíptico

### 5.3 Lectura 3 ((Moreno et al., 2001))

Se aporta información sobre el tratamiento de algunos fenómenos específicos, como:

1. SE
2. clíticos: doble etiquetación para los postclíticos (enclíticos)

3. expresiones idiomáticas: se tratan conjuntamente y reciben el rasgo IDIOM
4. palabras extranjeras: se les asigna la categoría correspondiente y el rasgo FOREIGNWORD
5. Fechas y horas: se tratan como dos categorías especiales: DATE y HOUR

## 6 ISST

### 6.1 Lectura 1 ((Montemagni et al., 2001))

Proponen cuatro niveles de anotación

1. morfosintáctico
2. sintáctico: constituyentes
3. sintáctico: relaciones funcionales
4. semántico: anotación léxica

el nivel base sobre el que se anota es la forma ortográfica.

Sobre este nivel se anota morfosintácticamente.

Sobre la morfosintaxis se llevan a cabo, por una parte, la anotación sintáctica (ambos aspectos por separado) y la semántica.

Anotación de constituyentes: identificación de límites e identificación de tipos de constituyentes

Anotación de las relaciones funcionales: relaciones binarias entre núcleos léxicos. Se trata también la coordinación, la correferencia interclausal y la elipsis.

## 7 MINIPAR

### 7.1 Lectura 1 ((Lin, 2001))

Objetivo: evaluación de pársers con un método basado en dependencias

MINIPAR: parser de amplia cobertura probado sobre el corpus SUSAN-NE. Resultados: 79% cobertura y 89% precisión.

Una cuestión de terminología:

**answers (a)** = árboles de análisis generados por el pársers

**keys (k)** = árboles de análisis construidos manualmente

En principio la evaluación debería ser la comparación entre  $a$  y  $k$ , pero este sistema presenta problemas, por lo que se propone un sistema de evaluación en que  $a$  y  $k$  se convierten a árboles de dependencias y los valores de la evaluación se computan según el conjunto de relaciones de dependencia en  $a$  y  $k$ .

Los árboles de dependencias se representan como tuplas

`word category [head] [relationship]`

donde los dos últimos elementos son opcionales.

`[head]` es la palabra que es modificada por `word`; es un indicador de posición.

`[relationship]` es la etiqueta asignada a la relación de dependencia.

La evaluación se hace comparando palabra a palabra

MINIPAR representa la gramática como una red donde los nodos ( $\# = 35$ ) son las categorías gramaticales y los links ( $\# = 59$ ) los tipos de relaciones de dependencia sintáctica. El lexicón que utiliza se ha derivado de WN.

## 8 PENN Treebank

### 8.1 Lecturas 1 ((Taylor et al., 2001)) y 2 ((Marcus et al., 1993))

CORPUS: 7M/words anotadas morfológicamente; 3M anotadas a nivel de constituyentes; y 2M anotadas con las estructura predicado-argumentos, y 1'6M de texto oral con anotaciones de disfluencias.

1. Anotación morfológica: reducción del tagset del Brown Corpus
  - 36 etiquetas para categorías y rasgos morfológicos
  - 12 etiquetas para signos de puntuación y símbolosAnotación es dos fases: tagger automático + corrección manual
  
2. Anotación sintáctica
  - Dos fases:
    - 1) **etiquetación de constituyentes** (parentización):  
Proceso de anotación: automático (que sólo proporciona un análisis para cada frase; sólo en caso de ambigüedad se presentan diferentes análisis del fragmento ambiguo ), + corrección manual
  
    - 2) **estructura predicado-argumentos:**
      - (a) funciones sintácticas (sujeto, objeto, ...)
      - (b) tratamiento de los elementos vacíos que incluye huellas (sujetos de infinitivo, movimiento de Q, pasivas, ...)
      - (c) correferencias (para elementos vacíos y huellas con coindización marcada numéricamente en la categoría no-terminal; los índices se rellenan con la referencia)
      - (d) tipos anómalos de coordinación
      - (e) tratamiento de los constituyentes discontinuos con coindización para señalar la estructura discontinua (pseudo-attachment).
      - (f) papeles semánticos (agente, paciente, ...)
      - (g) tipos de adverbios (lugar, tiempo, modo, ...)
      - (h) ambigüedad sintáctica: attach (se resuelve manualmente)

Etiquetas utilizadas:

#### **17 en la primera fase**

**segunda fase:** aparecen nuevas etiquetas para elementos que sólo se anotan en caso de que la distinción sea muy clara. Pueden ser de 4 tipos diferentes:

- 1) categorías textuales (3; p.ej.: títulos)
- 2) funciones gramaticales (8; p.ej.: tópico, sujeto lógico en las pasivas, ...)
- 3) papeles semánticos (6; p.ej.: vocativos, dirección, lugar, ...)
- 4) pseudo-attachment (4; p.ej.: expletivos, ambigüedades irresolubles (*I saw the man with the telescope*), *right node raising*, *interpret constituent here*)

La última parte del proyecto PennTreebank es una transcripción de conversaciones telefónicas, donde se anotan las disfluencias más comunes del habla (repeticiones, interrupciones, frases incompletas...). 1,6M/words.

## 9 Susanne corpus

### 9.1 Lectura 1 ((Sampson, 1995))

(*English for computers*) 128.000 palabras anotadas morfológica y sintácticamente de modo manual.

Etiquetas morfológicas: 355 (aprox.)

Anotación sintáctica de constituyentes y funciones

Etiquetas sintácticas: de constituyentes: 55 para sintagmas + 17 para subordinadas + 7 para nodos raíz/oración + 23 para funciones sintácticas.

Proceso de anotación totalmente manual.

Tratamiento de los elementos nulos (elipsis) con coindización.

## 10 Parser Evaluation

### 10.1 Lectura 1 ((Carroll et al., 1998))

Repaso a los diferentes métodos de evaluación de parsers, que queda resumido en la figura 2.

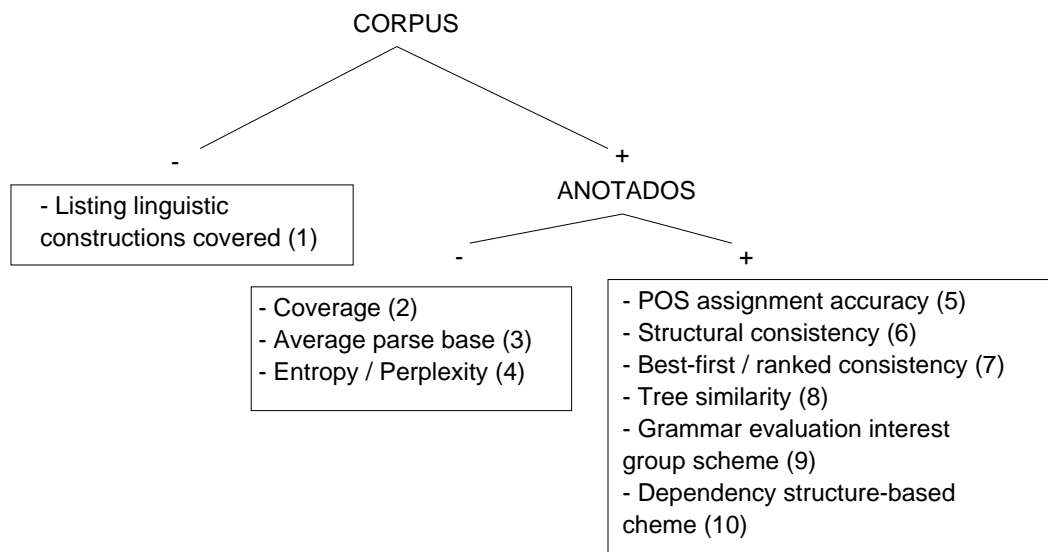


Figura 2: Resumen de los diferentes métodos de evaluación de parsers

En general puede decirse que los métodos de evaluación sirven para:

(a) guiar-controlar el desarrollo de un sistema particular de análisis o de una gramática particular (éste sería el caso de los métodos (1), (2), (3), (6) y probablemente (7)).

(b) comparar diferentes sistemas. En este campo el principal problema es que por lo general no suelen relacionarse los análisis con posibles aplicaciones o tareas de parsing y que un mismo esquema de evaluación debería poder aplicarse al output de parsers basados en diferentes teorías y/o aplicados a distintas lenguas.

Proponen entender las relaciones gramaticales (GR) como dependencias entre *head-dependent*. Su propuesta presupone que se dispone de un detector de heads. El esquema de relaciones es el que aparece en la figura 3.

(ncX = non-clausalX)

(cX = X controlado desde dentro)

(xX = X controlado desde fuera)

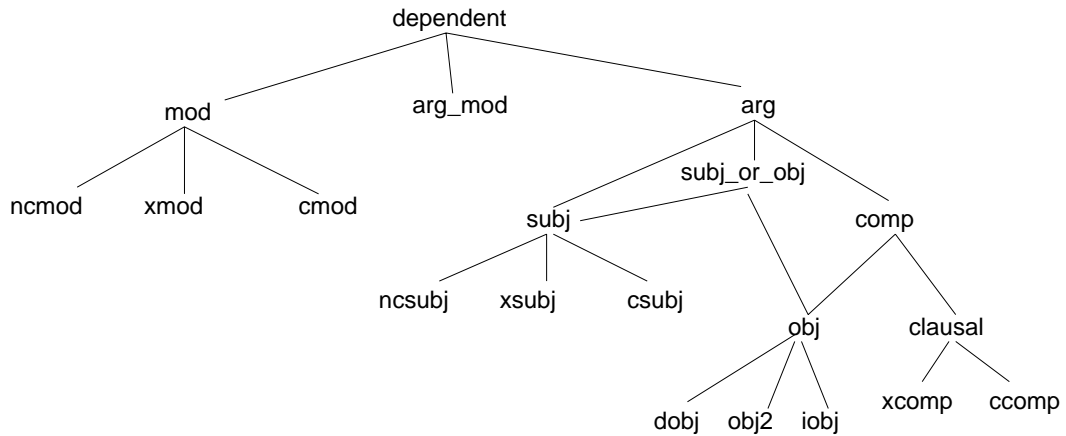


Figura 3: Jerarquía de relaciones

La evaluación se lleva a cabo con las medidas estándar de precisión y recall, pero teniendo en cuenta que el match puede ser total (relación exacta) o parcial (con un nodo madre = subespecificación)

El sistema se ha utilizado en el proyecto SPARKLE para evaluar parsers del inglés, francés, alemán e italiano diseñados según diferentes puntos de vista.

## 10.2 Lectura 2 ((Carroll et al., 1999a))

Es un subconjunto de la siguiente

## 10.3 Lecturas 3 ((Carroll et al., 2001)) y 4 ((Carroll et al., 1999b))

El sistema de evaluación de parsers más utilizado es PARSEVAL, que, entre otras cosas, no es aplicable a las representaciones de dependencias. Sin embargo, Lin98 propone un sistema en que los constituyentes se convierten en dependencias, con lo que ambos métodos pueden ya compararse.

Por lo demás, este artículo vuelve a presentar la jerarquía anteriormente mencionada.

## 11 Completing parsed corpora

### 11.1 Lectura 1 ((Wallis, 1999))

Básicamente igual que la siguiente, aunque aquí el corpus se presenta de forma más detallada.

### 11.2 Lectura 2 ((Wallis, 2001))

ICE-GB: treebank del inglés con 1 millón de palabras procedentes de fuentes orales y escritas. El tipo de marcaje sintáctico es superficial y se anotan los constituyentes, etiquetados con:

- categoría
- rol funcional
- rasgos adicionales (transitividad, coordinación, número, ...)

Comentarios sobre los diferentes usos que ya se han dado a este corpus.

Se ha editado un documento de más de 200 páginas como guía para los anotadores.

Fase de parsing automático: se ha utilizado TOSCA, un parser top-down basado en reglas que produce diferentes análisis entre los que hay que elegir uno manualmente. Proporciona análisis para el 75% de las oraciones. El 25% restante se analiza con un parser probabilístico que genera un único análisis.

Cuando se anota corpus hay que hacer frente a dos problemas básicos:

**the decision problem:** a saber, cuál es el análisis correcto, y

**the consistency problem:** que depende mucho del esquema de anotación adoptado, porque:

*if the guidelines are strictly deterministic, then surely they could be automated*

Sinclair92 argumenta contra la anotación manual:

- a) los anotadores introducen errores
- b) sus esfuerzos no aumentan a escala

Contraargumentos:

- 1) el estado del arte: los resultados de los pársers automáticos hacen necesaria la intervención humana
- 2) los correctores deben tomar decisiones concretas que hay que generalizar para incorporar a la futura corrección

Hay que llevar a cabo una síntesis entre una y otra postura.

**Métodos de corrección:**

**longitudinal:** durante el parsing o después (checking). En este caso no hay forma de saber cuál es la precisión (*accuracy*).

**transversal:** permite una especialización en la tarea.

**ventajas del método transversal:**

1. refuerza la consistencia
2. facilita el problema de la decisión:

*if a decision can be made once, it can be made several times*

3. el número de errores obvios y de inconsistencias cae

**inconvenientes del método transversal:**

1. se requiere un software más sofisticado, por ejemplo, un sistema de *queries*
2. es más difícil controlar el proceso, porque hay que actualizar constantemente los datos
3. puede producirse solapamiento en las correcciones
4. a veces las búsquedas pueden resultar difíciles

De qué depende la corrección transversal?

1. del tamaño del corpus
2. de la complejidad en el análisis: cuantos más detalles, más posibles inconsistencias
3. de las aplicaciones: algunas pueden requerir más precisión que otras

Pero hay un camino intermedio:

- 1) corrección transversal sólo para algunos casos concretos
- 2) corrección longitudinal para el resto

## A Anotación del corpus francés

**Objetivo:** anotación de constituyentes: categorías, límites izquierdos y límites derechos. Se trata de corregir la anotación hecha por un shallow parser que ha realizado agrupaciones mínimas, por lo que la tarea consiste básicamente en ampliar constituyentes.

Si aparecen errores residuales de la fase de tagging, no se corrigen.

No se pueden modificar los límites de frase existentes.

Los ":" marcan final de frase ssi introducen discurso directo.

### A.1 Principios generales

Como el objetivo es utilizar el corpus tanto para fines puramente informáticos como lingüísticos o psicolingüísticos, no se pretende aplicar ninguna teoría concreta.

El marcaje de los constituyentes es el primer paso hacia el marcaje posterior de funciones sintácticas.

1. No hay constituyentes discontinuos ni cruce de sintagmas

```
Jean <VN> ne veut </VN> pas <VPinf> <VN> venir </VN> </VPinf>
Jean <VN> veut </VN> <VPinf> ne_pas:Adv <VN> venir </VN> </VPinf>
Jean <VN> n'est pas venu </VN>
Jean <VN> n'a <NP> rien:Pro </NP> vu </VN> .
```

2. No se anotan categorías vacías para constituyentes elípticos o desplazados
3. Sintagmas exocéntricos

No hay categoría vacía para sintagmas elípticos o sin *head*:  
tagging

```
un ingénieur maison_N
il est très famille_N
une rouge_A
les meilleurs_A
```

syntax

```
<NP> les meilleurs:A </NP>
```

Y lo mismo para frases sin verbo:

```
<SSub> que toi </SSub>
<SRel> dont trois idiots </SRel>
```

Nunca se inserta un verbo elíptico

4. No hay ambigüedad residual: si es necesario se recorre al conocimiento enciclopédico.

```
<NP> Bill Clinton
  <COORD> , <NP> le maire de New York </NP> </COORD> ,
  <COORD> et <NP> Madeleine Albright </NP> </COORD>
</NP>
```

```
<NP> Bill Clinton, <NP> le président des Etats-Unis </NP> ,
<COORD> et <NP> Madeleine Albright </NP> </COORD> </NP>
```

Si se da el caso de que se puede optar por dos o más construcciones, se opta por la que implica menos anidamiento:

```
Jean <VN> a commis </VN>
  <NP> une agression </NP>
  <PP> contre <NP> Marie </NP> </PP>
```

5. Los sintagmas unarios

Se intenta reducirlos al máximo, aunque evidentemente se dan bastantes casos de este tipo de sintagmas.

6. Categorías léxicas y sus sintagmas correspondientes

En la tabla 1 pueden observarse los 12 tipos de sintagmas utilizados.

Tal como puede observarse no hay sintagma verbal, puesto que, o bien lo incluye todo, con lo que resulta inútil, o bien sólo incluye complementos y es discontinuo.

Tampoco hay sintagma determinante:

```
<NP> presque:Adv tous:A les:D enfants:NC </NP>
<NP> près-de:P trois-cents:D personnes </NP>
<NP> deux:D <COORD> ou:CC trois:D </COORD> enfants:NC </NP>
<NP> deux:D à:P trois:D enfants:NC </NP> : laisser plat.
```

## A.2 Los diferentes tipos de sintagmas

AP	syntagme adjectival
AdP	syntagme adverbial
COORD	syntagme (ou phrase) coordonné(e)
NP	syntagme nominal
VN	noyau verbal (verbes, clitiques, auxiliaires, faire)
PP	syntagme prépositionnel
SENT	phrase indépendante (On note aussi ainsi tout fragment indépendant - isolé par une ponctuation forte, un saut de ligne etc. - par exemple les titres des articles.)
VPpart	proposition participiale (part passé ou part présent)
VPinf	proposition infinitive (pouvant commencer par une préposition)
Srel	proposition relative (commençant par un pronom relatif ou un PP incluant un Pro rel)
Ssub	proposition subordonnée (complétive, interrogative indirecte, subordonnée circonstancielle)
Sint	proposition conjuguée interne (coordonnée, discours direct, incise)

Tabla 1: 12 tipos de sintagmas

## Referencias

- Abeillé, A., Toussanel, F., and Chéradame, M. (2001). Corpus le monde. annotations en constituants. guide pour les correcteurs. Technical report, LLF, UFRL. dernière mise à jour: 12-oct-2001.
- Abeillé, A., Clément, L., and Kinyon, A. (2000). Building a treebank for French. In *Proceedings of the Second Conference on Language Resources and Evaluation (LREC2000)*, pages 87–94, Athens, Greece.
- Abeillé, A., Clément, L., and Kinyon, A. (2001). *Building and Using syntactically annotated corpora*, chapter Building a treebank for French. Language and Speech. Kluwer, Dordrecht.
- Bemova, A., Hajic, J., Hladka, B., and Panevova, J. (1999). Morphological and Syntactic Tagging of The Prague Dependency Treebank. Journées Atala, Corpus annotés pour la syntaxe, Paris. available: <http://talana.linguist.jussieu.fr/treebanks99/>.
- Böhmova, A. and Hajicova, E. (1999). How Much of the Underlying Syntactic Structure can be Tagged Automatically. Jour-

nées Atala, Corpus annotés pour la syntaxe, Paris. available: <http://talana.linguist.jussieu.fr/treebanks99/>.

- Böhmová, A., Panevová, J., and Sgall, P. (1999). Syntactic Tagging: Procedure for the Transition from the Analytic to the Tectogrammatical Tree Structures. In *Proceedings of the Second Workshop on Text, Speech, Dialogue*, Mariánské lázně, Czech Republic. available [http://ufal.mff.cuni.cz/pdt/pdt\\_05.html](http://ufal.mff.cuni.cz/pdt/pdt_05.html).
- Böhmová, A. and Sgall, P. Automatic procedures in tectogrammatical tagging. available [http://ufal.mff.cuni.cz/pdt/pdt\\_05.html](http://ufal.mff.cuni.cz/pdt/pdt_05.html).
- Bosco, C., Lombardo, V., Vassallo, D., and Lesmo, L. (2000). Building a treebank for Italian: a Data-driven Annotation Schema. In *Proceedings of the Second Conference on Language Resources and Evaluation (LREC2000)*, pages 99–105, Athens, Greece.
- Brants, T. (2000). Inter-Annotator Agreement for a German Newspaper Corpus. In *Proceedings of the Second International Conference on Language and Evaluation LREC-2000*, Athens, Greece.
- Brants, T. and Plaehn, O. (2000). Interactive Corpus Annotation. In *Proceedings of the Second International Conference on Language and Evaluation LREC-2000*, Athens, Greece.
- Brants, T., Skut, W., and Uszkoreit, H. (1999). Syntactic Annotation of a German Newspaper Corpus. Journées Atala, Corpus annotés pour la syntaxe, Paris. available: <http://talana.linguist.jussieu.fr/treebanks99/>.
- Brants, T., Skut, W., and Uszkoreit, H. (2001). *Building and Using syntactically annotated corpora*, chapter Syntactic Annotation of a German Newspaper Corpus. Language and Speech. Kluwer, Dordrecht. available: <http://treebank.linguist.jussieu.fr/toc.html>.
- Carroll, J., Briscoe, T., and Sanfilippo, A. (1998). Parser Evaluation: a Survey and a New proposal. In *Proceedings of the First Conference on Language Resources and Evaluation. LREC'98*, pages 447–454, Granada.
- Carroll, J., Minnen, G., and Briscoe, T. (1999a). Corpus Annotation for Parser Evaluation. Journées Atala, Corpus annotés pour la syntaxe, Paris. available: <http://talana.linguist.jussieu.fr/treebanks99/>.

- Carroll, J., Minnen, G., and Briscoe, T. (1999b). Corpus Annotation for Parser Evaluation. In *LINC-99, Workshop at the 9th Conference of the EACL (EACL-99)*, Bergen, Norway.
- Carroll, J., Minnen, G., and Briscoe, T. (2001). *Building and Using syntactically annotated corpora*, chapter Parser evaluation using a grammatical relation annotation scheme. Language and Speech. Kluwer, Dordrecht. available: <http://treebank.linguist.jussieu.fr/toc.html>.
- Hajic, J. (1998). Building a Syntactically Annotated Corpus: the Prague dependency Treebank. *Issues of Valency and meaning*, pages 106–132.
- Hajic, J. and Hladká, B. (1998). Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *Proceedings of the 36th Annual Meeting of the ACL and the 17th ICCL*, pages 483–490, Montréal. available [http://ufal.mff.cuni.cz/pdt/pdt\\_05.html](http://ufal.mff.cuni.cz/pdt/pdt_05.html).
- Lin, D. (2001). *Building and Using syntactically annotated corpora*, chapter Dependency-based Evaluation of MINIPAR. Language and Speech. Kluwer, Dordrecht. available: <http://treebank.linguist.jussieu.fr/toc.html>.
- Marcus, M., Santorini, B., and Marcinkiewicz, M. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*. available: <http://www.cis.upenn.edu/treebank/home.html>.
- Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Lenci, A., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M., D.Saracino, Zanzotto, F., Mana, N., Pianesi, F., and Delmonte, R. (2001). *Building and Using syntactically annotated corpora*, chapter Building the Italian Syntactic-Semantic Treebank. Language and Speech. Kluwer, Dordrecht. available: <http://treebank.linguist.jussieu.fr/toc.html>.
- Moreno, A., Grishman, R., López, S., Sánchez, F., and Sekine, S. (2000). A Treebank of Spanish and its Application to Parsing. In *Proceedings of the Second Conference on Language Resources and Evaluation (LREC2000)*, pages 107–111, Athens, Greece.
- Moreno, A. and López, S. (1999). Developing a Spanish TreeBank. Journées Atala, Corpus annotés pour la syntaxe, Paris. available: <http://talana.linguist.jussieu.fr/treebanks99/>.

- Moreno, A., López, S., Sánchez, F., and Grishman, R. (2001). *Building and Using syntactically annotated corpora*, chapter Developing a Spanish Treebank. Language and Speech. Kluwer, Dordrecht. available: <http://treebank.linguist.jussieu.fr/toc.html>.
- Sampson, G. (1995). *English for the Computer. The SUSANNE corpus and Analytic Scheme*. Clarendon Press, Oxford.
- Taylor, A., Marcus, M., and Santorini, B. (2001). *Building and Using syntactically annotated corpora*, chapter The Penn Treebank: an overview. Language and Speech. Kluwer, Dordrecht. available: <http://treebank.linguist.jussieu.fr/toc.html>.
- Wallis, S. (1999). Completing parsed corpora: from Correction to Evolution. Journées Atala, Corpus annotés pour la syntaxe, Paris. available: <http://talana.linguist.jussieu.fr/treebanks99/>.
- Wallis, S. (2001). *Building and Using syntactically annotated corpora*, chapter Completing parsed corpora: from correction to evolution. Language and Speech. Kluwer, Dordrecht. available: <http://treebank.linguist.jussieu.fr/toc.html>.