# Corpus Linguistics, Resources and Normalisation

Sylvain Pogodalla
sylvain.pogodalla@loria.fr
Bureau : B. 242

# Outline

# Epistemological Aspects

- “The study of language based on examples of ’real life’ language use” [McEnery and Wilson(1996)]
- Not only a *methodology*, but also a *theory*
- Not by itself a linguistic branch (unlike phonology, syntax or semantics), but *transversal* (in particular form the methodological point of view)

## Some History

Before it was named: "The approach began with a large collection of recorded utterances from some language, a corpus. The corpus was subject to a clear, stepwise, bottom-up strategy of analysis" [Harris(1993)]

1876 – 1926 Studies of the language of the children [Preyer(1889)] (a lot of examples from a lot of children, a lot of examples from few children—longitudinal)

1897 Study of the frequency distribution of letters and letter sequences in German (stenography) : 11 millions words, 5000 analysts [Käding(1897)]

1940 Corpus for research on foreign language learning, vocabulary lists, word frequency

1940 Comparative linguistics, comparison of word sense in different languages [Eaton(1940)]. Corpora enabling the same kind of analyses was recreated only from 1996 [McEnery and Oakes(1996)]

1950 Syntax and semantics: a descriptive grammar of English based on a corpus [Fries(1952)]. For French, [Georges Gougenheim and Sauvageot(1956)] describes a grammar based on grammatical choices and lexical frequency computed from 275 speakers.

From the 50's to the 80's, less work. In particular because Chomsky's criticism.

## Some History

Before it was named: "The approach began with a large collection of recorded utterances from some language, a corpus. The corpus was subject to a clear, stepwise, bottom-up strategy of analysis" [Harris(1993)]

1876 – 1926 Studies of the language of the children [Preyer(1889)] (a lot of examples from a lot of children, a lot of examples from few children—longitudinal)

1897 Study of the frequency distribution of letters and letter sequences in German (stenography) : 11 millions words, 5000 analysts [Käding(1897)]

1940 Corpus for research on foreign language learning, vocabulary lists, word frequency

1940 Comparative linguistics, comparison of word sense in different languages [Eaton(1940)]. Corpora enabling the same kind of analyses was recreated only from 1996 [McEnery and Oakes(1996)]

1950 Syntax and semantics: a descriptive grammar of English based on a corpus [Fries(1952)]. For French, [Georges Gougenheim and Sauvageot(1956)] describes a grammar based on grammatical choices and lexical frequency computed from 275 speakers.

From the 50's to the 80's, less work. In particular because Chomsky's criticism.

# Some History

Before it was named: "The approach began with a large collection of recorded utterances from some language, a corpus. The corpus was subject to a clear, stepwise, bottom-up strategy of analysis" [Harris(1993)]

- **1876 – 1926** Studies of the language of the children [Preyer(1889)] (a lot of examples from a lot of children, a lot of examples from few children—longitudinal)
- **1897** Study of the frequency distribution of letters and letter sequences in German (stenography) : 11 millions words, 5000 analysts [Käding(1897)]
- **1940** Corpus for research on foreign language learning, vocabulary lists, word frequency
- **1940** Comparative linguistics, comparison of word sense in different languages [Eaton(1940)]. Corpora enabling the same kind of analyses was recreated only from 1996 [McEnery and Oakes(1996)]
- **1950** Syntax and semantics: a descriptive grammar of English based on a corpus [Fries(1952)]. For French, [Georges Gougenheim and Sauvageot(1956)] describes a grammar based on grammatical choices and lexical frequency computed from 275 speakers.

From the 50's to the 80's, less work. In particular because Chomsky's criticism.

## Some History

Before it was named: "The approach began with a large collection of recorded utterances from some language, a corpus. The corpus was subject to a clear, stepwise, bottom-up strategy of analysis" [Harris(1993)]

**1876 – 1926** Studies of the language of the children [Preyer(1889)] (a lot of examples from a lot of children, a lot of examples from few children—longitudinal)

**1897** Study of the frequency distribution of letters and letter sequences in German (stenography) : 11 millions words, 5000 analysts [Käding(1897)]

**1940** Corpus for research on foreign language learning, vocabulary lists, word frequency

**1940** Comparative linguistics, comparison of word sense in different languages [Eaton(1940)]. Corpora enabling the same kind of analyses was recreated only from 1996 [McEnery and Oakes(1996)]

**1950** Syntax and semantics: a descriptive grammar of English based on a corpus [Fries(1952)]. For French, [Georges Gougenheim and Sauvageot(1956)] describes a grammar based on grammatical choices and lexical frequency computed from 275 speakers.

From the 50's to the 80's, less work. In particular because Chomsky's criticism.

## Some History

Before it was named: "The approach began with a large collection of recorded utterances from some language, a corpus. The corpus was subject to a clear, stepwise, bottom-up strategy of analysis" [Harris(1993)]

- **1876 – 1926** Studies of the language of the children [Preyer(1889)] (a lot of examples from a lot of children, a lot of examples from few children—longitudinal)
- **1897** Study of the frequency distribution of letters and letter sequences in German (stenography) : 11 millions words, 5000 analysts [Käding(1897)]
- **1940** Corpus for research on foreign language learning, vocabulary lists, word frequency
- **1940** Comparative linguistics, comparison of word sense in different languages [Eaton(1940)]. Corpora enabling the same kind of analyses was recreated only from 1996 [McEnery and Oakes(1996)]
- **1950** Syntax and semantics: a descriptive grammar of English based on a corpus [Fries(1952)]. For French, [Georges Gougenheim and Sauvageot(1956)] describes a grammar based on grammatical choices and lexical frequency computed from 275 speakers.

From the 50's to the 80's, less work. In particular because Chomsky's criticism.

## Some History

Before it was named: "The approach began with a large collection of recorded utterances from some language, a corpus. The corpus was subject to a clear, stepwise, bottom-up strategy of analysis" [Harris(1993)]

- **1876 – 1926** Studies of the language of the children [Preyer(1889)] (a lot of examples from a lot of children, a lot of examples from few children—longitudinal)
- **1897** Study of the frequency distribution of letters and letter sequences in German (stenography) : 11 millions words, 5000 analysts [Käding(1897)]
- **1940** Corpus for research on foreign language learning, vocabulary lists, word frequency
- **1940** Comparative linguistics, comparison of word sense in different languages [Eaton(1940)]. Corpora enabling the same kind of analyses was recreated only from 1996 [McEnery and Oakes(1996)]
- **1950** Syntax and semantics: a descriptive grammar of English based on a corpus [Fries(1952)]. For French, [Georges Gougenheim and Sauvageot(1956)] describes a grammar based on grammatical choices and lexical frequency computed from 275 speakers.

From the 50's to the 80's, less work. In particular because Chomsky's criticism.

## Some History

Before it was named: "The approach began with a large collection of recorded utterances from some language, a corpus. The corpus was subject to a clear, stepwise, bottom-up strategy of analysis" [Harris(1993)]

- **1876 – 1926** Studies of the language of the children [Preyer(1889)] (a lot of examples from a lot of children, a lot of examples from few children—longitudinal)
- **1897** Study of the frequency distribution of letters and letter sequences in German (stenography) : 11 millions words, 5000 analysts [Käding(1897)]
- **1940** Corpus for research on foreign language learning, vocabulary lists, word frequency
- **1940** Comparative linguistics, comparison of word sense in different languages [Eaton(1940)]. Corpora enabling the same kind of analyses was recreated only from 1996 [McEnery and Oakes(1996)]
- **1950** Syntax and semantics: a descriptive grammar of English based on a corpus [Fries(1952)]. For French, [Georges Gougenheim and Sauvageot(1956)] describes a grammar based on grammatical choices and lexical frequency computed from 275 speakers.

From the 50's to the 80's, less work. In particular because Chomsky's criticism.

# Chomsky's Criticism

## About the use of coprpora in linguistics

- Using a corpus, what is being analysed? A property of the grammar or a "social" phenomenon?
- *Competence* and *performance*. ▸ Example Theoretical *linguists* study competence.
- Corpora reflect performance. Hence, competence ⇒ no corpus
- Competence ⇒ Introspection

## Some exceptions

- The study of language learning by children (because requires meta-knowledge on the language)
- Languages that are not spoken anymore
- Phonetics
- Historical linguistics
- ...

## But what about NLP?

# Chomsky's Criticism

## About the use of coprpora in linguistics

- Using a corpus, what is being analysed? A property of the grammar or a "social" phenomenon?
- *Competence* and *performance*. ▸ Example Theoretical *linguists* study competence.
- Corpora reflect performance. Hence, competence ⇒ no corpus
- Competence ⇒ Introspection

## Some exceptions

- The study of language learning by children (because requires meta-knowledge on the language)
- Languages that are not spoken anymore
- Phonetics
- Historical linguistics
- ...

## But what about NLP?

# Chomsky's Criticism

## About the use of coprpora in linguistics

- Using a corpus, what is being analysed? A property of the grammar or a "social" phenomenon?
- *Competence* and *performance*. ▸ Example Theoretical *linguists* study competence.
- Corpora reflect performance. Hence, competence $\Rightarrow$ no corpus
- Competence $\Rightarrow$ Introspection

## Some exceptions

- The study of language learning by children (because requires meta-knowledge on the language)
- Languages that are not spoken anymore
- Phonetics
- Historical linguistics
- ...

## But what about NLP?

# Chomsky's Criticism

## About the use of coprpora in linguistics

- Using a corpus, what is being analysed? A property of the grammar or a "social" phenomenon?
- *Competence* and *performance*. ▸ Example Theoretical *linguists* study competence.
- Corpora reflect performance. Hence, competence ⇒ no corpus
- Competence ⇒ Introspection

## Some exceptions

- The study of language learning by children (because requires meta-knowledge on the language)
- Languages that are not spoken anymore
- Phonetics
- Historical linguistics
- ...

## But what about NLP?

# Chomsky's Criticism

## About the use of coprpora in linguistics

- Using a corpus, what is being analysed? A property of the grammar or a "social" phenomenon?
- *Competence* and *performance*. Example Theoretical *linguists* study competence.
- Corpora reflect performance. Hence, competence $\Rightarrow$ no corpus
- Competence $\Rightarrow$ Introspection

## Some exceptions

- The study of language learning by children (because requires meta-knowledge on the language)
- Languages that are not spoken anymore
- Phonetics
- Historical linguistics
- ...

## But what about NLP?

# Chomsky's Criticism

## About the use of coprpora in linguistics

- Using a corpus, what is being analysed? A property of the grammar or a "social" phenomenon?
- *Competence* and *performance*. ▸ Example Theoretical *linguists* study competence.
- Corpora reflect performance. Hence, competence ⇒ no corpus
- Competence ⇒ Introspection

## Some exceptions

- The study of language learning by children (because requires meta-knowledge on the language)
- Languages that are not spoken anymore
- Phonetics
- Historical linguistics
- ...

## But what about NLP?

# Competence and performance

[Morrill(2000)]

Skip this slide

Go to the end

**Do you understand?**

The dog that chased the cat that saw the rat that ate the cheese barked

**Is it grammatically correct?**

The dog that chased the cat that saw the rat that ate the cheese barked

**Do you understand?**

The cheese that the rat that the cat that the dog chased saw ate stank

**Is it grammatically correct?**

The cheese that the rat that the cat that the dog chased saw ate stank

# Competence and performance

[Morrill(2000)]

Skip this slide

Go to the end

## Do you understand?

The dog that chased the cat that saw the rat that ate the cheese barked

## Is it grammatically correct?

The dog that chased the cat that saw the rat that ate the cheese barked

## Do you understand?

The cheese that the rat that the cat that the dog chased saw ate stank

## Is it grammatically correct?

The cheese that the rat that the cat that the dog chased saw ate stank

# Competence and performance

[Morrill(2000)]

Skip this slide

Go to the end

### Do you understand?

The dog that chased the cat that saw the rat that ate the cheese barked

### Is it grammatically correct?

The dog that chased the cat that saw the rat that ate the cheese barked

### Do you understand?

The cheese that the rat that the cat that the dog chased saw ate stank

### Is it grammatically correct?

The cheese that the rat that the cat that the dog chased saw ate stank

# Competence and performance
[Morrill(2000)]

Skip this slide

Go to the end

**Do you understand?**

The dog that chased the cat that saw the rat that ate the cheese barked

**Is it grammatically correct?**

The dog that chased the cat that saw the rat that ate the cheese barked

**Do you understand?**

The cheese that the rat that the cat that the dog chased saw ate stank

**Is it grammatically correct?**

The cheese that the rat that the cat that the dog chased saw ate stank

# Competence and performance

[Morrill(2000)]

Skip this slide

Go to the end

### Do you understand?

The dog that chased the cat that saw the rat that ate the cheese barked

### Is it grammatically correct?

The dog that chased the cat that saw the rat that ate the cheese barked

### Do you understand?

The cheese that the rat that the cat that the dog chased saw ate stank

### Is it grammatically correct?

The cheese that the rat that the cat that the dog chased saw ate stank

# Competence and performance [Morrill(2000)]

Skip this slide

Go to the end

**Do you understand?**

The dog that chased the cat that saw the rat that ate the cheese barked

**Is it grammatically correct?**

The dog that chased the cat that saw the rat that ate the cheese barked

**Do you understand?**

The cheese that the rat that the cat that the dog chased saw ate stank

**Is it grammatically correct?**

The cheese that the rat that the cat that the dog chased saw ate stank

# Competence and performance
[Morrill(2000)]

Skip this slide

Go to the end

**Do you understand?**

The dog that chased the cat that saw the rat that ate the cheese barked

**Is it grammatically correct?**

The dog that chased the cat that saw the rat that ate the cheese barked

**Do you understand?**

The cheese that the rat that the cat that the dog chased saw ate stank

**Is it grammatically correct?**

The cheese that the rat that the cat that the dog chased saw ate stank

# Competence and performance

[Morrill(2000)]

Skip this slide

Go to the end

**Do you understand?**

The dog that chased the cat that saw the rat that ate the cheese barked

**Is it grammatically correct?**

The dog that chased the cat that saw the rat that ate the cheese barked

**Do you understand?**

The cheese that the rat that the cat that the dog chased saw ate stank

**Is it grammatically correct?**

The cheese that the rat that the cat that the dog chased saw ate stank

# Competence and performance [Morrill(2000)]

Skip this slide

Go to the end

## Do you understand?

The dog that chased the cat that saw the rat that ate the cheese barked

## Is it grammatically correct?

The dog that chased the cat that saw the rat that ate the cheese barked

## Do you understand?

The cheese that the rat that the cat that the dog chased saw ate stank

## Is it grammatically correct?

The cheese that the rat that the cat that the dog chased saw ate stank

# Competence and performance

[Morrill(2000)]

Skip this slide

Go to the end

### Do you understand?

The dog that chased the cat that saw the rat that ate the cheese barked

### Is it grammatically correct?

The dog that chased the cat that saw the rat that ate the cheese barked

### Do you understand?

The cheese that the rat that the cat that the dog chased saw ate stank

### Is it grammatically correct?

The cheese that the rat that the cat that the dog chased saw ate stank

# Competence and performance
[Morrill(2000)]

Skip this slide

Go to the end

## Do you understand?

The dog that chased the cat that saw the rat that ate the cheese barked

## Is it grammatically correct?

The dog that chased the cat that saw the rat that ate the cheese barked

## Do you understand?

The cheese that the rat that the cat that the dog chased saw ate stank

## Is it grammatically correct?

The cheese that the rat that the cat that the dog chased saw ate stank

# Competence and performance [Morrill(2000)]

Skip this slide

Go to the end

**Do you understand?**

The dog that chased the cat that saw the rat that ate the cheese barked

**Is it grammatically correct?**

The dog that chased the cat that saw the rat that ate the cheese barked

**Do you understand?**

The cheese that the rat that the cat that the dog chased saw ate stank

**Is it grammatically correct?**

The cheese that the rat that the cat that the dog chased saw ate stank

Back to the previous slide

## Introspection and Data

- What is the nature of data in linguistics (experimental data)?
- What benefits from a corpus data?

  *[The corpus linguist] He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus the second word of a sentence [Fillmore(1992)]*

- Nevertheless [Hill(1962)] about introspection

  *Chomsky: The verb 'perform' cannot be used with mass word objects: one can 'perform a task' but one cannot 'perform labour'*
  *Hatcher: How do you know, if you don't use a corpus and have not studied the verb 'perform'?*
  *Chomsky': How do I know'? Because I am a native speaker of the English language*

- But, from BNC corpus, it is possible to 'perform magic'

# Introduction and Data

- What is the nature of data in linguistics (experimental data)?
- What benefits from a corpus data?

> *[The corpus linguist] He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus the second word of a sentence [Fillmore(1992)]*

- Nevertheless [Hill(1962)] about introspection

> *Chomsky: The verb 'perform' cannot be used with mass word objects: one can 'perform a task' but one cannot 'perform labour'*
> *Hatcher: How do you know, if you don't use a corpus and have not studied the verb 'perform'?*
> *Chomsky': How do I know'? Because I am a native speaker of the English language*

- But, from BNC corpus, it is possible to 'perform magic'

# Introspection and Data

- What is the nature of data in linguistics (experimental data)?
- What benefits from a corpus data?

  *[The corpus linguist] He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus the second word of a sentence [Fillmore(1992)]*

- Nevertheless [Hill(1962)] about introspection

  *Chomsky: The verb 'perform' cannot be used with mass word objects: one can 'perform a task' but one cannot 'perform labour'*
  *Hatcher: How do you know, if you don't use a corpus and have not studied the verb 'perform'?*
  *Chomsky': How do I know'? Because I am a native speaker of the English language*

- But, from BNC corpus, it is possible to 'perform magic'

## Introspection and Data

- What is the nature of data in linguistics (experimental data)?
- What benefits from a corpus data?

> *[The corpus linguist] He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus the second word of a sentence [Fillmore(1992)]*

- Nevertheless [Hill(1962)] about introspection

> *Chomsky: The verb 'perform' cannot be used with mass word objects: one can 'perform a task' but one cannot 'perform labour'*
> *Hatcher: How do you know, if you don't use a corpus and have not studied the verb 'perform'?*
> *Chomsky': How do I know'? Because I am a native speaker of the English language*

- But, from BNC corpus, it is possible to 'perform magic'

## Introspection and Data

- What is the nature of data in linguistics (experimental data)?
- What benefits from a corpus data?

  *[The corpus linguist] He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus the second word of a sentence [Fillmore(1992)]*

- Nevertheless [Hill(1962)] about introspection

  *Chomsky: The verb 'perform' cannot be used with mass word objects: one can 'perform a task' but one cannot 'perform labour'*
  *Hatcher: How do you know, if you don't use a corpus and have not studied the verb 'perform'?*
  *Chomsky': How do I know'? Because I am a native speaker of the English language*

- But, from BNC corpus, it is possible to 'perform magic'

## Introspection and Data

- What is the nature of data in linguistics (experimental data)?
- What benefits from a corpus data?

  *[The corpus linguist] He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus the second word of a sentence [Fillmore(1992)]*

- Nevertheless [Hill(1962)] about introspection

  *Chomsky: The verb 'perform' cannot be used with mass word objects: one can 'perform a task' but one cannot 'perform labour'*
  *Hatcher: How do you know, if you don't use a corpus and have not studied the verb 'perform'?*
  *Chomsky': How do I know'? Because I am a native speaker of the English language*

- But, from BNC corpus, it is possible to 'perform magic'

# Introspection and Data

### Intuition is sometimes useful:

- In a corpus : *il parle à* + proper name
- Not in a corpus : *il mange à* + proper name

### How is it possible to say that *il mange à Pierre* is ungrammatical?

There is no reason *a priori* that *parle* is different from *mange*.

# Introspection and Data

## Intuition is sometimes useful:

- In a corpus : *il parle à* + proper name
- Not in a corpus : *il mange à* + proper name

## How is it possible to say that *il mange à Pierre* is ungrammatical?

There is no reason *a priori* that *parle* is different from *mange*.

# Introduction and Data

**Intuition is sometimes useful:**

- In a corpus : *il parle à* + proper name
- Not in a corpus : *il mange à* + proper name

**How is it possible to say that *il mange à Pierre* is ungrammatical?**

There is no reason *a priori* that *parle* is different from *mange*.

## Introspection and Data

### Intuition is sometimes useful:

- In a corpus : *il parle à* + proper name
- Not in a corpus : *il mange à* + proper name

### How is it possible to say that *il mange à Pierre* is ungrammatical?

There is no reason *a priori* that *parle* is different from *mange*.

# Nature of Language

## Linguistics as data analysis

But also a little bit more:

- *[purpose of linguists] is not simply to account for all utterances which comprise his corpus [but rather] to account for utterances which are not in his corpus at a given time [Hockett(1948)]*

- *[introspective judgements] when it agrees [with corpus data non-corpus data is] superfluous; when it disagrees [with corpus data non-corpus data is] obnoxious [Hockett(1964)]*

## Infinite nature of the language

- Recursion
- Competence/performance example
- Nevertheless, learning results (in Gold sense) for CFG, Categorial Grammars...

# Nature of Language

## Linguistics as data analysis

But also a little bit more:

- *[purpose of linguists] is not simply to account for all utterances which comprise his corpus [but rather] to account for utterances which are not in his corpus at a given time [Hockett(1948)]*

- *[introspective judgements] when it agrees [with corpus data non-corpus data is] superfluous; when it disagrees [with corpus data non-corpus data is] obnoxious [Hockett(1964)]*

## Infinite nature of the language

- Recursion
- Competence/performance example
- Nevertheless, learning results (in Gold sense) for CFG, Categorial Grammars. . .

## Some other criticisms

- About representativity (biased and incomplete), Cf frequencies
- Pseudo-procedure methodology [Abercrombie(1965)]: manual analyses and mistakes

# Some other criticisms

- About representativity (biased and incomplete), Cf frequencies
- Pseudo-procedure methodology [Abercrombie(1965)]: manual analyses and mistakes

# But intrinsic advantages

## When introspection fails

- Phonology
- Acquisition
- Languages variation (regional expressions, sociolects, register)
- Data can be observed and checked
- Usefulness of the frequency measure (unavailable from introspection)

## Even when introspection can happen

- Failing to find some sentences or grammatical constructions in a corpus may also be an interesting comment on their frequency
- Introspection lacks systematicity (mistakes can also occur during introspection)
- "Corpus is a more powerful methodology from the point of view of the scientific method" [Leech(1992)]

# But intrinsic advantages

## When introspection fails

- Phonology
- Acquisition
- Languages variation (regional expressions, sociolects, register)
- Data can be observed and checked
- Usefulness of the frequency measure (unavailable from introspection)

## Even when introspection can happen

- Failing to find some sentences or grammatical constructions in a corpus may also be an interesting comment on their frequency
- Introspection lacks systematicity (mistakes can also occur during introspection)
- "Corpus is a more powerful methodology from the point of view of the scientific method" [Leech(1992)]

# But intrinsic advantages

## Mixing the two approaches

- Corpus of introspection examples (Cf Partee, famous examples, etc.).
- “What look like the cute examples and arbitrary infatuations of linguists often, though not always, represent a distillation of important and wide-ranging issues”(Martin Kay 2006) vs “[the non-corpus linguist] sits in a deep soft armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, 'Wow, what a neat fact!', grab his pencil, and writes something down... having come still no closer to knowing what language is really like.” [Fillmore(1992)]

## Benefits

- Evidence of usefulness of quantitative data: in the success of building efficient tools (taggers)
- Today, NLP applications?

## End of the debate?

Nevertheless, criticisms allowed and still allow to have an idea of *what a corpus should be* and *how to work with a corpus*.

# But intrinsic advantages

## Mixing the two approaches

- Corpus of introspection examples (Cf Partee, famous examples, etc.).
- "What look like the cute examples and arbitrary infatuations of linguists often, though not always, represent a distillation of important and wide-ranging issues"(Martin Kay 2006) vs "[the non-corpus linguist] sits in a deep soft armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, 'Wow, what a neat fact!', grab his pencil, and writes something down... having come still no closer to knowing what language is really like." [Fillmore(1992)]

## Benefits

- Evidence of usefulness of quantitative data: in the success of building efficient tools (taggers)
- Today, NLP applications?

## End of the debate?

Nevertheless, criticisms allowed and still allow to have an idea of *what a corpus should be* and *how to work with a corpus*.

# But intrinsic advantages

## Mixing the two approaches

- Corpus of introspection examples (Cf Partee, famous examples, etc.).
- “What look like the cute examples and arbitrary infatuations of linguists often, though not always, represent a distillation of important and wide-ranging issues”(Martin Kay 2006) vs “[the non-corpus linguist] sits in a deep soft armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, 'Wow, what a neat fact!', grab his pencil, and writes something down... having come still no closer to knowing what language is really like.” [Fillmore(1992)]

## Benefits

- Evidence of usefulness of quantitative data: in the success of building efficient tools (taggers)
- Today, NLP applications?

## End of the debate?

Nevertheless, criticisms allowed and still allow to have an idea of *what a corpus should be* and *how to work with a corpus*.

# Characteristics of a corpus

## Which Prerequisites to Work with a Corpus?

### Escaping from the pseudo-procedure criticism

- Data are *exploitable* by computers
- Date are reliable (at least with a measurable reliability)
  - OK for some automatic annotations (POS tagging)
  - Still pseudo-procedure for other non-annotated corpora (NP recognition)
- Enable searching, sorting, computing. . .

### Some examples

Frequency, Concordancer

# Characteristics of a corpus

## Which Prerequisites to Work with a Corpus?

### Escaping from the pseudo-procedure criticism

- Data are *exploitable* by computers
- Date are reliable (at least with a measurable reliability)
  - OK for some automatic annotations (POS tagging)
  - Still pseudo-procedure for other non-annotated corpora (NP recognition)
- Enable searching, sorting, computing. . .

### Some examples

Frequency, Concordancer

# Characteristics of a corpus

## Designed for a usage

- [McEnery and Wilson(1996)]

  > *If you think the language is finite, then your interpretation of the findings in a corpus may reflect that - if we can change the interpretation of the findings in a corpus to match the verities Chomsky revealed, then the natural data provided by the corpus can be a rich and powerful tool for the linguist. But we must understand what we are doing when we are looking in a corpus and building one. (...) A corpus and an introspection-based approach to linguistic are not mutually exclusive. In a very real sense they can be gainfully viewed as being complementary.*

- Difference between linguistic goals and NLP goals (in particular, as NLP tool user, on non introspective data)

# What is a Corpus

## As for now

- Data readable from computer
- But also:
  - Sampling and representativity
  - Finite size
  - Reference to a standard

# Necessity of Sampling

- **Infinite language** (except for dead languages, texts from a single author...). Hence scarcely a finite set ⇒ sampling
- **Sampling and representativity** Cf. Chomsky criticism
- **Aim** to get the greatest representativity inside a variety under study

## What does it mean?

- What are the *genres* of texts, the *number* of texts, what particular texts, how to select some abstracts, of what size.
- "Representativeness refers to the extent to which a sample includes the full range of variability in a population" [Biber(1993)]

## Consequence

Hence, a corpus has to be evaluated on its ability to include:

- The possible genres of texts of a language
- The linguistic variety inside a language

# Necessity of Sampling

- **Infinite language** (except for dead languages, texts from a single author...). Hence scarcely a finite set ⇒ sampling
- **Sampling and representativity** Cf. Chomsky criticism
- **Aim** to get the greatest representativity inside a variety under study

## What does it mean?

- What are the *genres* of texts, the *number* of texts, what particular texts, how to select some abstracts, of what size.
- "Representativeness refers to the extent to which a sample includes the full range of variability in a population" [Biber(1993)]

## Consequence

Hence, a corpus has to be evaluated on its ability to include:

- The possible genres of texts of a language
- The linguistic variety inside a language

# Necessity of Sampling

- **Infinite language** (except for dead languages, texts from a single author...). Hence scarcely a finite set $\Rightarrow$ sampling
- **Sampling and representativity** Cf. Chomsky criticism
- **Aim** to get the greatest representativity inside a variety under study

## What does it mean?

- What are the *genres* of texts, the *number* of texts, what particular texts, how to select some abstracts, of what size.
- "Representativeness refers to the extent to which a sample includes the full range of variability in a population" [Biber(1993)]

## Consequence

Hence, a corpus has to be evaluated on its ability to include:

- The possible genres of texts of a language
- The linguistic variety inside a language

## Necessity of Sampling

- **Infinite language** (except for dead languages, texts from a single author...). Hence scarcely a finite set ⇒ sampling
- **Sampling and representativity** Cf. Chomsky criticism
- **Aim** to get the greatest representativity inside a variety under study

### What does it mean?

- What are the *genres* of texts, the *number* of texts, what particular texts, how to select some abstracts, of what size.
- "Representativeness refers to the extent to which a sample includes the full range of variability in a population" [Biber(1993)]

### Consequence

Hence, a corpus has to be evaluated on its ability to include:

- The possible genres of texts of a language
- The linguistic variety inside a language

# Necessity of Sampling

- **Infinite language** (except for dead languages, texts from a single author...). Hence scarcely a finite set ⇒ sampling
- **Sampling and representativity** Cf. Chomsky criticism
- **Aim** to get the greatest representativity inside a variety under study

## What does it mean?

- What are the *genres* of texts, the *number* of texts, what particular texts, how to select some abstracts, of what size.
- "Representativeness refers to the extent to which a sample includes the full range of variability in a population" [Biber(1993)]

## Consequence

Hence, a corpus has to be evaluated on its ability to include:

- The possible genres of texts of a language
- The linguistic variety inside a language

# Representativity: an example

**Despite genres are not defined on a linguistic basis, there are big linguistic differences between genres**

Example: Brown corpus

- Classification
- Modalities example (`modal-ratio.py`),
- Token/type ratio example (`token-type-ratio.py`)

A corpus has to ensure, or at least to justify, representativity.

# Size of a Corpus

- Some corpus always increases (new data are added). It gives an idea of the evolution
- But, because of the changing size, sampling and representativity are more difficult to ensure.
- Examples: NLTK Corpora:
  - brown corpus (1 000 000 words)
  - treebank 10% of the WSJ section of the Penn Treebank (1 000 000 words)

# Machine-Readable Corpus

## Summary

Advantages: querying and computing

Examples: raw text pattern matching (grep)

## Consequences

- But what about all the flexed forms? (→ concordancer example )
- What about all the Ns? (→ Brown corpus example (POS tagging))
- What about all the NPs ? (→ treebank example)

⇒ Crucial role of *annotation*

# Other Examples

French Corpora

# Annotations

What we already know:

- Raw text $\longrightarrow$ regexp, concordancer
- Unstructured annotations (glues with the word)
- Annotations and trees
- Other (semantics, time and aspects, pragmatics, etc.)

Annotations are increasingly complex to read and to create

# Principles

Aims of annotations:

1. Enhancing (raw) data with relevant linguistic annotations (relevant with what respect? Depends on the usage)
2. Reusability: annotation cost (time, money), correction and accuracy. But:
    - Juridical problems (text ownership)
    - Technical problems (formats, recommendations, standardisation)

# Principles

Aims of annotations:

1. Enhancing (raw) data with relevant linguistic annotations (relevant with what respect? Depends on the usage)
2. Reusability: annotation cost (time, money), correction and accuracy. But:
   - Juridical problems (text ownership)
   - Technical problems (formats, recommendations, standardisation)

# Principles

Aims of annotations:

1. Enhancing (raw) data with relevant linguistic annotations (relevant with what respect? Depends on the usage)
2. Reusability: annotation cost (time, money), correction and accuracy. But:
   - Juridical problems (text ownership)
   - Technical problems (formats, recommendations, standardisation)

# Principles

Aims of annotations:

1. Enhancing (raw) data with relevant linguistic annotations (relevant with what respect? Depends on the usage)
2. Reusability: annotation cost (time, money), correction and accuracy. But:
   - Juridical problems (text ownership)
   - Technical problems (formats, recommendations, standardisation)

# Principles

Hence some rules [Leech(1993)] :

1. From an annotated corpus, it should always be possible to come back to initial data (example BC). Remark: may be difficult after normalisation ("l'arbre" → "le arbre", etc.)
2. Annotations should be extractable from the text in order to save them at another place (DB), to present them in another way, etc.
3. The annotation procedure should be documented (example Brown Corpus annotation guide , Penn Tree Bank annotation guide )
4. Mention should be made of the annotator and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected. . . )
5. Annotation schemas should be as independent as possible on formalisms
6. No annotation schema should consider itself a standard (it possibly becomes one)

# Principles

Hence some rules [Leech(1993)] :

1. From an annotated corpus, it should always be possible to come back to initial data (example BC). Remark: may be difficult after normalisation ("l'arbre" → "le arbre", etc.)
2. Annotations should be extractable from the text in order to save them at another place (DB), to present them in another way, etc.
3. The annotation procedure should be documented (example Brown Corpus annotation guide , Penn Tree Bank annotation guide )
4. Mention should be made of the annotator and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected. . . )
5. Annotation schemas should be as independent as possible on formalisms
6. No annotation schema should consider itself a standard (it possibly becomes one)

# Principles

Hence some rules [Leech(1993)] :

1. From an annotated corpus, it should always be possible to come back to initial data (example BC). Remark: may be difficult after normalisation ("l'arbre" → "le arbre", etc.)
2. Annotations should be extractable from the text in order to save them at another place (DB), to present them in another way, etc.
3. The annotation procedure should be documented (example Brown Corpus annotation guide , Penn Tree Bank annotation guide )
4. Mention should be made of the annotator and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected. . . )
5. Annotation schemas should be as independent as possible on formalisms
6. No annotation schema should consider itself a standard (it possibly becomes one)

# Principles

Hence some rules [Leech(1993)] :

1. From an annotated corpus, it should always be possible to come back to initial data (example BC). Remark: may be difficult after normalisation ("l'arbre" → "le arbre", etc.)
2. Annotations should be extractable from the text in order to save them at another place (DB), to present them in another way, etc.
3. The annotation procedure should be documented (example Brown Corpus annotation guide , Penn Tree Bank annotation guide )
4. Mention should be made of the annotator and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected...)
5. Annotation schemas should be as independent as possible on formalisms
6. No annotation schema should consider itself a standard (it possibly becomes one)

# Principles

Hence some rules [Leech(1993)] :

1. From an annotated corpus, it should always be possible to come back to initial data (example BC). Remark: may be difficult after normalisation ("l'arbre" → "le arbre", etc.)
2. Annotations should be extractable from the text in order to save them at another place (DB), to present them in another way, etc.
3. The annotation procedure should be documented (example Brown Corpus annotation guide , Penn Tree Bank annotation guide )
4. Mention should be made of the annotator and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected. . . )
5. Annotation schemas should be as independent as possible on formalisms
6. No annotation schema should consider itself a standard (it possibly becomes one)

# Principles

Hence some rules [Leech(1993)] :

1. From an annotated corpus, it should always be possible to come back to initial data (example BC). Remark: may be difficult after normalisation ("l'arbre" → "le arbre", etc.)
2. Annotations should be extractable from the text in order to save them at another place (DB), to present them in another way, etc.
3. The annotation procedure should be documented (example Brown Corpus annotation guide , Penn Tree Bank annotation guide )
4. Mention should be made of the annotator and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected. . . )
5. Annotation schemas should be as independent as possible on formalisms
6. No annotation schema should consider itself a standard (it possibly becomes one)

# What Kind of Annotation?

- Everything which is potentially interesting!
- Hence, depends on the application

## Non linguistic annotation

- Genre, variety (french from France, from North, from South, age and sex of the speaker, publication date, etc.)
- Form. Example:
  - For a written text: italic, bold letters (example: Alice)
  - But also line changes (verse)
  - For dialogs: sentences pronounced at the same time, etc.

# What Kind of Annotation?

- Everything which is potentially interesting!
- Hence, depends on the application

## Non linguistic annotation

- Genre, variety (french from France, from North, from South, age and sex of the speaker, publication date, etc.)
- Form. Example:
  - For a written text: italic, bold letters (example: Alice)
  - But also line changes (verse)
  - For dialogs: sentences pronounced at the same time, etc.

# What Kind of Annotation?

- Everything which is potentially interesting!
- Hence, depends on the application

## Non linguistic annotation

- Genre, variety (french from France, from North, from South, age and sex of the speaker, publication date, etc.)
- Form. Example:
  - For a written text, italic, bold letters (example: Alice)
  - But also line changes (verse)
  - For dialogs, sentences pronounced at the same time, etc.

# What Kind of Annotation?

- Everything which is potentially interesting!
- Hence, depends on the application

## Non linguistic annotation

- Genre, variety (french from France, from North, from South, age and sex of the speaker, publication date, etc.)
- Form. Example:
  - For a written text, italic, bold letters (example: Alice)
  - But also line changes (verse)
  - For dialogs, sentences pronounced at the same time, etc.

# What Kind of Annotation?

- Everything which is potentially interesting!
- Hence, depends on the application

## Non linguistic annotation

- Genre, variety (french from France, from North, from South, age and sex of the speaker, publication date, etc.)
- Form. Example:
  - For a written text, italic, bold letters (example: Alice)
  - But also line changes (verse)
  - For dialogs, sentences pronounced at the same time, etc.

# What Kind of Annotation?

- Everything which is potentially interesting!
- Hence, depends on the application

## Non linguistic annotation

- Genre, variety (french from France, from North, from South, age and sex of the speaker, publication date, etc.)
- Form. Example:
  - For a written text, italic, bold letters (example: Alice)
  - But also line changes (verse)
  - For dialogs, sentences pronounced at the same time, etc.

# What Kind of Annotation?

- Everything which is potentially interesting!
- Hence, depends on the application

## Non linguistic annotation

- Genre, variety (french from France, from North, from South, age and sex of the speaker, publication date, etc.)
- Form. Example:
  - For a written text, italic, bold letters (example: Alice)
  - But also line changes (verse)
  - For dialogs, sentences pronounced at the same time, etc.

## What Kind of Annotation?

- Everything which is potentially interesting!
- Hence, depends on the application

### Non linguistic annotation

- Genre, variety (french from France, from North, from South, age and sex of the speaker, publication date, etc.)
- Form. Example:
  - For a written text, italic, bold letters (example: Alice)
  - But also line changes (verse)
  - For dialogs, sentences pronounced at the same time, etc.

# What Kind of Annotation?

## Linguistic annotation

- POS tags, brown corpus example (tags inside text). cf `brown-tags.py`, `brown-tags-freq.py`
- POS tagging (lemonde talana example) : first stage. (before parsing, semantic annotation). With respect to raw text, gives homographs disambiguation (corpus lemonde example: “allemands” adj/n, “exécutif” adj/n, “parlé”, “pensé” NC/V. `french-tags.py`, `french-tags-freq.py`
- Corpus lemonde tags example (tei xml tags example)
- Treebank tags example (inside text, bracketting). `np-treebank.py`
- Different tag name example:
  - Example: “so that” brown (line 36) “won't (line 1836), “so that” (treebank num 3292), “won't” (treebank num 3292).
    See brown et treebank annotation guide (`chunk-parse-treebank.py`)

Remark: unification and standardisation (EAGLES). Three annotation levels:

# What Kind of Annotation?

## Linguistic annotation

- POS tags, brown corpus example (tags inside text). cf `brown-tags.py`, `brown-tags-freq.py`
- POS tagging (lemonde talana example) : first stage. (before parsing, semantic annotation). With respect to raw text, gives homographs disambiguation (corpus lemonde example: “allemands” adj/n, “exécutif” adj/n, “parlé”, “pensé” NC/V. `french-tags.py`, `french-tags-freq.py`
- Corpus lemonde tags example (tei xml tags example)
- Treebank tags example (inside text, bracketting). `np-treebank.py`
- Different tag name example:
  - Example: “so that” brown (line 36) “won't (line 1836), “so that” (treebank num 3292), “won't” (treebank num 3292).
    See brown et treebank annotation guide (`chunk-parse-treebank.py`)

Remark: unification and standardisation (EAGLES). Three annotation levels:

# What Kind of Annotation?

## Linguistic annotation

- POS tags, brown corpus example (tags inside text). cf `brown-tags.py`, `brown-tags-freq.py`
- POS tagging (lemonde talana example) : first stage. (before parsing, semantic annotation). With respect to raw text, gives homographs disambiguation (corpus lemonde example: "allemands" adj/n, "exécutif" adj/n, "parlé", "pensé" NC/V. `french-tags.py`, `french-tags-freq.py`
- Corpus lemonde tags example (tei xml tags example)
- Treebank tags example (inside text, bracketting). `np-treebank.py`
- Different tag name example:
  - Example: "so that" brown (line 36) "won't (line 1836), "so that" (treebank num 3292), "won't" (treebank num 3292).
    See brown et treebank annotation guide (`chunk-parse-treebank.py`)

Remark: unification and standardisation (EAGLES). Three annotation levels:

# What Kind of Annotation?

## Linguistic annotation

- POS tags, brown corpus example (tags inside text). cf `brown-tags.py`, `brown-tags-freq.py`
- POS tagging (lemonde talana example) : first stage. (before parsing, semantic annotation). With respect to raw text, gives homographs disambiguation (corpus lemonde example: “allemands” adj/n, “exécutif” adj/n, “parlé”, “pensé” NC/V. `french-tags.py`, `french-tags-freq.py`
- Corpus lemonde tags example (tei xml tags example)
- Treebank tags example (inside text, bracketting). `np-treebank.py`
- Different tag name example:
  - Example: “so that” brown (line 36) “won't (line 1836), “so that” (treebank num 3292), “won't” (treebank num 3292).
    See brown et treebank annotation guide (`chunk-parse-treebank.py`)

Remark: unification and standardisation (EAGLES). Three annotation levels:

# What Kind of Annotation?

## Linguistic annotation

- POS tags, brown corpus example (tags inside text). cf `brown-tags.py`, `brown-tags-freq.py`
- POS tagging (lemonde talana example) : first stage. (before parsing, semantic annotation). With respect to raw text, gives homographs disambiguation (corpus lemonde example: "allemands" adj/n, "exécutif" adj/n, "parlé", "pensé" NC/V. `french-tags.py`, `french-tags-freq.py`
- Corpus lemonde tags example (tei xml tags example)
- Treebank tags example (inside text, bracketting). `np-treebank.py`
- Different tag name example:
  - Example: "so that" brown (line 36) "won't (line 1836), "so that" (treebank num 3292), "won't" (treebank num 3292).
    See brown et treebank annotation guide (`chunk-parse-treebank.py`)
- Remark: unification and standardisation (EAGLES). Three annotation levels:

# What Kind of Annotation?

## Linguistic annotation

- POS tags, brown corpus example (tags inside text). cf `brown-tags.py`, `brown-tags-freq.py`
- POS tagging (lemonde talana example) : first stage. (before parsing, semantic annotation). With respect to raw text, gives homographs disambiguation (corpus lemonde example: "allemands" adj/n, "exécutif" adj/n, "parlé", "pensé" NC/V. `french-tags.py`, `french-tags-freq.py`
- Corpus lemonde tags example (tei xml tags example)
- Treebank tags example (inside text, bracketting). `np-treebank.py`
- Different tag name example:
  - Example: "so that" brown (line 36) "won't (line 1836), "so that" (treebank num 3292), "won't" (treebank num 3292).
    See brown et treebank annotation guide (`chunk-parse-treebank.py`)

Remark: unification and standardisation (EAGLES). Three annotation levels:

- Mandatory tags
- Recommended tags
- Optional tags

# What Kind of Annotation?

## Linguistic annotation

- POS tags, brown corpus example (tags inside text). cf `brown-tags.py`, `brown-tags-freq.py`
- POS tagging (lemonde talana example) : first stage. (before parsing, semantic annotation). With respect to raw text, gives homographs disambiguation (corpus lemonde example: “allemands” adj/n, “exécutif” adj/n, “parlé”, “pensé” NC/V. `french-tags.py`, `french-tags-freq.py`
- Corpus lemonde tags example (tei xml tags example)
- Treebank tags example (inside text, bracketting). `np-treebank.py`
- Different tag name example:
  - Example: “so that” brown (line 36) “won't (line 1836), “so that” (treebank num 3292), “won't” (treebank num 3292).
    See brown et treebank annotation guide (`chunk-parse-treebank.py`)

Remark: unification and standardisation (EAGLES). Three annotation levels:

- Mandatory tags
- Recommended tags
- Optional tags

# What Kind of Annotation?

## Linguistic annotation

- POS tags, brown corpus example (tags inside text). cf `brown-tags.py`, `brown-tags-freq.py`
- POS tagging (lemonde talana example) : first stage. (before parsing, semantic annotation). With respect to raw text, gives homographs disambiguation (corpus lemonde example: “allemands” adj/n, “exécutif” adj/n, “parlé”, “pensé” NC/V. `french-tags.py`, `french-tags-freq.py`
- Corpus lemonde tags example (tei xml tags example)
- Treebank tags example (inside text, bracketting). `np-treebank.py`
- Different tag name example:
  - Example: “so that” brown (line 36) “won’t (line 1836), “so that” (treebank num 3292), “won’t” (treebank num 3292).
    See brown et treebank annotation guide (`chunk-parse-treebank.py`)

Remark: unification and standardisation (EAGLES). Three annotation levels:

- Mandatory tags
- Recommended tags
- Optional tags

# What Kind of Annotation?

## Linguistic annotation

- POS tags, brown corpus example (tags inside text). cf `brown-tags.py`, `brown-tags-freq.py`
- POS tagging (lemonde talana example) : first stage. (before parsing, semantic annotation). With respect to raw text, gives homographs disambiguation (corpus lemonde example: “allemands” adj/n, “exécutif” adj/n, “parlé”, “pensé” NC/V. `french-tags.py`, `french-tags-freq.py`
- Corpus lemonde tags example (tei xml tags example)
- Treebank tags example (inside text, bracketting). `np-treebank.py`
- Different tag name example:
  - Example: “so that” brown (line 36) “won’t (line 1836), “so that” (treebank num 3292), “won’t” (treebank num 3292).
    See brown et treebank annotation guide (`chunk-parse-treebank.py`)

  Remark: unification and standardisation (EAGLES). Three annotation levels:
  - Mandatory tags
  - Recommended tags
  - Optional tags

# What Kind of Annotation?

## Linguistic annotation

- POS tags, brown corpus example (tags inside text). cf `brown-tags.py`, `brown-tags-freq.py`
- POS tagging (lemonde talana example) : first stage. (before parsing, semantic annotation). With respect to raw text, gives homographs disambiguation (corpus lemonde example: “allemands” adj/n, “exécutif” adj/n, “parlé”, “pensé” NC/V. `french-tags.py`, `french-tags-freq.py`
- Corpus lemonde tags example (tei xml tags example)
- Treebank tags example (inside text, bracketting). `np-treebank.py`
- Different tag name example:
  - Example: “so that” brown (line 36) “won’t (line 1836), “so that” (treebank num 3292), “won’t” (treebank num 3292).
    See brown et treebank annotation guide (`chunk-parse-treebank.py`)

Remark: unification and standardisation (EAGLES). Three annotation levels:

  - Mandatory tags
  - Recommended tags
  - Optional tags

# What Kind of Annotation?

## Linguistic annotation

- POS tags, brown corpus example (tags inside text). cf `brown-tags.py`, `brown-tags-freq.py`
- POS tagging (lemonde talana example) : first stage. (before parsing, semantic annotation). With respect to raw text, gives homographs disambiguation (corpus lemonde example: “allemands” adj/n, “exécutif” adj/n, “parlé”, “pensé” NC/V. `french-tags.py`, `french-tags-freq.py`
- Corpus lemonde tags example (tei xml tags example)
- Treebank tags example (inside text, bracketting). `np-treebank.py`
- Different tag name example:
  - Example: “so that” brown (line 36) “won't (line 1836), “so that” (treebank num 3292), “won't” (treebank num 3292).
    See brown et treebank annotation guide (`chunk-parse-treebank.py`)

Remark: unification and standardisation (EAGLES). Three annotation levels:

- Mandatory tags
- Recommended tags
- Optional tags

# What Kind of Annotation?

## Various informations:

- Lemmatisation (example: concordancer ), cf talana/lemonde corpus
- Parsing. Usually, a constituent tree. But, again:
  - Differences between tagsets, differences between structures (trees or chunks), cf example and Penn Tree Bank annotation guide
  - Differences between formalisms (dependency grammars (trees, but unordered leaves), or graphs, not trees (Word grammars [Hudson(1990)]). cf DGA example (data ,visualization
- Semantics:
  - Specific meaning of a word (among all its possible meanings) in a text, senseval/wordnet example. (Application: disambiguation, translation, etc.)
    Some of the difficulties of annotating: not automatic, definition of the tagset, appropriateness tag [Véronis(1998)]
  - Relations (agent, patient, reference string)

# What Kind of Annotation?

## Various informations:

- Lemmatisation (example: concordancer ), cf talana/lemonde corpus
- Parsing. Usually, a constituent tree. But, again:
  - Differences between tagsets, differences between structures (trees or chunks), cf example and Penn Tree Bank annotation guide
  - Differences between formalisms (dependency grammars (trees, but unordered leaves), or graphs, not trees (Word grammars [Hudson(1990)]). cf DGA example (data ,visualization
- Semantics:
  - Specific meaning of a word (among all its possible meanings) in a text, senseval/wordnet example. (Application: disambiguation, translation, etc.)
    Some of the difficulties of annotating: not automatic, definition of the tagset, appropriateness tag [Véronis(1998)]
  - Relations (agent, patient, reference string)

# What Kind of Annotation?

## Various informations:

- Lemmatisation (example: concordancer ), cf talana/lemonde corpus
- Parsing. Usually, a constituent tree. But, again:
  - Differences between tagsets, differences between structures (trees or chunks), cf example and Penn Tree Bank annotation guide
  - Differences between formalisms (dependency grammars (trees, but unordered leaves), or graphs, not trees (Word grammars [Hudson(1990)]). cf DGA example (data ,visualization
- Semantics:
  - Specific meaning of a word (among all its possible meanings) in a text, senseval/wordnet example. (Application: disambiguation, translation, etc.)
    Some of the difficulties of annotating: not automatic, definition of the tagset, appropriateness tag [Véronis(1998)]
  - Relations (agent, patient, reference string)

# What Kind of Annotation?

## Various informations:

- Lemmatisation (example: concordancer ), cf talana/lemonde corpus
- Parsing. Usually, a constituent tree. But, again:
    - Differences between tagsets, differences between structures (trees or chunks), cf example and Penn Tree Bank annotation guide
    - Differences between formalisms (dependency grammars (trees, but unordered leaves), or graphs, not trees (Word grammars [Hudson(1990)]). cf DGA example (data ,visualization
- Semantics:
    - Specific meaning of a word (among all its possible meanings) in a text, senseval/wordnet example. (Application: disambiguation, translation, etc.)
      Some of the difficulties of annotating: not automatic, definition of the tagset, appropriateness tag [Véronis(1998)]
    - Relations (agent, patient, reference string)

# What Kind of Annotation?

## Various informations:

- Lemmatisation (example: concordancer ), cf talana/lemonde corpus
- Parsing. Usually, a constituent tree. But, again:
  - Differences between tagsets, differences between structures (trees or chunks), cf example and Penn Tree Bank annotation guide
  - Differences between formalisms (dependency grammars (trees, but unordered leaves), or graphs, not trees (Word grammars [Hudson(1990)]). cf DGA example (data ,visualization
- Semantics:
  - Specific meaning of a word (among all its possible meanings) in a text, senseval/wordnet example. (Application: disambiguation, translation, etc.)
    Some of the difficulties of annotating: not automatic, definition of the tagset, appropriateness tag [Véronis(1998)]
  - Relations (agent, patient, reference string)

# What Kind of Annotation?

## Various informations:

- Lemmatisation (example: concordancer ), cf talana/lemonde corpus
- Parsing. Usually, a constituent tree. But, again:
  - Differences between tagsets, differences between structures (trees or chunks), cf example and Penn Tree Bank annotation guide
  - Differences between formalisms (dependency grammars (trees, but unordered leaves), or graphs, not trees (Word grammars [Hudson(1990)]). cf DGA example (data ,visualization
- Semantics:
  - Specific meaning of a word (among all its possible meanings) in a text, senseval/wordnet example. (Application: disambiguation, translation, etc.)
    Some of the difficulties of annotating: not automatic, definition of the tagset, appropriateness tag [Véronis(1998)]
  - Relations (agent, patient, reference string)

# What Kind of Annotation?

## Various informations:

- Lemmatisation (example: concordancer ), cf talana/lemonde corpus
- Parsing. Usually, a constituent tree. But, again:
  - Differences between tagsets, differences between structures (trees or chunks), cf example and Penn Tree Bank annotation guide
  - Differences between formalisms (dependency grammars (trees, but unordered leaves), or graphs, not trees (Word grammars [Hudson(1990)]). cf DGA example (data ,visualization
- Semantics:
  - Specific meaning of a word (among all its possible meanings) in a text, senseval/wordnet example. (Application: disambiguation, translation, etc.)
    Some of the difficulties of annotating: not automatic, definition of the tagset, appropriateness tag [Véronis(1998)]
  - Relations (agent, patient, reference string)

# What Kind of Annotation?

## Various informations:

- Lemmatisation (example: concordancer ), cf talana/lemonde corpus
- Parsing. Usually, a constituent tree. But, again:
  - Differences between tagsets, differences between structures (trees or chunks), cf example and Penn Tree Bank annotation guide
  - Differences between formalisms (dependency grammars (trees, but unordered leaves), or graphs, not trees (Word grammars [Hudson(1990)]). cf DGA example (data ,visualization
- Semantics:
  - Specific meaning of a word (among all its possible meanings) in a text, senseval/wordnet example. (Application: disambiguation, translation, etc.)
    Some of the difficulties of annotating: not automatic, definition of the tagset, appropriateness tag [Véronis(1998)]
  - Relations (agent, patient, reference string)

# Obstacles and Difficulties

## Technical difficulties

- Character encoding
- Tokenisation
- Orthography

## The annotation process

- Accurate annotation (manual) $\longrightarrow$ time, cost
- But even if manual annotation, not easy (semantics, Co-reference. . . (dependency on annotator))
- Automatic annotation: possible for some linguistic levels (morphosyntax, parsing. . . )
- Some tools: Brill pos tagger, treetagger $\longrightarrow$ robustness
- Lemmatisation: automatic lemmatization, unknown words, etc.

# Obstacles and Difficulties

## Technical difficulties

- Character encoding
- Tokenisation
- Orthography

## The annotation process

- Accurate annotation (manual) → time, cost
- But even if manual annotation, not easy (semantics, Co-reference... (dependency on annotator))
- Automatic annotation: possible for some linguistic levels (morphosyntax, parsing...)
- Some tools: Brill pos tagger, treetagger → robustness
- Lemmatisation: automatic lemmatization, unknown words, etc.

# Obstacles and Difficulties

## Technical difficulties

- Character encoding
- Tokenisation
- Orthography

## The annotation process

- Accurate annotation (manual) $\longrightarrow$ time, cost
- But even if manual annotation, not easy (semantics, Co-reference. . . (dependency on annotator))
- Automatic annotation: possible for some linguistic levels (morphosyntax, parsing. . . )
- Some tools: Brill pos tagger, treetagger $\longrightarrow$ robustness
- Lemmatisation: automatic lemmatization, unknown words, etc.

# Obstacles and Difficulties

## Technical difficulties

- Character encoding
- Tokenisation
- Orthography

## The annotation process

- Accurate annotation (manual) $\longrightarrow$ time, cost
- But even if manual annotation, not easy (semantics, Co-reference. . . (dependency on annotator))
- Automatic annotation: possible for some linguistic levels (morphosyntax, parsing. . . )
- Some tools: Brill pos tagger, treetagger $\longrightarrow$ robustness
- Lemmatisation: automatic lemmatization, unknown words, etc.

# Obstacles and Difficulties

## Technical difficulties

- Character encoding
- Tokenisation
- Orthography

## The annotation process

- Accurate annotation (manual) $\rightarrow$ time, cost
- But even if manual annotation, not easy (semantics, Co-reference...(dependency on annotator))
- Automatic annotation: possible for some linguistic levels (morphosyntax, parsing...)
- Some tools: Brill pos tagger, treetagger $\rightarrow$ robustness
- Lemmatisation: automatic lemmatization, unknown words, etc.

# Obstacles and Difficulties

## Technical difficulties

- Character encoding
- Tokenisation
- Orthography

## The annotation process

- Accurate annotation (manual) → time, cost
- But even if manual annotation, not easy (semantics, Co-reference. . . (dependency on annotator))
- Automatic annotation: possible for some linguistic levels (morphosyntax, parsing. . . )
- Some tools: Brill pos tagger, treetagger → robustness
- Lemmatisation: automatic lemmatization, unknown words, etc.

# Obstacles and Difficulties

## Technical difficulties

- Character encoding
- Tokenisation
- Orthography

## The annotation process

- Accurate annotation (manual) $\longrightarrow$ time, cost
- But even if manual annotation, not easy (semantics, Co-reference. . . (dependency on annotator))
- Automatic annotation: possible for some linguistic levels (morphosyntax, parsing. . . )
- Some tools: Brill pos tagger, treetagger $\longrightarrow$ robustness
- Lemmatisation: automatic lemmatization, unknown words, etc.

# Obstacles and Difficulties

## Technical difficulties

- Character encoding
- Tokenisation
- Orthography

## The annotation process

- Accurate annotation (manual) $\longrightarrow$ time, cost
- But even if manual annotation, not easy (semantics, Co-reference. . . (dependency on annotator))
- Automatic annotation: possible for some linguistic levels (morphosyntax, parsing. . . )
- Some tools: Brill pos tagger, treetagger $\longrightarrow$ robustness
- Lemmatisation: automatic lemmatization, unknown words, etc.

# Obstacles and Difficulties

## Technical difficulties

- Character encoding
- Tokenisation
- Orthography

## The annotation process

- Accurate annotation (manual) → time, cost
- But even if manual annotation, not easy (semantics, Co-reference. . . (dependency on annotator))
- Automatic annotation: possible for some linguistic levels (morphosyntax, parsing. . . )
- Some tools: Brill pos tagger, treetagger → robustness
- Lemmatisation: automatic lemmatization, unknown words, etc.

# Obstacles and Difficulties

## Technical difficulties

- Character encoding
- Tokenisation
- Orthography

## The annotation process

- Accurate annotation (manual) → time, cost
- But even if manual annotation, not easy (semantics, Co-reference. . . (dependency on annotator))
- Automatic annotation: possible for some linguistic levels (morphosyntax, parsing. . . )
- Some tools: Brill pos tagger, treetagger → robustness
- Lemmatisation: automatic lemmatization, unknown words, etc.

# Obstacles and Difficulties (cont'd)

## The annotation process (cont'd)

- Annotation and linguistic theory:
  - Constituent vs dependency grammars
  - Subject/agent relations
- Makes NLP tool evaluation difficult (the background of annotation may not fit the tool's one)

## Reusability

cf normalisation

# Obstacles and Difficulties (cont'd)

## The annotation process (cont'd)

- Annotation and linguistic theory:
  - Constituent vs dependency grammars
  - Subject/agent relations
- Makes NLP tool evaluation difficult (the background of annotation may not fit the tool's one)

## Reusability

cf normalisation

# Obstacles and Difficulties (cont'd)

## The annotation process (cont'd)

- Annotation and linguistic theory:
  - Constituent vs dependency grammars
  - Subject/agent relations
- Makes NLP tool evaluation difficult (the background of annotation may not fit the tool's one)

## Reusability

cf normalisation

# Obstacles and Difficulties (cont'd)

## The annotation process (cont'd)

- Annotation and linguistic theory:
  - Constituent vs dependency grammars
  - Subject/agent relations
- Makes NLP tool evaluation difficult (the background of annotation may not fit the tool's one)

## Reusability

cf normalisation

# Obstacles and Difficulties (cont'd)

## The annotation process (cont'd)

- Annotation and linguistic theory:
  - Constituent vs dependency grammars
  - Subject/agent relations
- Makes NLP tool evaluation difficult (the background of annotation may not fit the tool's one)

## Reusability

cf normalisation

# Obstacles and Difficulties (cont'd)

## The annotation process (cont'd)

- Annotation and linguistic theory:
  - Constituent vs dependency grammars
  - Subject/agent relations
- Makes NLP tool evaluation difficult (the background of annotation may not fit the tool's one)

## Reusability

cf normalisation

# EAGLES/ISLE

## EAGLES : Expert Advisory Group on Language Engineering Standards

- European Union initiative
- Produced recommendation

In particular:

- Recommendations for morphosyntactic annotation of corpora
- Specifications for english , german, italienish and french morphosyntax.

# EAGLES/ISLE

**EAGLES : Expert Advisory Group on Language Engineering Standards**

- European Union initiative
- Produced recommendation

In particular:

- Recommendations for morphosyntactic annotation of corpora
- Specifications for english , german, italienish and french morphosyntax.

# EAGLES/ISLE

## EAGLES : Expert Advisory Group on Language Engineering Standards

- European Union initiative
- Produced recommendation

In particular:

- Recommendations for morphosyntactic annotation of corpora
- Specifications for english , german, italienish and french morphosyntax.

# EAGLES/ISLE

## EAGLES : Expert Advisory Group on Language Engineering Standards

- European Union initiative
- Produced recommendation

In particular:

- Recommendations for morphosyntactic annotation of corpora
- Specifications for english , german, italienish and french morphosyntax.

# EAGLES/ISLE

## EAGLES : Expert Advisory Group on Language Engineering Standards

- European Union initiative
- Produced recommendation

In particular:

- Recommendations for morphosyntactic annotation of corpora
- Specifications for english , german, italienish and french  morphosyntax.

# EAGLES/ISLE

## ISLE : International Standards for Language Engineering

- EU and NSF
- Workshops, conferences and documents

In particular:

- Multilingual lexicons
- Evaluation (MT evaluation)

# EAGLES/ISLE

## ISLE : International Standards for Language Engineering

- EU and NSF
- Workshops, conferences and documents

In particular:

- Multilingual lexicons
- Evaluation (MT evaluation)

# EAGLES/ISLE

## ISLE : International Standards for Language Engineering

- EU and NSF
- Workshops, conferences and documents

In particular:

- Multilingual lexicons
- Evaluation (MT evaluation)

# EAGLES/ISLE

## ISLE : International Standards for Language Engineering

- EU and NSF
- Workshops, conferences and documents

In particular:

- Multilingual lexicons
- Evaluation (MT evaluation)

# EAGLES/ISLE

### ISLE : International Standards for Language Engineering

- EU and NSF
- Workshops, conferences and documents

In particular:

- Multilingual lexicons
- Evaluation (MT evaluation)

# Meta-data and Normalisation

**Objectives:**

- Sharing resources (transcriptions, annotated corpus, lexicons)
- Sharing tools (annotation, visualisation, access)
- Sharing practices (data collections, annotation guides, evaluation methods)

Require agreement on standards, definitions of norms for corpora and resources

# Meta-data and Normalisation

**Objectives:**

- Sharing resources (transcriptions, annotated corpus, lexicons)
- Sharing tools (annotation, visualisation, access)
- Sharing practices (data collections, annotation guides, evaluation methods)

Require agreement on standards, definitions of norms for corpora and resources

# Meta-data and Normalisation

**Objectives:**

- Sharing resources (transcriptions, annotated corpus, lexicons)
- Sharing tools (annotation, visualisation, access)
- Sharing practices (data collections, annotation guides, evaluation methods)

Require agreement on standards, definitions of norms for corpora and resources

# Meta-data and Normalisation

**Objectives:**

- Sharing resources (transcriptions, annotated corpus, lexicons)
- Sharing tools (annotation, visualisation, access)
- Sharing practices (data collections, annotation guides, evaluation methods)

Require agreement on standards, definitions of norms for corpora and resources

# To Deal with Corpora

## Linguistic aspects:

- Text selection
- Informations (genres, paragraphs, sentences, words. . . )

## Technical aspects:

- Close formats (encoding, annotation)
- Semantics annotation (generic concepts and specific realization of a morphosyntactic tagging)

# To Deal with Corpora

## Linguistic aspects:

- Text selection
- Informations (genres, paragraphs, sentences, words. . . )

## Technical aspects:

- Close formats (encoding, annotation)
- Semantics annotation (generic concepts and specific realization of a morphosyntactic tagging)

# To Deal with Corpora

**Linguistic aspects:**

- Text selection
- Informations (genres, paragraphs, sentences, words. . . )

**Technical aspects:**

- Close formats (encoding, annotation)
- Semantics annotation (generic concepts and specific realization of a morphosyntactic tagging)

# To Deal with Corpora

## Linguistic aspects:

- Text selection
- Informations (genres, paragraphs, sentences, words. . . )

## Technical aspects:

- Close formats (encoding, annotation)
- Semantics annotation (generic concepts and specific realization of a morphosyntactic tagging)

# To Deal with Corpora

**Linguistic aspects:**

- Text selection
- Informations (genres, paragraphs, sentences, words. . . )

**Technical aspects:**

- Close formats (encoding, annotation)
- Semantics annotation (generic concepts and specific realization of a morphosyntactic tagging)

# To Deal with Corpora

**Linguistic aspects:**

- Text selection
- Informations (genres, paragraphs, sentences, words. . . )

**Technical aspects:**

- Close formats (encoding, annotation)
- Semantics annotation (generic concepts and specific realization of a morphosyntactic tagging)

# To Deal with Resources

## Lexical databases

- Representation of the informations
  - Specifying these informations
  - Comparing, adding to these informations with other resources (interoperability)
- General organisation of the data (meta-model): onomasiological (from the concept to its realizations in the language) or semasiological (science of the meaning, from an expression—word—of the language to its meanings)
- Choice of the data depending on the model

# To Deal with Resources

## Lexical databases

- Representation of the informations
  - Specifying these informations
  - Comparing, adding to these informations with other resources (interoperability)
- General organisation of the data (meta-model): onomasiological (from the concept to its realizations in the language) or semasiological (science of the meaning, from an expression—word—of the language to its meanings)
- Choice of the data depending on the model

# To Deal with Resources

## Lexical databases

- Representation of the informations
  - Specifying these informations
  - Comparing, adding to these informations with other resources (interoperability)
- General organisation of the data (meta-model): onomasiological (from the concept to its realizations in the language) or semasiological (science of the meaning, from an expression—word—of the language to its meanings)
- Choice of the data depending on the model

# To Deal with Resources

## Lexical databases

- Representation of the informations
  - Specifying these informations
  - Comparing, adding to these informations with other resources (interoperability)
- General organisation of the data (meta-model): onomasiological (from the concept to its realizations in the language) or semasiological (science of the meaning, from an expression—word—of the language to its meanings)
- Choice of the data depending on the model

# To Deal with Resources

## Lexical databases

- Representation of the informations
  - Specifying these informations
  - Comparing, adding to these informations with other resources (interoperability)
- General organisation of the data (meta-model): onomasiological (from the concept to its realizations in the language) or semasiological (science of the meaning, from an expression—word—of the language to its meanings)
- Choice of the data depending on the model

# To Deal with Resources

## Lexical databases

- Representation of the informations
  - Specifying these informations
  - Comparing, adding to these informations with other resources (interoperability)
- General organisation of the data (meta-model): onomasiological (from the concept to its realizations in the language) or semasiological (science of the meaning, from an expression—word—of the language to its meanings)
- Choice of the data depending on the model

# Normalisation and Standards

- *De jure* norm (formal): set by an organisation which is mandated to design norms (e.g. ISO, International Standardisation Organisation, AFNOR, Association Française de Normalization, etc.)
- *De facto* norm: after the practice of a group of people, industries, users, etc.

# Normalisation and Standards

- *De jure* norm (formal): set by an organisation which is mandated to design norms (e.g. ISO, International Standardisation Organisation, AFNOR, Association Française de Normalization, etc.)
- *De facto* norm: after the practice of a group of people, industries, users, etc.

# The Normative Process

[Lupovici(1993)]:

1. Descending process by vertical division
2. Ascending process by vertical division
3. Transversal process

**Descending process:**

Organised with divisions by large fields, subdivided by topics, subdivided by working groups. . .
As for classification. Difficult for rapidly evolving domains.

# The Normative Process

[Lupovici(1993)]:

1. Descending process by vertical division
2. Ascending process by vertical division
3. Transversal process

**Descending process:**

Organised with divisions by large fields, subdivided by topics, subdivided by working groups. . .
As for classification. Difficult for rapidly evolving domains.

# The Normative Process

[Lupovici(1993)]:

1. Descending process by vertical division
2. Ascending process by vertical division
3. Transversal process

**Descending process:**

Organised with divisions by large fields, subdivided by topics, subdivided by working groups. . .
As for classification. Difficult for rapidly evolving domains.

# The Normative Process

[Lupovici(1993)]:

1. Descending process by vertical division
2. Ascending process by vertical division
3. Transversal process

**Descending process:**

Organised with divisions by large fields, subdivided by topics, subdivided by working groups. . .
As for classification. Difficult for rapidly evolving domains.

# Example: ISO TC37SC4

**ISO TC37** (Technical committee): Terminology and other language resources

- **SC3** (Sub-committee): Computer applications in terminology
  - ISO 12200: Martif
  - ISO 12620: Data categories
  - ISO 16642: TMF (Terminological Markup Framework)
- **SC4**: Language Resource Management. (Chairman: Laurent Romary)

**Ascending process**

Normalisation results from agreement (for inst. out of *de facto*). But difficulties in case of mismatch with a descending classification.)

**Transversal process**

Study of shared problems and definition of shared strategies. Towards a conceptual modelling of a field that can get realized into compatible and homogeneous standards (XML, OWL, Resource Description Framework (RDF), etc.)

# Example: ISO TC37SC4

**ISO TC37** (Technical committee): Terminology and other language resources

- **SC3** (Sub-committee): Computer applications in terminology
  - ISO 12200: Martif
  - ISO 12620: Data categories
  - ISO 16642: TMF (Terminological Markup Framework)
- **SC4**: Language Resource Management. (Chairman: Laurent Romary)

**Ascending process**

Normalisation results from agreement (for inst. out of *de facto*). But difficulties in case of mismatch with a descending classification.)

**Transversal process**

Study of shared problems and definition of shared strategies. Towards a conceptual modelling of a field that can get realized into compatible and homogeneous standards (XML, OWL, Resource Description Framework (RDF), etc.)

# Example: ISO TC37SC4

## **ISO TC37** (Technical committee): Terminology and other language resources

- **SC3** (Sub-committee): Computer applications in terminology
  - ISO 12200: Martif
  - ISO 12620: Data categories
  - ISO 16642: TMF (Terminological Markup Framework)
- **SC4**: Language Resource Management. (Chairman: Laurent Romary)

## Ascending process

Normalisation results from agreement (for inst. out of *de facto*). But difficulties in case of mismatch with a descending classification.)

## Transversal process

Study of shared problems and definition of shared strategies. Towards a conceptual modelling of a field that can get realized into compatible and homogeneous standards (XML, OWL, Resource Description Framework (RDF), etc.)

# Example: ISO TC37SC4

**ISO TC37** (Technical committee): Terminology and other language resources

- **SC3** (Sub-committee): Computer applications in terminology
  - ISO 12200: Martif
  - ISO 12620: Data categories
  - ISO 16642: TMF (Terminological Markup Framework)
- **SC4**: Language Resource Management. (Chairman: Laurent Romary)

**Ascending process**

Normalisation results from agreement (for inst. out of *de facto*). But difficulties in case of mismatch with a descending classification.)

**Transversal process**

Study of shared problems and definition of shared strategies. Towards a conceptual modelling of a field that can get realized into compatible and homogeneous standards (XML, OWL, Resource Description Framework (RDF), etc.)

# Example: ISO TC37SC4

## **ISO TC37** (Technical committee): Terminology and other language resources

- **SC3** (Sub-committee): Computer applications in terminology
  - ISO 12200: Martif
  - ISO 12620: Data categories
  - ISO 16642: TMF (Terminological Markup Framework)
- **SC4**: Language Resource Management. (Chairman: Laurent Romary)

## Ascending process

Normalisation results from agreement (for inst. out of *de facto*). But difficulties in case of mismatch with a descending classification.)

## Transversal process

Study of shared problems and definition of shared strategies. Towards a conceptual modelling of a field that can get realized into compatible and homogeneous standards (XML, OWL, Resource Description Framework (RDF), etc.)

# Example: ISO TC37SC4

**ISO TC37** (Technical committee): Terminology and other language resources

- **SC3** (Sub-committee): Computer applications in terminology
  - ISO 12200: Martif
  - ISO 12620: Data categories
  - ISO 16642: TMF (Terminological Markup Framework)
- **SC4**: Language Resource Management. (Chairman: Laurent Romary)

**Ascending process**

Normalisation results from agreement (for inst. out of *de facto*). But difficulties in case of mismatch with a descending classification.)

**Transversal process**

Study of shared problems and definition of shared strategies. Towards a conceptual modelling of a field that can get realized into compatible and homogeneous standards (XML, OWL, Resource Description Framework (RDF), etc.)

# Example: ISO TC37SC4

**ISO TC37** (Technical committee): Terminology and other language resources

- **SC3** (Sub-committee): Computer applications in terminology
  - ISO 12200: Martif
  - ISO 12620: Data categories
  - ISO 16642: TMF (Terminological Markup Framework)
- **SC4**: Language Resource Management. (Chairman: Laurent Romary)

**Ascending process**

Normalisation results from agreement (for inst. out of *de facto*). But difficulties in case of mismatch with a descending classification.)

**Transversal process**

Study of shared problems and definition of shared strategies. Towards a conceptual modelling of a field that can get realized into compatible and homogeneous standards (XML, OWL, Resource Description Framework (RDF), etc.)

# Example: ISO TC37SC4

## **ISO TC37** (Technical committee): Terminology and other language resources

- **SC3** (Sub-committee): Computer applications in terminology
  - ISO 12200: Martif
  - ISO 12620: Data categories
  - ISO 16642: TMF (Terminological Markup Framework)
- **SC4**: Language Resource Management. (Chairman: Laurent Romary)

## Ascending process

Normalisation results from agreement (for inst. out of *de facto*). But difficulties in case of mismatch with a descending classification.)

## Transversal process

Study of shared problems and definition of shared strategies. Towards a conceptual modelling of a field that can get realized into compatible and homogeneous standards (XML, OWL, Resource Description Framework (RDF), etc.)

# Work of TC37/SC4

- Linguistic Annotation Framework (general principles and conceptual representation for annotation schemas, pointer mechanisms, links for external annotation)
- Morphosyntactic tagging (reference model including tokenization issues, ambiguity issues, etc.)
- Lexical Markup Framework (platform for specifying lexical data)
- Other... (representing feature structures, register for data categories, etc.)

# Work of TC37/SC4

- Linguistic Annotation Framework (general principles and conceptual representation for annotation schemas, pointer mechanisms, links for external annotation)
- Morphosyntactic tagging (reference model including tokenization issues, ambiguity issues, etc.)
- Lexical Markup Framework (platform for specifying lexical data)
- Other... (representing feature structures, register for data categories, etc.)

# Work of TC37/SC4

- Linguistic Annotation Framework (general principles and conceptual representation for annotation schemas, pointer mechanisms, links for external annotation)
- Morphosyntactic tagging (reference model including tokenization issues, ambiguity issues, etc.)
- Lexical Markup Framework (platform for specifying lexical data)
- Other... (representing feature structures, register for data categories, etc.)

# Work of TC37/SC4

- Linguistic Annotation Framework (general principles and conceptual representation for annotation schemas, pointer mechanisms, links for external annotation)
- Morphosyntactic tagging (reference model including tokenization issues, ambiguity issues, etc.)
- Lexical Markup Framework (platform for specifying lexical data)
- Other... (representing feature structures, register for data categories, etc.)

# Specifications

- No unique format
- Means to specify models (to be instantiated for a given application), independent from any particular syntax, in order to preserve interoperability

Hence:

- Meta-models
- Data categories (elementary descriptors to be used in a linguistic description or an annotation schema. E.g. /Part of speech/, /Grammatical gender/, /Grammatical number/, /Feminine/, /Ablative/...)

# Specifications

- No unique format
- Means to specify models (to be instantiated for a given application), independent from any particular syntax, in order to preserve interoperability

Hence:

- Meta-models
- Data categories (elementary descriptors to be used in a linguistic description or an annotation schema. E.g. /Part of speech/, /Grammatical gender/, /Grammatical number/, /Feminine/, /Ablative/...)

# Specifications

- No unique format
- Means to specify models (to be instantiated for a given application), independent from any particular syntax, in order to preserve interoperability

Hence:

- Meta-models
- Data categories (elementary descriptors to be used in a linguistic description or an annotation schema. E.g. /Part of speech/, /Grammatical gender/, /Grammatical number/, /Feminine/, /Ablative/. . . )

# Specifications

- No unique format
- Means to specify models (to be instantiated for a given application), independent from any particular syntax, in order to preserve interoperability

Hence:

- Meta-models
- Data categories (elementary descriptors to be used in a linguistic description or an annotation schema. E.g. /Part of speech/, /Grammatical gender/, /Grammatical number/, /Feminine/, /Ablative/...)

# Example: morphosyntactic annotation

# Example: morphosyntactic annotation

## The MAF model (p. 14)

# Example: morphosyntactic annotation

## The MAF model (p. 14)

## Example: morphosyntactic annotation

### The MAF model (p. 14)

- Notion of *token* (p. 17)
- Notion of *word-form* (p. 19)
- Morphosyntactic information (p. 22)
- Ambiguity (p. 28)

# Example: morphosyntactic annotation

## The MAF model (p. 14)

# Example: lexical resources

## The MLF model

- Core model (p. 15)
- Extensions:
  - Extensional morphological lexicon (p. 21)
  - Intensional morphological lexicon (pp. 55–58)
  - Syntactic lexicon (p. 33)
  - Semantics lexicon (p. 39)

# Example: semantic annotation

## Time and events

### SemAF: Semantic annotation framework — Part 1: Time and events

- ISO-TimeML
- LINKS
- Semantics for TimeML
- Examples (p. 66 (74))

# Example: semantic annotation
## Time and events

### SemAF: Semantic annotation framework — Part 1: Time and events

- ISO-TimeML
- LINKS
- Semantics for TimeML
- Examples (p. 66 (74))

# Example: semantic annotation

## Time and events

### SemAF: Semantic annotation framework — Part 1: Time and events

- ISO-TimeML
- LINKS
- Semantics for TimeML
- Examples (p. 66 (74))

# Example: semantic annotation

## Time and events

### SemAF: Semantic annotation framework — Part 1: Time and events

- ISO-TimeML
- `LINKS`
- Semantics for TimeML
- Examples (p. 66 (74))

# Example: semantic annotation

## Time and events

### SemAF: Semantic annotation framework — Part 1: Time and events

- ISO-TimeML
- `LINKS`
- Semantics for TimeML
- Examples (p. 66 (74))

# Text Encoding Initiative (TEI)

## Chronological aspects

- At the beginning (1987):
  - Association for Computers and the Humanities
  - Association for Computational Linguistics
  - Association for Literary and Linguistic Computing
- Since 2000, consortium for maintaining and developing the TEI standard
- Academic consortium with a important human science part

## Activity

Standardisation activity: P3 (1992), P4 (XML, 2002), P5 (modular, 2004)

# Text Encoding Initiative (TEI)

## Chronological aspects

- At the beginning (1987):
  - Association for Computers and the Humanities
  - Association for Computational Linguistics
  - Association for Literary and Linguistic Computing
- Since 2000, consortium for maintaining and developing the TEI standard
- Academic consortium with a important human science part

## Activity

Standardisation activity: P3 (1992), P4 (XML, 2002), P5 (modular, 2004)

# Text Encoding Initiative (TEI)

## Chronological aspects

- At the beginning (1987):
  - Association for Computers and the Humanities
  - Association for Computational Linguistics
  - Association for Literary and Linguistic Computing
- Since 2000, consortium for maintaining and developing the TEI standard
- Academic consortium with a important human science part

## Activity

Standardisation activity: P3 (1992), P4 (XML, 2002), P5 (modular, 2004)

# Text Encoding Initiative (TEI)

## Chronological aspects

- At the beginning (1987):
  - Association for Computers and the Humanities
  - Association for Computational Linguistics
  - Association for Literary and Linguistic Computing
- Since 2000, consortium for maintaining and developing the TEI standard
- Academic consortium with a important human science part

## Activity

Standardisation activity: P3 (1992), P4 (XML, 2002), P5 (modular, 2004)

# Text Encoding Initiative (TEI)

## Chronological aspects

- At the beginning (1987):
  - Association for Computers and the Humanities
  - Association for Computational Linguistics
  - Association for Literary and Linguistic Computing
- Since 2000, consortium for maintaining and developing the TEI standard
- Academic consortium with a important human science part

## Activity

Standardisation activity: P3 (1992), P4 (XML, 2002), P5 (modular, 2004)

# Text Encoding Initiative (TEI)

## Chronological aspects

- At the beginning (1987):
  - Association for Computers and the Humanities
  - Association for Computational Linguistics
  - Association for Literary and Linguistic Computing
- Since 2000, consortium for maintaining and developing the TEI standard
- Academic consortium with a important human science part

## Activity

Standardisation activity: P3 (1992), P4 (XML, 2002), P5 (modular, 2004)

# Text Encoding Initiative (TEI)

## Chronological aspects

- At the beginning (1987):
  - Association for Computers and the Humanities
  - Association for Computational Linguistics
  - Association for Literary and Linguistic Computing
- Since 2000, consortium for maintaining and developing the TEI standard
- Academic consortium with a important human science part

## Activity

Standardisation activity: P3 (1992), P4 (XML, 2002), P5 (modular, 2004)

# Text Encoding Initiative (TEI)

## Chronological aspects

- At the beginning (1987):
  - Association for Computers and the Humanities
  - Association for Computational Linguistics
  - Association for Literary and Linguistic Computing
- Since 2000, consortium for maintaining and developing the TEI standard
- Academic consortium with a important human science part

## Activity

Standardisation activity: P3 (1992), P4 (XML, 2002), P5 (modular, 2004)

## Objectives

- To give a standardised format for data exchange
- To give guidelines for encoding (into these formats)
- To enable the encoding of any kind of information for any kind of text
- To be independent from applications

## Objectives

- To give a standardised format for data exchange
- To give guidelines for encoding (into these formats)
- To enable the encoding of any kind of information for any kind of text
- To be independent from applications

# Objectives

- To give a standardised format for data exchange
- To give guidelines for encoding (into these formats)
- To enable the encoding of any kind of information for any kind of text
- To be independent from applications

# Objectives

- To give a standardised format for data exchange
- To give guidelines for encoding (into these formats)
- To enable the encoding of any kind of information for any kind of text
- To be independent from applications

# Architectural choices

- Choice of SGML, XML, ISO 646 (ASCII), Unicode
- A large number of pre-defined annotations
- Distinction between required practices, recommended practices and optional practices
- Encoding for various points of view on the text
- Provide ways for users to extend basic schemas

*This mandate is fulfilled by the explicit specification, in the reference section for each tag, that the tag is required, mandatory when applicable but otherwise omissible, recommended generally, recommended when applicable but not always applicable, or optional.*
*However, the TEI Guidelines make (with relatively rare exceptions) no suggestions or restrictions as to the relative importance of textual features. The philosophy of the Guidelines is 'if you want to encode this feature, do it this way'— but very few features are mandatory.*

# Architectural choices

- Choice of SGML, XML, ISO 646 (ASCII), Unicode
- A large number of pre-defined annotations
- Distinction between required practices, recommended practices and optional practices
- Encoding for various points of view on the text
- Provide ways for users to extend basic schemas

*This mandate is fulfilled by the explicit specification, in the reference section for each tag, that the tag is required, mandatory when applicable but otherwise omissible, recommended generally, recommended when applicable but not always applicable, or optional.*
*However, the TEI Guidelines make (with relatively rare exceptions) no suggestions or restrictions as to the relative importance of textual features. The philosophy of the Guidelines is 'if you want to encode this feature, do it this way'— but very few features are mandatory.*

# Architectural choices

- Choice of SGML, XML, ISO 646 (ASCII), Unicode
- A large number of pre-defined annotations
- Distinction between required practices, recommended practices and optional practices
- Encoding for various points of view on the text
- Provide ways for users to extend basic schemas

*This mandate is fulfilled by the explicit specification, in the reference section for each tag, that the tag is required, mandatory when applicable but otherwise omissible, recommended generally, recommended when applicable but not always applicable, or optional.*
*However, the TEI Guidelines make (with relatively rare exceptions) no suggestions or restrictions as to the relative importance of textual features. The philosophy of the Guidelines is 'if you want to encode this feature, do it this way'— but very few features are mandatory.*

# Architectural choices

- Choice of SGML, XML, ISO 646 (ASCII), Unicode
- A large number of pre-defined annotations
- Distinction between required practices, recommended practices and optional practices
- Encoding for various points of view on the text
- Provide ways for users to extend basic schemas

*This mandate is fulfilled by the explicit specification, in the reference section for each tag, that the tag is required, mandatory when applicable but otherwise omissible, recommended generally, recommended when applicable but not always applicable, or optional.*
*However, the TEI Guidelines make (with relatively rare exceptions) no suggestions or restrictions as to the relative importance of textual features. The philosophy of the Guidelines is 'if you want to encode this feature, do it this way'— but very few features are mandatory.*

# Architectural choices

- Choice of SGML, XML, ISO 646 (ASCII), Unicode
- A large number of pre-defined annotations
- Distinction between required practices, recommended practices and optional practices
- Encoding for various points of view on the text
- Provide ways for users to extend basic schemas

*This mandate is fulfilled by the explicit specification, in the reference section for each tag, that the tag is required, mandatory when applicable but otherwise omissible, recommended generally, recommended when applicable but not always applicable, or optional.*
*However, the TEI Guidelines make (with relatively rare exceptions) no suggestions or restrictions as to the relative importance of textual features. The philosophy of the Guidelines is 'if you want to encode this feature, do it this way'— but very few features are mandatory.*

# Architectural choices

- Choice of SGML, XML, ISO 646 (ASCII), Unicode
- A large number of pre-defined annotations
- Distinction between required practices, recommended practices and optional practices
- Encoding for various points of view on the text
- Provide ways for users to extend basic schemas

*This mandate is fulfilled by the explicit specification, in the reference section for each tag, that the tag is required, mandatory when applicable but otherwise omissible, recommended generally, recommended when applicable but not always applicable, or optional.*
*However, the TEI Guidelines make (with relatively rare exceptions) no suggestions or restrictions as to the relative importance of textual features. The philosophy of the Guidelines is 'if you want to encode this feature, do it this way'— but very few features are mandatory.*

# Example of recommendations

From TEI P5. Guidelines for Electronic Text Encoding and Interchange

- Almost always present elements (paragraphs, ponctuation, citations, dates, lists, etc.)
- Description of the encoded document (TEI Header) (description of the file—title, author. . . —, of the encoding, if the profile—register. . . )
- Structure of the text. . . (TEI + TEIheader + text, example monde98), divisions (chapter, cf Alice)

# Syntactic function and semantic roles

## The Proposition Treebank [Palmer et al.(2005)Palmer, Gildea, and Kingsbury]

### Predicate-argument information/semantic roles

The same semantic role can be taken by different syntactic arguments

#### Example (Syntactic argument vs semantic argument)

- John broke the window
- The window broke

Note:

- Does not rely on an active/passive voice distinction
- Does not rely on transitive/intransitive distinction

# Syntactic function and semantic roles

## The Proposition Treebank [Palmer et al.(2005)Palmer, Gildea, and Kingsbury]

### Predicate-argument information/semantic roles

The same semantic role can be taken by different syntactic arguments

### Example (Syntactic argument vs semantic argument)

- John broke the window
- The window broke

Note:

- Does not rely on an active/passive voice distinction
- Does not rely on transitive/intransitive distinction

# Syntactic function and semantic roles

## The Proposition Treebank [Palmer et al.(2005)Palmer, Gildea, and Kingsbury]

### Predicate-argument information/semantic roles

The same semantic role can be taken by different syntactic arguments

### Example (Syntactic argument vs semantic argument)

- John broke the window
- The window broke

Note:

- Does not rely on an active/passive voice distinction
- Does not rely on transitive/intransitive distinction

# Syntactic function and semantic roles

## The Proposition Treebank [Palmer et al.(2005)Palmer, Gildea, and Kingsbury]

### Predicate-argument information/semantic roles

The same semantic role can be taken by different syntactic arguments

### Example (Syntactic argument vs semantic argument)

- John broke the window
- The window broke

Note:

- Does not rely on an active/passive voice distinction
- Does not rely on transitive/intransitive distinction

# Syntactic function and semantic roles

## The Proposition Treebank [Palmer et al.(2005)Palmer, Gildea, and Kingsbury]

### Predicate-argument information/semantic roles

The same semantic role can be taken by different syntactic arguments

### Example (Syntactic argument vs semantic argument)

- John broke the window
- The window broke

Note:

- Does not rely on an active/passive voice distinction
- Does not rely on transitive/intransitive distinction

# Syntactic function and semantic roles

## The Proposition Treebank [Palmer et al.(2005)Palmer, Gildea, and Kingsbury]

### Predicate-argument information/semantic roles

The same semantic role can be taken by different syntactic arguments

### Example (Syntactic argument vs semantic argument)

- John broke the window
- The window broke

Note:

- Does not rely on an active/passive voice distinction
- Does not rely on transitive/intransitive distinction

# Syntactic function and semantic roles

**Example (Roles of the subject)**

- The sergeant played taps
- The sergeant played
- Taps played quietly in the background

**Example (Roles of the object)**

- The sergeant played taps
- The sergeant played a beat-up old bugle

# Syntactic function and semantic roles

**Example (Roles of the subject)**

- The sergeant played taps
- The sergeant played
- Taps played quietly in the background

**Example (Roles of the object)**

- The sergeant played taps
- The sergeant played a beat-up old bugle

# Syntactic function and semantic roles

**Example (Roles of the subject)**

- The sergeant played taps
- The sergeant played
- Taps played quietly in the background

**Example (Roles of the object)**

- The sergeant played taps
- The sergeant played a beat-up old bugle

# Syntactic function and semantic roles

**Example (Roles of the subject)**

- The sergeant played taps
- The sergeant played
- Taps played quietly in the background

**Example (Roles of the object)**

- The sergeant played taps
- The sergeant played a beat-up old bugle

# Syntactic function and semantic roles

## Example (Roles of the subject)

- The sergeant played taps
- The sergeant played
- Taps played quietly in the background

## Example (Roles of the object)

- The sergeant played taps
- The sergeant played a beat-up old bugle

# Syntactic function and semantic roles

**Example (Roles of the subject)**

- The sergeant played taps
- The sergeant played
- Taps played quietly in the background

**Example (Roles of the object)**

- The sergeant played taps
- The sergeant played a beat-up old bugle

# Syntactic function and semantic roles

**Example (Roles of the subject)**

- The sergeant played taps
- The sergeant played
- Taps played quietly in the background

**Example (Roles of the object)**

- The sergeant played taps
- The sergeant played a beat-up old bugle

# Semantic Annotation: objective

## Getting roles

- Automatically: not a simple translation from syntax...
- Hence, the need to build a corpus

## General objectives

- Get data on the actual frequency of syntactic variations
- Help statistical systems based on shallow levels of semantic representation

# Semantic Annotation: objective

## Getting roles

- Automatically: not a simple translation from syntax. . .
- Hence, the need to build a corpus

## General objectives

- Get data on the actual frequency of syntactic variations
- Help statistical systems based on shallow levels of semantic representation

# Semantic Annotation: objective

## Getting roles

- Automatically: not a simple translation from syntax...
- Hence, the need to build a corpus

## General objectives

- Get data on the actual frequency of syntactic variations
- Help statistical systems based on shallow levels of semantic representation

# Semantic Annotation: objective

## Getting roles

- Automatically: not a simple translation from syntax...
- Hence, the need to build a corpus

## General objectives

- Get data on the actual frequency of syntactic variations
- Help statistical systems based on shallow levels of semantic representation

# Semantic Annotation: objective

## Getting roles

- Automatically: not a simple translation from syntax...
- Hence, the need to build a corpus

## General objectives

- Get data on the actual frequency of syntactic variations
- Help statistical systems based on shallow levels of semantic representation

# Semantic Annotation: objective

## Getting roles

- Automatically: not a simple translation from syntax. . .
- Hence, the need to build a corpus

## General objectives

- Get data on the actual frequency of syntactic variations
- Help statistical systems based on shallow levels of semantic representation

# What kind of annotations

## Example (What is expected?)

- ... [$_{\text{Arg0}}$the company] to ... *offer* [$_{\text{Arg1}}$a 15% to a 20% stake] [$_{\text{Arg2}}$to the public]
- ... [$_{\text{Arg0}}$Sotheby's] ... *offered* [$_{\text{Arg2}}$the Dorrance heirs][$_{\text{Arg1}}$a money-back guarantee]
- ... [$_{\text{Arg1}}$an amendment] *offered* [$_{\text{Arg0}}$by Rep. Peter DeFazio] ...
- ... [$_{\text{Arg2}}$Subcontractors] will be *offered* [$_{\text{Arg1}}$a settlement]...

## What is needed: the possible roles for a verb

## Example (Frameset kick.01 "drive or impel with the foot")

| | |
|---|---|
| Arg0: | Kicker |
| Arg1: | Thing kicked |
| Arg2: | Instrument (default to foot) |
| Ex1: | But [$_{\text{Arg0}}$two big New York banks$_i$] seem [$_{\text{Arg0}}$*trace*$_i$] to have *kicked* [$_{\text{Arg1}}$those chances], for the moment, [$_{\text{Arg2}}$with the embarrassing failure of Citicorp and Chase Manhattan Corp. ...] |
| Ex2: | [$_{\text{Arg0}}$John$_i$] tried [$_{\text{Arg0}}$*trace*$_i$] to *kick* [$_{\text{Arg1}}$the football], but Mary pulled it away at the last moment. |

# What kind of annotations

## Example (What is expected?)

- . . . [$_{\text{Arg0}}$the company] to . . . *offer* [$_{\text{Arg1}}$a 15% to a 20% stake] [$_{\text{Arg2}}$to the public]
- . . . [$_{\text{Arg0}}$Sotheby's] . . . *offered* [$_{\text{Arg2}}$the Dorrance heirs][$_{\text{Arg1}}$a money-back guarantee]
- . . . [$_{\text{Arg1}}$an amendment] *offered* [$_{\text{Arg0}}$by Rep. Peter DeFazio] . . .
- . . . [$_{\text{Arg2}}$Subcontractors] will be *offered* [$_{\text{Arg1}}$a settlement]. . .

## What is needed: the possible roles for a verb

## Example (Frameset kick.01 "drive or impel with the foot")

| | |
|---|---|
| Arg0: | Kicker |
| Arg1: | Thing kicked |
| Arg2: | Instrument (default to foot) |
| Ex1: | But [$_{\text{Arg0}}$two big New York banks$_i$] seem [$_{\text{Arg0}}$*trace*$_i$] to have *kicked* [$_{\text{Arg1}}$those chances], for the moment, [$_{\text{Arg2}}$with the embarrassing failure of Citicorp and Chase Manhattan Corp. . . . ] |
| Ex2: | [$_{\text{Arg0}}$John$_i$] tried [$_{\text{Arg0}}$*trace*$_i$] to *kick* [$_{\text{Arg1}}$the football], but Mary pulled it away at the last moment. |

# What kind of annotations

## Example (What is expected?)

- . . . [$_{\text{Arg0}}$the company] to . . . *offer* [$_{\text{Arg1}}$a 15% to a 20% stake] [$_{\text{Arg2}}$to the public]
- . . . [$_{\text{Arg0}}$Sotheby's] . . . *offered* [$_{\text{Arg2}}$the Dorrance heirs][$_{\text{Arg1}}$a money-back guarantee]
- . . . [$_{\text{Arg1}}$an amendment] *offered* [$_{\text{Arg0}}$by Rep. Peter DeFazio] . . .
- . . . [$_{\text{Arg2}}$Subcontractors] will be *offered* [$_{\text{Arg1}}$a settlement]. . .

## What is needed: the possible roles for a verb

## Example (Frameset kick.01 "drive or impel with the foot")

| | |
|---|---|
| Arg0: | Kicker |
| Arg1: | Thing kicked |
| Arg2: | Instrument (default to foot) |
| Ex1: | But [$_{\text{Arg0}}$two big New York banks$_i$] seem [$_{\text{Arg0}}$*trace*$_i$] to have *kicked* [$_{\text{Arg1}}$those chances], for the moment, [$_{\text{Arg2}}$with the embarrassing failure of Citicorp and Chase Manhattan Corp. . . . ] |
| Ex2: | [$_{\text{Arg0}}$John$_i$] tried [$_{\text{Arg0}}$*trace*$_i$] to *kick* [$_{\text{Arg1}}$the football], but Mary pulled it away at the last moment. |

# What kind of annotations

## Example (What is expected?)

- … [$_{\text{Arg0}}$the company] to … *offer* [$_{\text{Arg1}}$a 15% to a 20% stake] [$_{\text{Arg2}}$to the public]
- … [$_{\text{Arg0}}$Sotheby's] … *offered* [$_{\text{Arg2}}$the Dorrance heirs][$_{\text{Arg1}}$a money-back guarantee]
- … [$_{\text{Arg1}}$an amendment] *offered* [$_{\text{Arg0}}$by Rep. Peter DeFazio] …
- … [$_{\text{Arg2}}$Subcontractors] will be *offered* [$_{\text{Arg1}}$a settlement]…

## What is needed: the possible roles for a verb

### Example (Frameset kick.01 "drive or impel with the foot")

| | |
|---|---|
| Arg0: | Kicker |
| Arg1: | Thing kicked |
| Arg2: | Instrument (default to foot) |
| Ex1: | But [$_{\text{Arg0}}$two big New York banks$_i$] seem [$_{\text{Arg0}}$*trace*$_i$] to have *kicked* [$_{\text{Arg1}}$those chances], for the moment, [$_{\text{Arg2}}$with the embarrassing failure of Citicorp and Chase Manhattan Corp. …] |
| Ex2: | [$_{\text{Arg0}}$John$_i$] tried [$_{\text{Arg0}}$*trace*$_i$] to *kick* [$_{\text{Arg1}}$the football], but Mary pulled it away at the last moment. |

# What kind of annotations

## Example (What is expected?)

- . . . [$_{\text{Arg0}}$the company] to . . . *offer* [$_{\text{Arg1}}$a 15% to a 20% stake] [$_{\text{Arg2}}$to the public]
- . . . [$_{\text{Arg0}}$Sotheby's] . . . *offered* [$_{\text{Arg2}}$the Dorrance heirs][$_{\text{Arg1}}$a money-back guarantee]
- . . . [$_{\text{Arg1}}$an amendment] *offered* [$_{\text{Arg0}}$by Rep. Peter DeFazio] . . .
- . . . [$_{\text{Arg2}}$Subcontractors] will be *offered* [$_{\text{Arg1}}$a settlement]. . .

## What is needed: the possible roles for a verb

### Example (Frameset kick.01 "drive or impel with the foot")

| | |
|---|---|
| Arg0: | Kicker |
| Arg1: | Thing kicked |
| Arg2: | Instrument (default to foot) |
| Ex1: | But [$_{\text{Arg0}}$two big New York banks$_i$] seem [$_{\text{Arg0}}$*trace*$_i$] to have *kicked* [$_{\text{Arg1}}$those chances], for the moment, [$_{\text{Arg2}}$with the embarrassing failure of Citicorp and Chase Manhattan Corp. . . . ] |
| Ex2: | [$_{\text{Arg0}}$John$_i$] tried [$_{\text{Arg0}}$*trace*$_i$] to *kick* [$_{\text{Arg1}}$the football], but Mary pulled it away at the last moment. |

# What kind of annotations

## Example (What is expected?)

- . . . [$_{\text{Arg0}}$the company] to . . . *offer* [$_{\text{Arg1}}$a 15% to a 20% stake] [$_{\text{Arg2}}$to the public]
- . . . [$_{\text{Arg0}}$Sotheby's] . . . *offered* [$_{\text{Arg2}}$the Dorrance heirs][$_{\text{Arg1}}$a money-back guarantee]
- . . . [$_{\text{Arg1}}$an amendment] *offered* [$_{\text{Arg0}}$by Rep. Peter DeFazio] . . .
- . . . [$_{\text{Arg2}}$Subcontractors] will be *offered* [$_{\text{Arg1}}$a settlement]. . .

## What is needed: the possible roles for a verb

### Example (Frameset **kick.01** "drive or impel with the foot")

Arg0: Kicker
Arg1: Thing kicked
Arg2: Instrument (default to foot)
Ex1: But [$_{\text{Arg0}}$two big New York banks$_i$] seem [$_{\text{Arg0}}$*trace*$_i$] to have *kicked* [$_{\text{Arg1}}$those chances], for the moment, [$_{\text{Arg2}}$with the embarrassing failure of Citicorp and Chase Manhattan Corp. . . ]
Ex2: [$_{\text{Arg0}}$John$_i$] tried [$_{\text{Arg0}}$*trace*$_i$] to *kick* [$_{\text{Arg1}}$the football], but Mary pulled it away at the last moment.

# What kind of annotations

## Example (What is expected?)

- . . . [$_{\text{Arg0}}$the company] to . . . *offer* [$_{\text{Arg1}}$a 15% to a 20% stake] [$_{\text{Arg2}}$to the public]
- . . . [$_{\text{Arg0}}$Sotheby's] . . . *offered* [$_{\text{Arg2}}$the Dorrance heirs][$_{\text{Arg1}}$a money-back guarantee]
- . . . [$_{\text{Arg1}}$an amendment] *offered* [$_{\text{Arg0}}$by Rep. Peter DeFazio] . . .
- . . . [$_{\text{Arg2}}$Subcontractors] will be *offered* [$_{\text{Arg1}}$a settlement]. . .

## What is needed: the possible roles for a verb

## Example (Frameset **kick.01** "drive or impel with the foot")

| | |
|---|---|
| Arg0: | Kicker |
| Arg1: | Thing kicked |
| Arg2: | Instrument (default to foot) |
| Ex1: | But [$_{\text{Arg0}}$two big New York banks$_i$] seem [$_{\text{Arg0}}$*trace*$_i$] to have *kicked* [$_{\text{Arg1}}$those chances], for the moment, [$_{\text{Arg2}}$with the embarrassing failure of Citicorp and Chase Manhattan Corp. . . ] |
| Ex2: | [$_{\text{Arg0}}$John$_i$] tried [$_{\text{Arg0}}$*trace*$_i$] to *kick* [$_{\text{Arg1}}$the football], but Mary pulled it away at the last moment. |

# Framesets

- Frameset of a verb: a set of syntactic variations associated with a set of roles for this verb (see the Framing guidelines )
- Note:
  - This lexicon need to be build before the annotation (at least conceptually). Some of the realizations are not attested in the corpus.
  - The semantic of Arg0, Arg1, etc. depends on the verb
  - When to have multiple framesets, single framesets?
  - Both on a semantic and syntactical ground
  - Examples (decline , break , call )
- For PropBank:
  - 3,300 verbs
  - 4,500 framesets (average polysemy: 1.36)
  - only 21.5% of the verbs have more than 1 framesets
  - less than 100 verbs have 4 framesets or more

# Framesets

- Frameset of a verb: a set of syntactic variations associated with a set of roles for this verb (see the Framing guidelines )
- Note:
  - This lexicon need to be build before the annotation (at least conceptually). Some of the realizations are not attested in the corpus.
  - The semantic of Arg0, Arg1, etc. depends on the verb
  - When to have multiple framesets, single framesets?
  - Both on a semantic and syntactical ground
  - Examples (decline , break , call )
- For PropBank:
  - 3,300 verbs
  - 4,500 framesets (average polysemy: 1.36)
  - only 21.5% of the verbs have more than 1 framesets
  - less than 100 verbs have 4 framesets or more

# Framesets

- Frameset of a verb: a set of syntactic variations associated with a set of roles for this verb (see the Framing guidelines )
- Note:
  - This lexicon need to be build before the annotation (at least conceptually). Some of the realizations are not attested in the corpus.
  - The semantic of Arg0, Arg1, etc. depends on the verb
  - When to have multiple framesets, single framesets?
  - Both on a semantic and syntactical ground
  - Examples (decline , break , call )
- For PropBank:
  - 3,300 verbs
  - 4,500 framesets (average polysemy: 1.36)
  - only 21.5% of the verbs have more than 1 framesets
  - less than 100 verbs have 4 framesets or more

# Framesets

- Frameset of a verb: a set of syntactic variations associated with a set of roles for this verb (see the Framing guidelines )
- Note:
  - This lexicon need to be build before the annotation (at least conceptually). Some of the realizations are not attested in the corpus.
  - The semantic of Arg0, Arg1, etc. depends on the verb
  - When to have multiple framesets, single framesets?
  - Both on a semantic and syntactical ground
  - Examples (decline , break , call )
- For PropBank:
  - 3,300 verbs
  - 4,500 framesets (average polysemy: 1.36)
  - only 21.5% of the verbs have more than 1 framesets
  - less than 100 verbs have 4 framesets or more

# Framesets

- Frameset of a verb: a set of syntactic variations associated with a set of roles for this verb (see the Framing guidelines )
- Note:
  - This lexicon need to be build before the annotation (at least conceptually). Some of the realizations are not attested in the corpus.
  - The semantic of Arg0, Arg1, etc. depends on the verb
  - When to have multiple framesets, single framesets?
  - Both on a semantic and syntactical ground
  - Examples (decline , break , call )
- For PropBank:
  - 3,300 verbs
  - 4,500 framesets (average polysemy: 1.36)
  - only 21.5% of the verbs have more than 1 framesets
  - less than 100 verbs have 4 framesets or more

# Framesets

- Frameset of a verb: a set of syntactic variations associated with a set of roles for this verb (see the Framing guidelines )
- Note:
  - This lexicon need to be build before the annotation (at least conceptually). Some of the realizations are not attested in the corpus.
  - The semantic of Arg0, Arg1, etc. depends on the verb
  - When to have multiple framesets, single framesets?
  - Both on a semantic and syntactical ground
  - Examples (decline , break , call )
- For PropBank:
  - 3,300 verbs
  - 4,500 framesets (average polysemy: 1.36)
  - only 21.5% of the verbs have more than 1 framesets
  - less than 100 verbs have 4 framesets or more

# Framesets

- Frameset of a verb: a set of syntactic variations associated with a set of roles for this verb (see the Framing guidelines )
- Note:
  - This lexicon need to be build before the annotation (at least conceptually). Some of the realizations are not attested in the corpus.
  - The semantic of Arg0, Arg1, etc. depends on the verb
  - When to have multiple framesets, single framesets?
  - Both on a semantic and syntactical ground
  - Examples (decline , break , call )
- For PropBank:
  - 3,300 verbs
  - 4,500 framesets (average polysemy: 1.36)
  - only 21.5% of the verbs have more than 1 framesets
  - less than 100 verbs have 4 framesets or more

# Framesets

- Frameset of a verb: a set of syntactic variations associated with a set of roles for this verb (see the Framing guidelines )
- Note:
  - This lexicon need to be build before the annotation (at least conceptually). Some of the realizations are not attested in the corpus.
  - The semantic of Arg0, Arg1, etc. depends on the verb
  - When to have multiple framesets, single framesets?
  - Both on a semantic and syntactical ground
  - Examples (decline , break , call )
- For PropBank:
  - 3,300 verbs
  - 4,500 framesets (average polysemy: 1.36)
  - only 21.5% of the verbs have more than 1 framesets
  - less than 100 verbs have 4 framesets or more

# Framesets

- Frameset of a verb: a set of syntactic variations associated with a set of roles for this verb (see the Framing guidelines )
- Note:
  - This lexicon need to be build before the annotation (at least conceptually). Some of the realizations are not attested in the corpus.
  - The semantic of Arg0, Arg1, etc. depends on the verb
  - When to have multiple framesets, single framesets?
  - Both on a semantic and syntactical ground
  - Examples (decline , break , call )
- For PropBank:
  - 3,300 verbs
  - 4,500 framesets (average polysemy: 1.36)
  - only 21.5% of the verbs have more than 1 framesets
  - less than 100 verbs have 4 framesets or more

# Framesets

- Frameset of a verb: a set of syntactic variations associated with a set of roles for this verb (see the Framing guidelines )
- Note:
  - This lexicon need to be build before the annotation (at least conceptually). Some of the realizations are not attested in the corpus.
  - The semantic of Arg0, Arg1, etc. depends on the verb
  - When to have multiple framesets, single framesets?
  - Both on a semantic and syntactical ground
  - Examples (decline , break , call )
- For PropBank:
  - 3,300 verbs
  - 4,500 framesets (average polysemy: 1.36)
  - only 21.5% of the verbs have more than 1 framesets
  - less than 100 verbs have 4 framesets or more

# Framesets

- Frameset of a verb: a set of syntactic variations associated with a set of roles for this verb (see the Framing guidelines )
- Note:
  - This lexicon need to be build before the annotation (at least conceptually). Some of the realizations are not attested in the corpus.
  - The semantic of Arg0, Arg1, etc. depends on the verb
  - When to have multiple framesets, single framesets?
  - Both on a semantic and syntactical ground
  - Examples (decline , break , call )
- For PropBank:
  - 3,300 verbs
  - 4,500 framesets (average polysemy: 1.36)
  - only 21.5% of the verbs have more than 1 framesets
  - less than 100 verbs have 4 framesets or more

# Framesets

- Frameset of a verb: a set of syntactic variations associated with a set of roles for this verb (see the Framing guidelines )
- Note:
  - This lexicon need to be build before the annotation (at least conceptually). Some of the realizations are not attested in the corpus.
  - The semantic of Arg0, Arg1, etc. depends on the verb
  - When to have multiple framesets, single framesets?
  - Both on a semantic and syntactical ground
  - Examples (decline , break , call )
- For PropBank:
  - 3,300 verbs
  - 4,500 framesets (average polysemy: 1.36)
  - only 21.5% of the verbs have more than 1 framesets
  - less than 100 verbs have 4 framesets or more

# The Annotation Process

## The semantic annotation is linked to the syntactic one

- Address of the syntactic node
- Help of a rule-based argument tagger (on pilot data, then error recovering by hand)
- A two-pass blind procedure, followed by an adjudication phase
- the annotators:
  - Different backgrounds (linguists, computer scientists, others)
  - Undergraduate students (inexpensive, but only few months so require frequent training)
  - Linguists have the best overall judgements
  - Very steep learning curves (comfortable after 3 days). Contrast with syntactic annotations
  - Over 30 annotators (from few weeks to 3 years)

# The Annotation Process

## The semantic annotation is linked to the syntactic one

- Address of the syntactic node
- Help of a rule-based argument tagger (on pilot data, then error recovering by hand)
- A two-pass blind procedure, followed by an adjudication phase
- the annotators:
  - Different backgrounds (linguists, computer scientists, others)
  - Undergraduate students (inexpensive, but only few months so require frequent training)
  - Linguists have the best overall judgements
  - Very steep learning curves (comfortable after 3 days). Contrast with syntactic annotations
  - Over 30 annotators (from few weeks to 3 years)

# The Annotation Process

## The semantic annotation is linked to the syntactic one

- Address of the syntactic node
- Help of a rule-based argument tagger (on pilot data, then error recovering by hand)
- A two-pass blind procedure, followed by an adjudication phase
- the annotators:
  - Different backgrounds (linguists, computer scientists, others)
  - Undergraduate students (inexpensive, but only few months so require frequent training)
  - Linguists have the best overall judgements
  - Very steep learning curves (comfortable after 3 days). Contrast with syntactic annotations
  - Over 30 annotators (from few weeks to 3 years)

# The Annotation Process

## The semantic annotation is linked to the syntactic one

- Address of the syntactic node
- Help of a rule-based argument tagger (on pilot data, then error recovering by hand)
- A two-pass blind procedure, followed by an adjudication phase
- the annotators:
  - Different backgrounds (linguists, computer scientists, others)
  - Undergraduate students (inexpensive, but only few months so require frequent training)
  - Linguists have the best overall judgements
  - Very steep learning curves (comfortable after 3 days). Contrast with syntactic annotations
  - Over 30 annotators (from few weeks to 3 years)

# The Annotation Process

## The semantic annotation is linked to the syntactic one

- Address of the syntactic node
- Help of a rule-based argument tagger (on pilot data, then error recovering by hand)
- A two-pass blind procedure, followed by an adjudication phase
- the annotators:
  - Different backgrounds (linguists, computer scientists, others)
  - Undergraduate students (inexpensive, but only few months so require frequent training)
  - Linguists have the best overall judgements
  - Very steep learning curves (comfortable after 3 days). Contrast with syntactic annotations
  - Over 30 annotators (from few weeks to 3 years)

# The Annotation Process

## The semantic annotation is linked to the syntactic one

- Address of the syntactic node
- Help of a rule-based argument tagger (on pilot data, then error recovering by hand)
- A two-pass blind procedure, followed by an adjudication phase
- the annotators:
  - Different backgrounds (linguists, computer scientists, others)
  - Undergraduate students (inexpensive, but only few months so require frequent training)
  - Linguists have the best overall judgements
  - Very steep learning curves (comfortable after 3 days). Contrast with syntactic annotations
  - Over 30 annotators (from few weeks to 3 years)

# The Annotation Process

## The semantic annotation is linked to the syntactic one

- Address of the syntactic node
- Help of a rule-based argument tagger (on pilot data, then error recovering by hand)
- A two-pass blind procedure, followed by an adjudication phase
- the annotators:
  - Different backgrounds (linguists, computer scientists, others)
  - Undergraduate students (inexpensive, but only few months so require frequent training)
  - Linguists have the best overall judgements
  - Very steep learning curves (comfortable after 3 days). Contrast with syntactic annotations
  - Over 30 annotators (from few weeks to 3 years)

# The Annotation Process

## The semantic annotation is linked to the syntactic one

- Address of the syntactic node
- Help of a rule-based argument tagger (on pilot data, then error recovering by hand)
- A two-pass blind procedure, followed by an adjudication phase
- the annotators:
  - Different backgrounds (linguists, computer scientists, others)
  - Undergraduate students (inexpensive, but only few months so require frequent training)
  - Linguists have the best overall judgements
  - Very steep learning curves (comfortable after 3 days). Contrast with syntactic annotations
  - Over 30 annotators (from few weeks to 3 years)

# The Annotation Process

## The semantic annotation is linked to the syntactic one

- Address of the syntactic node
- Help of a rule-based argument tagger (on pilot data, then error recovering by hand)
- A two-pass blind procedure, followed by an adjudication phase
- the annotators:
  - Different backgrounds (linguists, computer scientists, others)
  - Undergraduate students (inexpensive, but only few months so require frequent training)
  - Linguists have the best overall judgements
  - Very steep learning curves (comfortable after 3 days). Contrast with syntactic annotations
  - Over 30 annotators (from few weeks to 3 years)

# The Annotation Process

## The semantic annotation is linked to the syntactic one

- Address of the syntactic node
- Help of a rule-based argument tagger (on pilot data, then error recovering by hand)
- A two-pass blind procedure, followed by an adjudication phase
- the annotators:
  - Different backgrounds (linguists, computer scientists, others)
  - Undergraduate students (inexpensive, but only few months so require frequent training)
  - Linguists have the best overall judgements
  - Very steep learning curves (comfortable after 3 days). Contrast with syntactic annotations
  - Over 30 annotators (from few weeks to 3 years)

# Some results

## Syntactic position/role distribution

- When there is an Arg0, then it is a subject (96.9%), another verbal argument (2.4%) or an object (0.2%)
- The subject of a verb has Arg0 role (79.0%), Arg1 role (16.8%), Arg2 role (2.4%)...

## Automatic semantic labelling:Comparison with FrameNet

| | FrameNet | PropBank | PropBank (>10 ex) |
|---|---|---|---|
| Automatic parses | 82.0 | 79.9 | 80.9 |
| Gold-standard parses | | 82.0 | 82.8 |

Accuracy of semantic role prediction for known boundaries

# Some results

## Syntactic position/role distribution

- When there is an Arg0, then it is a subject (96.9%), another verbal argument (2.4%) or an object (0.2%)
- The subject of a verb has Arg0 role (79.0%), Arg1 role (16.8%), Arg2 role (2.4%). . .

## Automatic semantic labelling:Comparison with FrameNet

| | FrameNet | PropBank | PropBank (>10 ex) |
|---|---|---|---|
| Automatic parses | 82.0 | 79.9 | 80.9 |
| Gold-standard parses | | 82.0 | 82.8 |

Accuracy of semantic role prediction for known boundaries

# Some results

## Syntactic position/role distribution

- When there is an Arg0, then it is a subject (96.9%), another verbal argument (2.4%) or an object (0.2%)
- The subject of a verb has Arg0 role (79.0%), Arg1 role (16.8%), Arg2 role (2.4%)...

## Automatic semantic labelling:Comparison with FrameNet

| | FrameNet | PropBank | PropBank (>10 ex) |
|---|---|---|---|
| Automatic parses | 82.0 | 79.9 | 80.9 |
| Gold-standard parses | | 82.0 | 82.8 |

Accuracy of semantic role prediction for known boundaries

# Some results

## Syntactic position/role distribution

- When there is an Arg0, then it is a subject (96.9%), another verbal argument (2.4%) or an object (0.2%)
- The subject of a verb has Arg0 role (79.0%), Arg1 role (16.8%), Arg2 role (2.4%)...

## Automatic semantic labelling:Comparison with FrameNet

| | FrameNet | PropBank | PropBank (>10 ex) |
|---|---|---|---|
| Automatic parses | 82.0 | 79.9 | 80.9 |
| Gold-standard parses | | 82.0 | 82.8 |

Accuracy of semantic role prediction for known boundaries

# Automatic semantic labelling

## Comparison with FrameNet

| | FrameNet | | PropBank | | PropBank >10 | |
|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall |
| Automatic parses | 64.6 | 61.2 | 68.6 | 57.8 | 69.9 | 61.1 |
| Gold-standard parses | | | 74.3 | 66.4 | 76.0 | 69.9 |
| Gold-standard parses w/ traces | | | 80.6 | 71.6 | 82.0 | 74.7 |

Accuracy for unknown boundaries

## Comparison using shallow parses

| | Precision | Recall |
|---|---|---|
| Gold. parse | 74.3 | 66.4 |
| Auto parse | 68.6 | 57.8 |
| Chunks | 49.5 | 35.1 |

Accuracy for unknown boundaries

# Automatic semantic labelling

## Comparison with FrameNet

| | FrameNet | | PropBank | | PropBank >10 | |
|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall |
| Automatic parses | 64.6 | 61.2 | 68.6 | 57.8 | 69.9 | 61.1 |
| Gold-standard parses | | | 74.3 | 66.4 | 76.0 | 69.9 |
| Gold-standard parses w/ traces | | | 80.6 | 71.6 | 82.0 | 74.7 |

Accuracy for unknown boundaries

## Comparison using shallow parses

| | Precision | Recall |
|---|---|---|
| Gold. parse | 74.3 | 66.4 |
| Auto parse | 68.6 | 57.8 |
| Chunks | 49.5 | 35.1 |

Accuracy for unknown boundaries

# Time information in corpora

## Motivations[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

### Q&A Systems

Q What is the current unemployment rate?

A Depends on the day (month) the question was asked

Q When did the Berlin Wall fall?

A Thursday

> *East German border workers began dismantling the Berlin Wall at the historic Brandenburg gate on Thursday night to make a new crossing.*

### Formal semantics

Relation between time events and tensed clauses

- Yesterday John saw a girl who was running **this morning**
- This morning John saw a girl who was running yesterday
- Tomorrow John will see a girl who was running earlier

Absolute, relative, anaphoric

# Time information in corpora

Motivations[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

## Q&A Systems

Q What is the current unemployment rate?

A Depends on the day (month) the question was asked

Q When did the Berlin Wall fall?

A Thursday

*East German border workers began dismantling the Berlin Wall at the historic Brandenburg gate on Thursday night to make a new crossing.*

## Formal semantics

Relation between time events and tensed clauses

- Yesterday John saw a girl who was running this morning
- This morning John saw a girl who was running yesterday
- Tomorrow John will see a girl who was running earlier

Absolute, relative, anaphoric

# Time information in corpora

Motivations[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

## Q&A Systems

Q What is the current unemployment rate?

A Depends on the day (month) the question was asked

Q When did the Berlin Wall fall?

A Thursday

*East German border workers began dismantling the Berlin Wall at the historic Brandenburg gate on Thursday night to make a new crossing.*

## Formal semantics

Relation between time events and tensed clauses

- Yesterday John saw a girl who was running this morning
- This morning John saw a girl who was running yesterday
- Tomorrow John will see a girl who was running earlier

Absolute, relative, anaphoric

# Time information in corpora

## Motivations[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

### Q&A Systems

- Q What is the current unemployment rate?
- A Depends on the day (month) the question was asked
- Q When did the Berlin Wall fall?
- A Thursday

*East German border workers began dismantling the Berlin Wall at the historic Brandenburg gate on Thursday night to make a new crossing.*

### Formal semantics

Relation between time events and tensed clauses

- Yesterday John saw a girl who was running **this morning**
- This morning John saw a girl who was running yesterday
- Tomorrow John will see a girl who was running earlier

Absolute, relative, anaphoric

# Time information in corpora

## Motivations[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

### Q&A Systems

Q What is the current unemployment rate?

A Depends on the day (month) the question was asked

Q When did the Berlin Wall fall?

A Thursday

*East German border workers began dismantling the Berlin Wall at the historic Brandenburg gate on Thursday night to make a new crossing.*

### Formal semantics

Relation between time events and tensed clauses

- Yesterday John saw a girl who was running this morning
- This morning John saw a girl who was running yesterday
- Tomorrow John will see a girl who was running earlier

Absolute, relative, anaphoric

# Time information in corpora

## Motivations[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

### Q&A Systems

Q What is the current unemployment rate?

A Depends on the day (month) the question was asked

Q When did the Berlin Wall fall?

A Thursday

> *East German border workers began dismantling the Berlin Wall at the historic Brandenburg gate on Thursday night to make a new crossing.*

### Formal semantics

Relation between time events and tensed clauses

- Yesterday John saw a girl who was running this morning
- This morning John saw a girl who was running yesterday
- Tomorrow John will see a girl who was running earlier

Absolute, relative, anaphoric

# Time information in corpora

## Motivations[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

### Q&A Systems

- Q What is the current unemployment rate?
- A Depends on the day (month) the question was asked
- Q When did the Berlin Wall fall?
- A Thursday

  *East German border workers began dismantling the Berlin Wall at the historic Brandenburg gate on Thursday night to make a new crossing.*

### Formal semantics

Relation between time events and tensed clauses

- Yesterday John saw a girl who was running this morning
- This morning John saw a girl who was running yesterday
- Tomorrow John will see a girl who was running earlier

Absolute, relative, anaphoric

# Time information in corpora

Motivations[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

## Q&A Systems

- Q What is the current unemployment rate?
- A Depends on the day (month) the question was asked
- Q When did the Berlin Wall fall?
- A Thursday

*East German border workers began dismantling the Berlin Wall at the historic Brandenburg gate on Thursday night to make a new crossing.*

## Formal semantics

Relation between time events and tensed clauses

- Yesterday John saw a girl who was running this morning
- This morning John saw a girl who was running yesterday
- Tomorrow John will see a girl who was running earlier

Absolute, relative, anaphoric

# Time information in corpora

Motivations[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

## Q&A Systems

Q What is the current unemployment rate?

A Depends on the day (month) the question was asked

Q When did the Berlin Wall fall?

A Thursday

*East German border workers began dismantling the Berlin Wall at the historic Brandenburg gate on Thursday night to make a new crossing.*

## Formal semantics

Relation between time events and tensed clauses

- Yesterday John saw a girl who was running this morning
- This morning John saw a girl who was running yesterday
- Tomorrow John will see a girl who was running earlier

Absolute, relative, anaphoric

# Time information in corpora

Motivations[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

## Q&A Systems

- Q What is the current unemployment rate?
- A Depends on the day (month) the question was asked
- Q When did the Berlin Wall fall?
- A Thursday

*East German border workers began dismantling the Berlin Wall at the historic Brandenburg gate on Thursday night to make a new crossing.*

## Formal semantics

Relation between time events and tensed clauses

- Yesterday John saw a girl who was running this morning
- This morning John saw a girl who was running yesterday
- Tomorrow John will see a girl who was running earlier

Absolute, relative, anaphoric

# Time information in corpora

Motivations[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

## Q&A Systems

Q What is the current unemployment rate?

A Depends on the day (month) the question was asked

Q When did the Berlin Wall fall?

A Thursday

> *East German border workers began dismantling the Berlin Wall at the historic Brandenburg gate on Thursday night to make a new crossing.*

## Formal semantics

Relation between time events and tensed clauses

- Yesterday John saw a girl who was running this morning
- This morning John saw a girl who was running yesterday
- Tomorrow John will see a girl who was running earlier

Absolute, relative, anaphoric

# Time information in corpora

## Requirements[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

- Theories of tense (and aspects and event structure)
- Annotation scheme and tools: TimeML

**Conceptual and linguistics basis**

Event expressions: introduced by tensed verbs, stative adjectives (on board, stalled), event nominals (World Championship, election)

Dependencies between events and time: anchoring (on Monday), ordering (yesterday), embedding (John said Mary left)

**Anchoring example:** *John taught on Monday*

$$\exists e_1.\mathbf{teach}(e_1, j) \wedge \mathbf{on}(e_1, \mathbf{Monday}) \wedge \mathbf{Past}(e_1)$$

# Time information in corpora

## Requirements[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

- Theories of tense (and aspects and event structure)
- Annotation scheme and tools: TimeML

**Conceptual and linguistics basis**

Event expressions: introduced by tensed verbs, stative adjectives (on board, stalled), event nominals (World Championship, election)

Dependencies between events and time: anchoring (on Monday), ordering (yesterday), embedding (John said Mary left)

**Anchoring example:** *John taught on Monday*

$$\exists e_1.\mathbf{teach}(e_1, j) \wedge \mathbf{on}(e_1, \mathbf{Monday}) \wedge \mathbf{Past}(e_1)$$

# Time information in corpora

## Requirements[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

- Theories of tense (and aspects and event structure)
- Annotation scheme and tools: TimeML

### Conceptual and linguistics basis

Event expressions: introduced by tensed verbs, stative adjectives (on board, stalled), event nominals (World Championship, election)

Dependencies between events and time: anchoring (on Monday), ordering (yesterday), embedding (John said Mary left)

### Anchoring example: *John taught on Monday*

$$\exists e_1.\mathbf{teach}(e_1, j) \wedge \mathbf{on}(e_1, \mathbf{Monday}) \wedge \mathbf{Past}(e_1)$$

# Time information in corpora

## Requirements[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

- Theories of tense (and aspects and event structure)
- Annotation scheme and tools: TimeML

### Conceptual and linguistics basis

**Event expressions:** introduced by tensed verbs, stative adjectives (on board, stalled), event nominals (World Championship, election)

**Dependencies between events and time:** anchoring (on Monday), ordering (yesterday), embedding (John said Mary left)

### Anchoring example: *John taught on Monday*

$$\exists e_1.\mathbf{teach}(e_1, j) \wedge \mathbf{on}(e_1, \mathbf{Monday}) \wedge \mathbf{Past}(e_1)$$

# Time information in corpora

Requirements[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

- Theories of tense (and aspects and event structure)
- Annotation scheme and tools: TimeML

## Conceptual and linguistics basis

Event expressions: introduced by tensed verbs, stative adjectives (on board, stalled), event nominals (World Championship, election)

Dependencies between events and time: anchoring (on Monday), ordering (yesterday), embedding (John said Mary left)

## Anchoring example: *John taught on Monday*

$$\exists e_1.\textbf{teach}(e_1, j) \wedge \textbf{on}(e_1, \textbf{Monday}) \wedge \textbf{Past}(e_1)$$

# Time information in corpora

## Requirements[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

- Theories of tense (and aspects and event structure)
- Annotation scheme and tools: TimeML

### Conceptual and linguistics basis

Event expressions: introduced by tensed verbs, stative adjectives (on board, stalled), event nominals (World Championship, election)

Dependencies between events and time: anchoring (on Monday), ordering (yesterday), embedding (John said Mary left)

### Anchoring example: *John taught on Monday*

$$\exists e_1.\mathbf{teach}(e_1, j) \wedge \mathbf{on}(e_1, \mathbf{Monday}) \wedge \mathbf{Past}(e_1)$$

# Time information in corpora

Requirements[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

**Relation example:** *John said he taught*

$$\exists e_1 e_2 . \mathbf{say}(e_1, j) \wedge \mathbf{teach}(e_2, j) \wedge \mathbf{Past}(e_1) \wedge \mathbf{Past}(e_2) \wedge e_2 < e_1$$

**Temporal expressions**

- Fully specified (November 29, 2007)
- Underspecified (Monday, next month)
- Durations (four weeks)

# Time information in corpora

Requirements[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

## Relation example: *John said he taught*

$$\exists e_1 e_2 . \mathbf{say}(e_1, j) \wedge \mathbf{teach}(e_2, j) \wedge \mathbf{Past}(e_1) \wedge \mathbf{Past}(e_2) \wedge e_2 < e_1$$

## Temporal expressions

- Fully specified (November 29, 2007)
- Underspecified (Monday, next month)
- Durations (four weeks)

# Time information in corpora

Requirements[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

## Relation example: *John said he taught*

$$\exists e_1 e_2 . \mathbf{say}(e_1, j) \wedge \mathbf{teach}(e_2, j) \wedge \mathbf{Past}(e_1) \wedge \mathbf{Past}(e_2) \wedge e_2 < e_1$$

## Temporal expressions

- Fully specified (November 29, 2007)
- Underspecified (Monday, next month)
- Durations (four weeks)

# Time information in corpora

Requirements[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

**Relation example:** *John said he taught*

$$\exists e_1 e_2 . \mathbf{say}(e_1, j) \wedge \mathbf{teach}(e_2, j) \wedge \mathbf{Past}(e_1) \wedge \mathbf{Past}(e_2) \wedge e_2 < e_1$$

**Temporal expressions**

- Fully specified (November 29, 2007)
- Underspecified (Monday, next month)
- Durations (four weeks)

# Time information in corpora

Requirements[Pustejovsky et al.(2003)Pustejovsky, Katz, and Gaizauskas]

### Relation example: *John said he taught*

$$\exists e_1 e_2 . \mathbf{say}(e_1, j) \wedge \mathbf{teach}(e_2, j) \wedge \mathbf{Past}(e_1) \wedge \mathbf{Past}(e_2) \wedge e_2 < e_1$$

### Temporal expressions

- Fully specified (November 29, 2007)
- Underspecified (Monday, next month)
- Durations (four weeks)

# Evolving from TIMEX2 to TimeML

## <TIMEX2>

Attributes:

| Attribute | Function | Example |
|---|---|---|
| VAL | Contains a normalisez form of the date/time | VAL=''2007-11-29'' |
| MOD | Captures temporal modifiers | MOD=''APPROX'' |
| PERIODICITY | Period between regularly recurring times | PERIODICITY=''P1M'' |
| ... | ... | ... |

## Missing

- Underspecified temporal expressions
- Durations
- Expressions relative to the document itself

⇒ TIMEX3

# Evolving from TIMEX2 to TimeML

## <TIMEX2>

Attributes:

| Attribute | Function | Example |
|---|---|---|
| VAL | Contains a normalisez form of the date/time | VAL=''2007-11-29'' |
| MOD | Captures temporal modifiers | MOD=''APPROX'' |
| PERIODICITY | Period between regularly recurring times | PERIODICITY=''P1M'' |
| ... | ... | ... |

## Missing

- Underspecified temporal expressions
- Durations
- Expressions relative to the document itself

⇒ TIMEX3

# Evolving from TIMEX2 to TimeML

## <TIMEX2>

Attributes:

| Attribute | Function | Example |
|---|---|---|
| VAL | Contains a normalisez form of the date/time | VAL=''2007-11-29'' |
| MOD | Captures temporal modifiers | MOD=''APPROX'' |
| PERIODICITY | Period between regularly recurring times | PERIODICITY=''P1M'' |
| ... | ... | ... |

## Missing

- Underspecified temporal expressions
- Durations
- Expressions relative to the document itself

⇒ TIMEX3

# Evolving from TIMEX2 to TimeML

## <TIMEX2>

Attributes:

| Attribute | Function | Example |
|---|---|---|
| VAL | Contains a normalisez form of the date/time | VAL=''2007-11-29'' |
| MOD | Captures temporal modifiers | MOD=''APPROX'' |
| PERIODICITY | Period between regularly recurring times | PERIODICITY=''P1M'' |
| ... | ... | ... |

## Missing

- Underspecified temporal expressions
- Durations
- Expressions relative to the document itself

⇒ TIMEX3

# Evolving from TIMEX2 to TimeML

## <TIMEX2>

Attributes:

| Attribute | Function | Example |
|---|---|---|
| VAL | Contains a normalisez form of the date/time | VAL=''2007-11-29'' |
| MOD | Captures temporal modifiers | MOD=''APPROX'' |
| PERIODICITY | Period between regularly recurring times | PERIODICITY=''P1M'' |
| ... | ... | ... |

## Missing

- Underspecified temporal expressions
- Durations
- Expressions relative to the document itself

$\Rightarrow$ TIMEX3

# The TimeML language

## Markup Language for Temporal and Event Expressions

### TimeML: a specification language for events and temporal expressions in natural language

- Time stamping of events (identifying an event and anchoring it in time)
- Ordering events with respect to one another (lexical versus discourse properties of ordering). `tlink` tag.
- Reasoning with contextually underspecified temporal expressions (temporal functions such as 'last week' and 'two weeks before')
- Reasoning about the persistence of events (how long does an event or the outcome of an event last)

### Additional features

- a `TLINK` to relate events and times, to order events relative to each other and to order times relative to each other.
- `Make Instance` to create instances of event description

# The TimeML language

## Markup Language for Temporal and Event Expressions

### TimeML: a specification language for events and temporal expressions in natural language

- Time stamping of events (identifying an event and anchoring it in time)
- Ordering events with respect to one another (lexical versus discourse properties of ordering). `tlink` tag.
- Reasoning with contextually underspecified temporal expressions (temporal functions such as 'last week' and 'two weeks before')
- Reasoning about the persistence of events (how long does an event or the outcome of an event last)

### Additional features

- a `TLINK` to relate events and times, to order events relative to each other and to order times relative to each other.
- `Make Instance` to create instances of event description

# The TimeML language
## Markup Language for Temporal and Event Expressions

### TimeML: a specification language for events and temporal expressions in natural language

- Time stamping of events (identifying an event and anchoring it in time)
- Ordering events with respect to one another (lexical versus discourse properties of ordering). `tlink` tag.
- Reasoning with contextually underspecified temporal expressions (temporal functions such as 'last week' and 'two weeks before')
- Reasoning about the persistence of events (how long does an event or the outcome of an event last)

### Additional features

- a `TLINK` to relate events and times, to order events relative to each other and to order times relative to each other.
- `Make Instance` to create instances of event description

# The TimeML language

## Markup Language for Temporal and Event Expressions

### TimeML: a specification language for events and temporal expressions in natural language

- Time stamping of events (identifying an event and anchoring it in time)
- Ordering events with respect to one another (lexical versus discourse properties of ordering). `tlink` tag.
- Reasoning with contextually underspecified temporal expressions (temporal functions such as 'last week' and 'two weeks before')
- Reasoning about the persistence of events (how long does an event or the outcome of an event last)

### Additional features

- a `TLINK` to relate events and times, to order events relative to each other and to order times relative to each other.
- `Make Instance` to create instances of event description

# The TimeML language

## Markup Language for Temporal and Event Expressions

### TimeML: a specification language for events and temporal expressions in natural language

- Time stamping of events (identifying an event and anchoring it in time)
- Ordering events with respect to one another (lexical versus discourse properties of ordering). `tlink` tag.
- Reasoning with contextually underspecified temporal expressions (temporal functions such as 'last week' and 'two weeks before')
- Reasoning about the persistence of events (how long does an event or the outcome of an event last)

### Additional features

- a `TLINK` to relate events and times, to order events relative to each other and to order times relative to each other.
- `Make Instance` to create instances of event description

# The TimeML language

## Markup Language for Temporal and Event Expressions

### TimeML: a specification language for events and temporal expressions in natural language

- Time stamping of events (identifying an event and anchoring it in time)
- Ordering events with respect to one another (lexical versus discourse properties of ordering). `tlink` tag.
- Reasoning with contextually underspecified temporal expressions (temporal functions such as 'last week' and 'two weeks before')
- Reasoning about the persistence of events (how long does an event or the outcome of an event last)

### Additional features

- a `TLINK` to relate events and times, to order events relative to each other and to order times relative to each other.
- `Make Instance` to create instances of event description

# The TimeML language

## Markup Language for Temporal and Event Expressions

### TimeML: a specification language for events and temporal expressions in natural language

- Time stamping of events (identifying an event and anchoring it in time)
- Ordering events with respect to one another (lexical versus discourse properties of ordering). `tlink` tag.
- Reasoning with contextually underspecified temporal expressions (temporal functions such as 'last week' and 'two weeks before')
- Reasoning about the persistence of events (how long does an event or the outcome of an event last)

### Additional features

- a `TLINK` to relate events and times, to order events relative to each other and to order times relative to each other.
- `Make Instance` to create instances of event description

# The TimeML language

Markup Language for Temporal and Event Expressions

## TimeML: a specification language for events and temporal expressions in natural language

- Time stamping of events (identifying an event and anchoring it in time)
- Ordering events with respect to one another (lexical versus discourse properties of ordering). `tlink` tag.
- Reasoning with contextually underspecified temporal expressions (temporal functions such as 'last week' and 'two weeks before')
- Reasoning about the persistence of events (how long does an event or the outcome of an event last)

## Additional features

- a `TLINK` to relate events and times, to order events relative to each other and to order times relative to each other.
- `Make Instance` to create instances of event description

# TimeML example[Pratt-Hartmann(2007)]

*After his talk with Mary, John drove to Boston. During the drive he ate a donut.*

After `<EVENT eid=talkJM>`his talk with Mary`</EVENT>`, `<EVENT eid=driveJB>` John drove to Boston `</EVENT>` `<MAKEINST eid=talkJM eiid=`$I_1$`/>` `<MAKEINST eid=driveJB eiid=`$I_2$`/>` `<TLINK eventInst=`$I_1$ `relatedToEventInst=`$I_2$ `relType=BEFORE>`.
During the drive `<EVENT eid=eatJd` he ate a donut `/EVENT>` `<MAKEINST eid=eatJd eiid=`$I_3$`/>` `<TLINK eventInst=`$I_3$ `relatedToEventInst=` $I_2$ `relType=DURING/>`.

## Inference of implicit data

`<TLINK eventInst=`$I_1$ `relatedToEventInst=`$I_3$ `relType=BEFORe/>`

# TimeML example[Pratt-Hartmann(2007)]

*After his talk with Mary, John drove to Boston. During the drive he ate a donut.*

After <EVENT eid=talkJM>his talk with Mary</EVENT>, <EVENT eid=driveJB> John drove to Boston </EVENT> <MAKEINST eid=talkJM eiid=$I_1$/> <MAKEINST eid=driveJB eiid=$I_2$/> <TLINK eventInst=$I_1$ relatedToEventInst=$I_2$ relType=BEFORE>. During the drive <EVENT eid=eatJd he ate a donut /EVENT> <MAKEINST eid=eatJd eiid=$I_3$/> <TLINK eventInst=$I_3$ relatedToEventInst= $I_2$ relType=DURING/>.

Inference of implicit data

<TLINK eventInst=$I_1$ relatedToEventInst=$I_3$ relType=BEFORe/>

# The TimeBank corpus

- 300 texts of media sources from the news domain
- Annotation:
  - an automatic pre-processing step
  - a human annotation step (check out the pre-processing step, introduces new events, time expressions and links)
- statistics:

| | Count | Frequency |
|---|---|---|
| Words | 68555 | 100% |
| Tags (events, timex3, signal) | 11206 | 16.3% |
| Events | 7571 | 11.0% |
| Timex3 | 1423 | 2.1% |
| Signal | 2212 | 3.2 % |
| Links | 8851 | |
| Tlink | 5514 | 62.3% |
| Slink | 3068 | 34.7% |

# The TimeBank corpus

- 300 texts of media sources from the news domain
- Annotation:
  - an automatic pre-processing step
  - a human annotation step (check out the pre-processing step, introduces new events, time expressions and links)
- statistics:

| | Count | Frequency |
|---|---|---|
| Words | 68555 | 100% |
| Tags (events, timex3, signal) | 11206 | 16.3% |
| Events | 7571 | 11.0% |
| Timex3 | 1423 | 2.1% |
| Signal | 2212 | 3.2 % |
| Links | 8851 | |
| Tlink | 5514 | 62.3% |
| Slink | 3068 | 34.7% |

# The TimeBank corpus

- 300 texts of media sources from the news domain
- Annotation:
  - an automatic pre-processing step
  - a human annotation step (check out the pre-processing step, introduces new events, time expressions and links)
- statistics:

| | Count | Frequency |
|---|---|---|
| Words | 68555 | 100% |
| Tags (events, timex3, signal) | 11206 | 16.3% |
| Events | 7571 | 11.0% |
| Timex3 | 1423 | 2.1% |
| Signal | 2212 | 3.2 % |
| Links | 8851 | |
| Tlink | 5514 | 62.3% |
| Slink | 3068 | 34.7% |

# The TimeBank corpus

- 300 texts of media sources from the news domain
- Annotation:
  - an automatic pre-processing step
  - a human annotation step (check out the pre-processing step, introduces new events, time expressions and links)
- statistics:

| | Count | Frequency |
|---|---|---|
| Words | 68555 | 100% |
| Tags (events, timex3, signal) | 11206 | 16.3% |
| Events | 7571 | 11.0% |
| Timex3 | 1423 | 2.1% |
| Signal | 2212 | 3.2 % |
| Links | 8851 | |
| Tlink | 5514 | 62.3% |
| Slink | 3068 | 34.7% |

# The TimeBank corpus

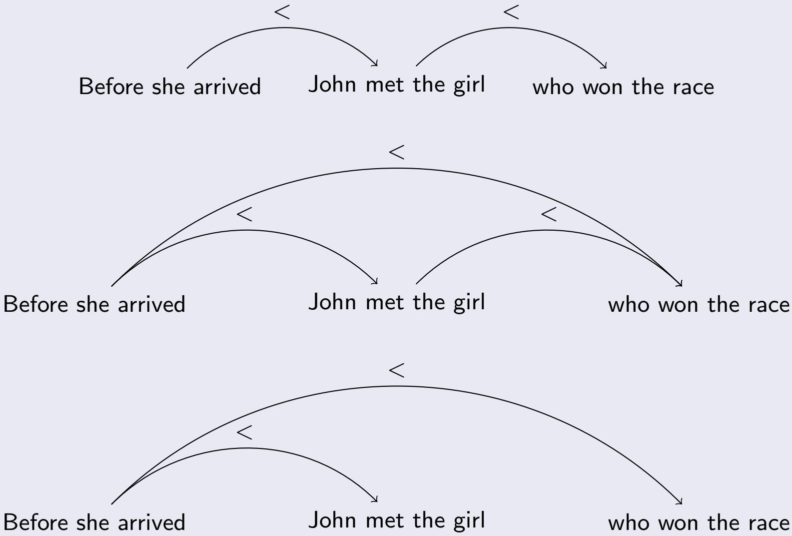
- 300 texts of media sources from the news domain
- Annotation:
  - an automatic pre-processing step
  - a human annotation step (check out the pre-processing step, introduces new events, time expressions and links)
- statistics:

| | **Count** | **Frequency** |
|---|---|---|
| Words | 68555 | 100% |
| Tags (events, timex3, signal) | 11206 | 16.3% |
| Events | 7571 | 11.0% |
| Timex3 | 1423 | 2.1% |
| Signal | 2212 | 3.2 % |
| Links | 8851 | |
| Tlink | 5514 | 62.3% |
| Slink | 3068 | 34.7% |

# Questions on the semantics of the annotation

## Inference properties

### Comparing annotation



# Questions on the semantics of the annotation

## About quantification

*During each of John's drive to Boston he ate a donut.*
*John drove to Boston. During his drive he did not eat a donut.*

## Expressivity and decidability

- Is it possible to make inference?
- Is it possible to add quantification?

→ suggestions on modifying TimeML [Pratt-Hartmann(2007)] in order to support the expressivity of an established temporal logic (Interval Temporal Logic)

# Questions on the semantics of the annotation

## About quantification

*During each of John's drive to Boston he ate a donut.*
*John drove to Boston. During his drive he did not eat a donut.*

## Expressivity and decidability

- Is it possible to make inference?
- Is it possible to add quantification?

→ suggestions on modifying TimeML [Pratt-Hartmann(2007)] in order to support the expressivity of an established temporal logic (Interval Temporal Logic)

# Questions on the semantics of the annotation

## About quantification

*During each of John's drive to Boston he ate a donut.*
*John drove to Boston. During his drive he did not eat a donut.*

## Expressivity and decidability

- Is it possible to make inference?
- Is it possible to add quantification?

→ suggestions on modifying TimeML [Pratt-Hartmann(2007)] in order to support the expressivity of an established temporal logic (Interval Temporal Logic)

# Questions on the semantics of the annotation

## About quantification

*During each of John's drive to Boston he ate a donut.*
*John drove to Boston. During his drive he did not eat a donut.*

## Expressivity and decidability

- Is it possible to make inference?
- Is it possible to add quantification?

→ suggestions on modifying TimeML [Pratt-Hartmann(2007)] in order to support the expressivity of an established temporal logic (Interval Temporal Logic)

# Questions on the semantics of the annotation

## About quantification

*During each of John's drive to Boston he ate a donut.*
*John drove to Boston. During his drive he did not eat a donut.*

## Expressivity and decidability

- Is it possible to make inference?
- Is it possible to add quantification?

→ suggestions on modifying TimeML [Pratt-Hartmann(2007)] in order to support the expressivity of an established temporal logic (Interval Temporal Logic)

# Annotating a corpus with discourse relations

## The Penn Discourse TreeBank example

### Motivations

- To annotate the WSJ corpus in the Penn TreeBank (1 million words) with discourse relations
- To have discourse relations independent from any theory

### Example

Even though critical, it was just the kind of attention they were seeking. **So** they fired back at the Goldman Sachs objections in their own economics letter, "The BMC report".

### Objectives

- Gives a layer of discourse structure
- Enables the comparison of discourse annotation with syntactic annotation in order to highlight relationship between syntactic structure and discourse structure
- Aims to be useful for inference tasks (Q&A)
- Gives a resource for building robust and automatic annotation tools

# Annotating a corpus with discourse relations

## The Penn Discourse TreeBank example

### Motivations

- To annotate the WSJ corpus in the Penn TreeBank (1 million words) with discourse relations
- To have discourse relations independent from any theory

### Example

Even though critical, it was just the kind of attention they were seeking. **So** they fired back at the Goldman Sachs objections in their own economics letter, “The BMC report”.

### Objectives

- Gives a layer of discourse structure
- Enables the comparison of discourse annotation with syntactic annotation in order to highlight relationship between syntactic structure and discourse structure
- Aims to be useful for inference tasks (Q&A)
- Gives a resource for building robust and automatic annotation tools

# Annotating a corpus with discourse relations

## The Penn Discourse TreeBank example

### Motivations

- To annotate the WSJ corpus in the Penn TreeBank (1 million words) with discourse relations
- To have discourse relations independent from any theory

### Example

Even though critical, it was just the kind of attention they were seeking. **So** they fired back at the Goldman Sachs objections in their own economics letter, "The BMC report".

### Objectives

- Gives a layer of discourse structure
- Enables the comparison of discourse annotation with syntactic annotation in order to highlight relationship between syntactic structure and discourse structure
- Aims to be useful for inference tasks (Q&A)
- Gives a resource for building robust and automatic annotation tools

# Annotating a corpus with discourse relations

## The Penn Discourse TreeBank example

### Motivations

- To annotate the WSJ corpus in the Penn TreeBank (1 million words) with discourse relations
- To have discourse relations independent from any theory

### Example

Even though critical, it was just the kind of attention they were seeking. **So** they fired back at the Goldman Sachs objections in their own economics letter, "The BMC report".

### Objectives

- Gives a layer of discourse structure
- Enables the comparison of discourse annotation with syntactic annotation in order to highlight relationship between syntactic structure and discourse structure
- Aims to be useful for inference tasks (Q&A)
- Gives a resource for building robust and automatic annotation tools

# Annotating a corpus with discourse relations

## The Penn Discourse TreeBank example

### Motivations

- To annotate the WSJ corpus in the Penn TreeBank (1 million words) with discourse relations
- To have discourse relations independent from any theory

### Example

Even though critical, it was just the kind of attention they were seeking. **So** they fired back at the Goldman Sachs objections in their own economics letter, "The BMC report".

### Objectives

- Gives a layer of discourse structure
- Enables the comparison of discourse annotation with syntactic annotation in order to highlight relationship between syntactic structure and discourse structure
- Aims to be useful for inference tasks (Q&A)
- Gives a resource for building robust and automatic annotation tools

# Annotating a corpus with discourse relations

## The Penn Discourse TreeBank example

### Motivations

- To annotate the WSJ corpus in the Penn TreeBank (1 million words) with discourse relations
- To have discourse relations independent from any theory

### Example

Even though critical, it was just the kind of attention they were seeking. **So** they fired back at the Goldman Sachs objections in their own economics letter, "The BMC report".

### Objectives

- Gives a layer of discourse structure
- Enables the comparison of discourse annotation with syntactic annotation in order to highlight relationship between syntactic structure and discourse structure
- Aims to be useful for inference tasks (Q&A)
- Gives a resource for building robust and automatic annotation tools

# Annotating a corpus with discourse relations

## The Penn Discourse TreeBank example

### Motivations

- To annotate the WSJ corpus in the Penn TreeBank (1 million words) with discourse relations
- To have discourse relations independent from any theory

### Example

Even though critical, it was just the kind of attention they were seeking. **So** they fired back at the Goldman Sachs objections in their own economics letter, "The BMC report".

### Objectives

- Gives a layer of discourse structure
- Enables the comparison of discourse annotation with syntactic annotation in order to highlight relationship between syntactic structure and discourse structure
- Aims to be useful for inference tasks (Q&A)
- Gives a resource for building robust and automatic annotation tools

# Annotating a corpus with discourse relations

## The Penn Discourse TreeBank example

### Motivations

- To annotate the WSJ corpus in the Penn TreeBank (1 million words) with discourse relations
- To have discourse relations independent from any theory

### Example

Even though critical, it was just the kind of attention they were seeking. **So** they fired back at the Goldman Sachs objections in their own economics letter, "The BMC report".

### Objectives

- Gives a layer of discourse structure
- Enables the comparison of discourse annotation with syntactic annotation in order to highlight relationship between syntactic structure and discourse structure
- Aims to be useful for inference tasks (Q&A)
- Gives a resource for building robust and automatic annotation tools

# Annotating a corpus with discourse relations
## The Penn Discourse TreeBank example

### Motivations

- To annotate the WSJ corpus in the Penn TreeBank (1 million words) with discourse relations
- To have discourse relations independent from any theory

### Example

Even though critical, it was just the kind of attention they were seeking. **So** they fired back at the Goldman Sachs objections in their own economics letter, "The BMC report".

### Objectives

- Gives a layer of discourse structure
- Enables the comparison of discourse annotation with syntactic annotation in order to highlight relationship between syntactic structure and discourse structure
- Aims to be useful for inference tasks (Q&A)
- Gives a resource for building robust and automatic annotation tools

# About discourse Data

**Example (Discourse connective as subordinate conjunction)**

John eats porridge for breakfast, **while** Mary eats muesli.

**Example (DC as adverbial)**

Eat your porridge. **Otherwise** you're not going to football practice.

**Example (DC as PP)**

You've eaten your porridge every day this week. **As a result**, you can go to football practice.

## About discourse Data

**Example (Discourse connective as subordinate conjunction)**

John eats porridge for breakfast, **while** Mary eats muesli.

**Example (DC as adverbial)**

Eat your porridge. **Otherwise** you're not going to football practice.

**Example (DC as PP)**

You've eaten your porridge every day this week. **As a result**, you can go to football practice.

## About discourse Data

**Example (Discourse connective as subordinate conjunction)**

John eats porridge for breakfast, **while** Mary eats muesli.

**Example (DC as adverbial)**

Eat your porridge. **Otherwise** you're not going to football practice.

**Example (DC as PP)**

You've eaten your porridge every day this week. **As a result**, you can go to football practice.

# Multiple relations

**Example**

John loves Barolo. So he ordered three cases of the '97. But he had to cancel the order **because** then he discovered he was broke.

**Example**

John loves Barolo. So he ordered three cases of the '97. But he had to cancel the order because **then** he discovered he was broke.

**Example**

Buyers can look forward to a double-digit annual returns if they are right. **But** they will have disappointing returns or even losses if interest rates rise instead.

**Example**

Buyers can look forward to a double-digit annual returns if they are right. But they will have disappointing returns or even losses if interest rates rise **instead**.

# Multiple relations

## Example

John loves Barolo. So he ordered three cases of the '97. But he had to cancel the order **because** then he discovered he was broke.

## Example

John loves Barolo. So he ordered three cases of the '97. But he had to cancel the order because **then** he discovered he was broke.

## Example

Buyers can look forward to a double-digit annual returns if they are right. **But** they will have disappointing returns or even losses if interest rates rise instead.

## Example

Buyers can look forward to a double-digit annual returns if they are right. But they will have disappointing returns or even losses if interest rates rise **instead**.

# Multiple relations

**Example**

John loves Barolo. So he ordered three cases of the '97. But he had to cancel the order **because** then he discovered he was broke.

**Example**

John loves Barolo. So he ordered three cases of the '97. But he had to cancel the order because **then** he discovered he was broke.

**Example**

Buyers can look forward to a double-digit annual returns if they are right. **But** they will have disappointing returns or even losses if interest rates rise instead.

**Example**

Buyers can look forward to a double-digit annual returns if they are right. But they will have disappointing returns or even losses if interest rates rise **instead**.

## Multiple relations

**Example**

John loves Barolo. So he ordered three cases of the '97. But he had to cancel the order **because** then he discovered he was broke.

**Example**

John loves Barolo. So he ordered three cases of the '97. But he had to cancel the order because **then** he discovered he was broke.

**Example**

Buyers can look forward to a double-digit annual returns if they are right. **But** they will have disappointing returns or even losses if interest rates rise instead.

**Example**

Buyers can look forward to a double-digit annual returns if they are right. But they will have disappointing returns or even losses if interest rates rise **instead**.

# Design choices

## Choice of the corpus

- The same Penn WSJ as the Penn TreeBank $\Rightarrow$ three levels of annotation of the same data (syntactic, semantic and discourse)
- Raw text annotation (not XML) because sometimes discourse args do not align with syntactic structures

## Annotation process

- Through the whole corpus, one connective at a time. Allow annotators to "immediately exploit the experience they were gaining in annotating a connective"
- Ten lists of possible discourse connectives
- Annotation tool to navigate in the instances of the discourse connectives and mark the arguments
- Annotation consist in a**discourse connective**, a first argument and a second argument (the latter being syntactically bound to the dc) ( According to [Webber et al.(2005)Webber, Joshi, Miltsakaki, Prasad, Dinesh, Lee, and Forbes], "In none of the work we carried out (...) did we find an English discourse connective that had anything other than **two** arguments [unlike verbs]")

# Design choices

## Choice of the corpus

- The same Penn WSJ as the Penn TreeBank ⇒ three levels of annotation of the same data (syntactic, semantic and discourse)
- Raw text annotation (not XML) because sometimes discourse args do not align with syntactic structures

## Annotation process

- Through the whole corpus, one connective at a time. Allow annotators to "immediately exploit the experience they were gaining in annotating a connective"
- Ten lists of possible discourse connectives
- Annotation tool to navigate in the instances of the discourse connectives and mark the arguments
- Annotation consist in a**discourse connective**, a first argument and a second argument (the latter being syntactically bound to the dc) ( According to [Webber et al.(2005)Webber, Joshi, Miltsakaki, Prasad, Dinesh, Lee, and Forbes], "In none of the work we carried out (...) did we find an English discourse connective that had anything other than **two** arguments [unlike verbs]")

# Design choices

## Choice of the corpus

- The same Penn WSJ as the Penn TreeBank ⇒ three levels of annotation of the same data (syntactic, semantic and discourse)
- Raw text annotation (not XML) because sometimes discourse args do not align with syntactic structures

## Annotation process

- Through the whole corpus, one connective at a time. Allow annotators to "immediately exploit the experience they were gaining in annotating a connective"
- Ten lists of possible discourse connectives
- Annotation tool to navigate in the instances of the discourse connectives and mark the arguments
- Annotation consist in a **discourse connective**, a first argument and a second argument (the latter being syntactically bound to the dc) ( According to [Webber et al.(2005)Webber, Joshi, Miltsakaki, Prasad, Dinesh, Lee, and Forbes], "In none of the work we carried out (...) did we find an English discourse connective that had anything other than **two** arguments [unlike verbs]")

# Design choices

## Choice of the corpus

- The same Penn WSJ as the Penn TreeBank ⇒ three levels of annotation of the same data (syntactic, semantic and discourse)
- Raw text annotation (not XML) because sometimes discourse args do not align with syntactic structures

## Annotation process

- Through the whole corpus, one connective at a time. Allow annotators to "immediately exploit the experience they were gaining in annotating a connective"
- Ten lists of possible discourse connectives
- Annotation tool to navigate in the instances of the discourse connectives and mark the arguments
- Annotation consist in a**discourse connective**, a first argument and a second argument (the latter being syntactically bound to the dc) ( According to [Webber et al.(2005)Webber, Joshi, Miltsakaki, Prasad, Dinesh, Lee, and Forbes], "In none of the work we carried out (...) did we find an English discourse connective that had anything other than **two** arguments [unlike verbs]")

# Design choices

## Choice of the corpus

- The same Penn WSJ as the Penn TreeBank ⇒ three levels of annotation of the same data (syntactic, semantic and discourse)
- Raw text annotation (not XML) because sometimes discourse args do not align with syntactic structures

## Annotation process

- Through the whole corpus, one connective at a time. Allow annotators to "immediately exploit the experience they were gaining in annotating a connective"
- Ten lists of possible discourse connectives
- Annotation tool to navigate in the instances of the discourse connectives and mark the arguments
- Annotation consist in a**discourse connective**, a first argument and a second argument (the latter being syntactically bound to the dc) ( According to [Webber et al.(2005)Webber, Joshi, Miltsakaki, Prasad, Dinesh, Lee, and Forbes], "In none of the work we carried out (...) did we find an English discourse connective that had anything other than **two** arguments [unlike verbs]" )

# Design choices

## Choice of the corpus

- The same Penn WSJ as the Penn TreeBank ⇒ three levels of annotation of the same data (syntactic, semantic and discourse)
- Raw text annotation (not XML) because sometimes discourse args do not align with syntactic structures

## Annotation process

- Through the whole corpus, one connective at a time. Allow annotators to "immediately exploit the experience they were gaining in annotating a connective"
- Ten lists of possible discourse connectives
- Annotation tool to navigate in the instances of the discourse connectives and mark the arguments
- Annotation consist in a**discourse connective**, a first argument and a second argument (the latter being syntactically bound to the dc) ( According to [Webber et al.(2005)Webber, Joshi, Miltsakaki, Prasad, Dinesh, Lee, and Forbes], "In none of the work we carried out (...) did we find an English discourse connective that had anything other than **two** arguments [unlike verbs]")

# Design choices

## Choice of the corpus

- The same Penn WSJ as the Penn TreeBank ⇒ three levels of annotation of the same data (syntactic, semantic and discourse)
- Raw text annotation (not XML) because sometimes discourse args do not align with syntactic structures

## Annotation process

- Through the whole corpus, one connective at a time. Allow annotators to "immediately exploit the experience they were gaining in annotating a connective"
- Ten lists of possible discourse connectives
- Annotation tool to navigate in the instances of the discourse connectives and mark the arguments
- Annotation consist in a**discourse connective**, a first argument and a second argument (the latter being syntactically bound to the dc) ( According to [Webber et al.(2005)Webber, Joshi, Miltsakaki, Prasad, Dinesh, Lee, and Forbes], "In none of the work we carried out (...) did we find an English discourse connective that had anything other than **two** arguments [unlike verbs]")

# Design choices

## Choice of the corpus

- The same Penn WSJ as the Penn TreeBank ⇒ three levels of annotation of the same data (syntactic, semantic and discourse)
- Raw text annotation (not XML) because sometimes discourse args do not align with syntactic structures

## Annotation process

- Through the whole corpus, one connective at a time. Allow annotators to "immediately exploit the experience they were gaining in annotating a connective"
- Ten lists of possible discourse connectives
- Annotation tool to navigate in the instances of the discourse connectives and mark the arguments
- Annotation consist in a**discourse connective**, a first argument and a second argument (the latter being syntactically bound to the dc) ( According to [Webber et al.(2005)Webber, Joshi, Miltsakaki, Prasad, Dinesh, Lee, and Forbes], "In none of the work we carried out (...) did we find an English discourse connective that had anything other than **two** arguments [unlike verbs]")

# About annotation

## Form of the annotation

- Span list (words)

**Example**

But, says Mr. Dinkins, he did get an office. So he shouldn't complain.

**Example**

On the one hand, Mr. Giuliani wants to cut into Mr. Dinkin's credibility. On the other, he seeks to convince voters he's the new Fiorello LaGuardia—affable, good-natured and ready to lead New-York out of the mess it's in.

- Automatically computed Gorn address list (browse an example )
- Possible sup(plementary)-arguments.

# About annotation

## Form of the annotation

- Span list (words)

### Example

But, says Mr. Dinkins, he did get an office. **So** he shouldn't complain.

### Example

**On the one hand**, Mr. Giuliani wants to cut into Mr. Dinkin's credibility. **On the other**, he seeks to convince voters he's the new Fiorello LaGuardia—affable, good-natured and ready to lead New-York out of the mess it's in.

- Automatically computed Gorn address list (browse an example )
- Possible sup(plementary)-arguments.

# About annotation

## Form of the annotation

- Span list (words)

### Example

But, says Mr. Dinkins, he did get an office. **So** he shouldn't complain.

### Example

**On the one hand**, Mr. Giuliani wants to cut into Mr. Dinkin's credibility. **On the other**, he seeks to convince voters he's the new Fiorello LaGuardia—affable, good-natured and ready to lead New-York out of the mess it's in.

- Automatically computed Gorn address list (browse an example )
- Possible sup(plementary)-arguments.

# About annotation

## Form of the annotation

- Span list (words)

### Example

But, says Mr. Dinkins, he did get an office. **So** he shouldn't complain.

### Example

**On the one hand**, Mr. Giuliani wants to cut into Mr. Dinkin's credibility. **On the other**, he seeks to convince voters he's the new Fiorello LaGuardia—affable, good-natured and ready to lead New-York out of the mess it's in.

- Automatically computed Gorn address list (browse an example )
- Possible sup(plementary)-arguments.

# About annotation

## Form of the annotation

- Span list (words)

### Example

But, says Mr. Dinkins, he did get an office. **So** he shouldn't complain.

### Example

**On the one hand**, Mr. Giuliani wants to cut into Mr. Dinkin's credibility. **On the other**, he seeks to convince voters he's the new Fiorello LaGuardia—affable, good-natured and ready to lead New-York out of the mess it's in.

- Automatically computed Gorn address list (browse an example )
- Possible sup(plementary)-arguments.

# About annotation

## Form of the annotation

- Span list (words)

### Example

But, says Mr. Dinkins, he did get an office. **So** he shouldn't complain.

### Example

**On the one hand**, Mr. Giuliani wants to cut into Mr. Dinkin's credibility. **On the other**, he seeks to convince voters he's the new Fiorello LaGuardia—affable, good-natured and ready to lead New-York out of the mess it's in.

- Automatically computed Gorn address list (browse an example )
- Possible sup(plementary)-arguments.

# Facts about the corpus

## Inter-annotator agreement

- 4 annotators, then 2
- 10 explicit connectives (2717 tokens)
- Exact match agreement: 90.2%
- Partial agreement (arg overlapping): 94.5%

### Example (Parenthetical in the middle of an argument)

Bankers said warrants for Hong Kong stocks are attractive **because** they give foreign investors, wary of volatility in the colony's stock market, an opportunity to buy shares without taking too great a risk.

Bankers said warrants for Hong Kong stocks are attractive **because** they give foreign investors, wary of volatility in the colony's stock market, an opportunity to buy shares without taking too great a risk.

Hence “minimality” principle and sup1 and sup2 tags.

# Facts about the corpus

## Inter-annotator agreement

- 4 annotators, then 2
- 10 explicit connectives (2717 tokens)
- Exact match agreement: 90.2%
- Partial agreement (arg overlapping): 94.5%

### Example (Parenthetical in the middle of an argument)

Bankers said warrants for Hong Kong stocks are attractive **because** they give foreign investors, wary of volatility in the colony's stock market, an opportunity to buy shares without taking too great a risk.

Bankers said warrants for Hong Kong stocks are attractive **because** they give foreign investors, wary of volatility in the colony's stock market, an opportunity to buy shares without taking too great a risk.

Hence "minimality" principle and sup1 and sup2 tags.

# Facts about the corpus

## Inter-annotator agreement

- 4 annotators, then 2
- 10 explicit connectives (2717 tokens)
- Exact match agreement: 90.2%
- Partial agreement (arg overlapping): 94.5%

### Example (Parenthetical in the middle of an argument)

Bankers said warrants for Hong Kong stocks are attractive **because** they give foreign investors, wary of volatility in the colony's stock market, an opportunity to buy shares without taking too great a risk.

Bankers said warrants for Hong Kong stocks are attractive **because** they give foreign investors, wary of volatility in the colony's stock market, an opportunity to buy shares without taking too great a risk.

Hence "minimality" principle and sup1 and sup2 tags.

# Facts about the corpus

## Inter-annotator agreement

- 4 annotators, then 2
- 10 explicit connectives (2717 tokens)
- Exact match agreement: 90.2%
- Partial agreement (arg overlapping): 94.5%

## Example (Parenthetical in the middle of an argument)

Bankers said warrants for Hong Kong stocks are attractive **because** they give foreign investors, wary of volatility in the colony's stock market, an opportunity to buy shares without taking too great a risk.

Bankers said warrants for Hong Kong stocks are attractive **because** they give foreign investors, wary of volatility in the colony's stock market, an opportunity to buy shares without taking too great a risk.

Hence "minimality" principle and sup1 and sup2 tags.

# Facts about the corpus

## Inter-annotator agreement

- 4 annotators, then 2
- 10 explicit connectives (2717 tokens)
- Exact match agreement: 90.2%
- Partial agreement (arg overlapping): 94.5%

## Example (Parenthetical in the middle of an argument)

Bankers said warrants for Hong Kong stocks are attractive **because** they give foreign investors, wary of volatility in the colony's stock market, an opportunity to buy shares without taking too great a risk.

Bankers said warrants for Hong Kong stocks are attractive **because** they give foreign investors, wary of volatility in the colony's stock market, an opportunity to buy shares without taking too great a risk.

Hence "minimality" principle and sup1 and sup2 tags.

# Facts about the corpus

## Inter-annotator agreement

- 4 annotators, then 2
- 10 explicit connectives (2717 tokens)
- Exact match agreement: 90.2%
- Partial agreement (arg overlapping): 94.5%

## Example (Parenthetical in the middle of an argument)

Bankers said warrants for Hong Kong stocks are attractive **because** they give foreign investors, wary of volatility in the colony's stock market, an opportunity to buy shares without taking too great a risk.

Bankers said warrants for Hong Kong stocks are attractive **because** they give foreign investors, wary of volatility in the colony's stock market, an opportunity to buy shares without taking too great a risk.

Hence "minimality" principle and sup1 and sup2 tags.

# Facts about the corpus

## Implicit connectives

Relations with no discourse connective

**Example (Implicit because)**

"We like to make our own judgements" about Mr. Morishita, says C. D. People have a different reputation country by country.

- Exact match: 72%
- Partial match: 92.6%

⇒ distribution of the location of the dc and its arguments [Prasad et al.(2004)Prasad, Miltsakaki, Joshi, and Webber].

# Facts about the corpus

## Implicit connectives

Relations with no discourse connective

**Example (Implicit because)**

"We like to make our own judgements" about Mr. Morishita, says C. D. People have a different reputation country by country.

- Exact match: 72%
- Partial match: 92.6%

⇒ distribution of the location of the dc and its arguments [Prasad et al.(2004)Prasad, Miltsakaki, Joshi, and Webber].

# Facts about the corpus

## Implicit connectives

Relations with no discourse connective

**Example (Implicit because)**

"We like to make our own judgements" about Mr. Morishita, says C. D. People have a different reputation country by country.

- Exact match: 72%
- Partial match: 92.6%

⇒ distribution of the location of the dc and its arguments [Prasad et al.(2004)Prasad, Miltsakaki, Joshi, and Webber].

# Statistics for Corpus Linguistics

- Introduction
- Representativity
- Hypthesis Testing

# Introduction

## Some terminology

# Central Tendancy Measures

Role: To give the "typical" value of the grouped data

## Which measure?

- Mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ or } \bar{x} = \frac{\sum_{i=1}^{n} f(x_i)x_1}{n}$$

  with $f$ frequency distribution.
- Median: let us assume $x_{i-1} \leq x_i$. Then

$$M = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}}}{2} & \text{if } n \text{ is even} \end{cases} \text{ or } x_{i_2} \text{s.t.} \sum_{1}^{i_2 - 1} f(x_i) < \frac{n}{2} \text{ and } \sum_{1}^{i_2} f(x_i) \geq \frac{n}{2}$$

- Mode: value with the highest frequency

# Central Tendancy Measure

| | Mean | Median | Mode |
|---|---|---|---|
| Ratio variable | • | • | • |
| Interval variables | • | • | • |
| Ordinal variables | — | • | • |
| Nominal variables | — | — | • |

## Do they give the right tendancy?

Etreme values
Skewed distribution
Symetric distribution $\Rightarrow$ mode $=$ median $=$ mean

# Central Measure Tendancy
## Measuring Variability

- Range: $w = \max_i(x_i) - \min_i(x_i)$
- Interquartile range:
  - $x_{i_1}$ s.t. $\sum_1^{i_1-1} f(x_i) < \frac{n}{2}$ and $\sum_1^{i_1} f(x_i) \geq \frac{n}{4}$
  - $x_{i_2}$ s.t. $\sum_1^{i_2-1} f(x_i) < \frac{n}{2}$ and $\sum_1^{i_2} f(x_i) \geq \frac{2n}{4}$
  - $x_{i_3}$ s.t. $\sum_1^{i_3-1} f(x_i) < \frac{n}{2}$ and $\sum_1^{i_3} f(x_i) \geq \frac{3n}{4}$

  $$w_{1,3} = x_{i_3} - x_{i_1}$$
- Variance: $\sigma^2 = \frac{\sum_1^n (x-\bar{x})^2}{n}$
- Standard deviation: $\sigma$

# The Laplace-Gauss Distribution

## A.k.a The normal distribution

$$LG(\mu, \sigma) : f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Tables

## Bilbliographie choisie

- [McEnery and Wilson(1996)]
- [Daille and Romary(2001)]
- [Oakes(1998)]
- [Butler(1985)]

D. Abercrombie.
*Studies in Phonetics and Linguistics*.
Oxford University Press, 1965.

D. Biber.
Representativeness in corpus design.
*Literary and Linguistic Computing*, 8(4):243–257, 1993.
URL http://llc.oxfordjournals.org/cgi/content/abstract/8/4/243.

C. Butler.
*Statistics in Linguistics*.
Blackwell Publishers, 1985.
Available at
http://www.uwe.ac.uk/hlss/llas/statistics-in-linguistics/bkindex.shtml.

B. Daille and L. Romary, editors.
*Les linguistiques de corpus*, volume 42(2) of *Traitement Automatique de Langues (T.A.L.)*.
Hermès, Paris, 2001.

H. S. Eaton, editor.
*Semantic Frequency List for English, French, German and Spanish: A Correlation of the First Six Thousand Words in Four Single-Language Frequency Lists*.
Chicago University Press, 1940.

C. J. Fillmore.
Corpus linguistics' vs. 'computer-aided armchair linguistics'.
In *Directions in Corpus Linguistics*. 1992.

C. C. Fries.
*The Structure of English: An Introduction to the Construction of Sentences*.
Harcourt-Brace, New-York, 1952.

P. R. Georges Gougenheim, René Michea and A. Sauvageot.
*L'élaboration du franç ais élémentaire : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*.
Didier, Paris, 1956.

R. A. Harris.
*The Linguistics Wars*.
Oxford University Press, 1993.

A. A. Hill, editor.
*The Third Texas Conference on Problems of Linguistic Analysis in English*, 1962.
University of Texas Press.

C. F. Hockett.
A note on structure.
*International Journal of American Linguistics*, 14, 1948.

C. F. Hockett.
Sound change.
*Language*, 41, 1964.

R. A. Hudson.
*English Word Grammar*.
B. Blackwell, Oxford, UK, 1990.

J. Käding.
*Häufigkeitswörterbuch der deutschen Sprache*.
privately published, 1897.

G. Leech.
Corpora and theories of linguistic performance.
In J. Svartvik, editor, *Directions in Corpus Linguistics*. 1992.

G. Leech.
Corpus annotation schemes.
*Literary and Linguistic Computing*, 8(4), 1993.

C. Lupovici.
Révolution électronique en normalisation.
*Bulletin des Bibliothéquaires de France*, 38(5), 1993.

T. McEnery and M. Oakes.

Sentence and word alignment in the CRATER project.
In J. Thomas and M. Short, editors, *Using Corpora for Language Research*. Longman, 1996.

T. McEnery and A. Wilson.
*Corpus Linguistics*.
Edinburgh University Press, 1996.

G. V. Morrill.
Incremental processing and acceptability.
*Computational Linguistics*, 26(3):319–338, Sept. 2000.

M. P. Oakes.
*Statistics for Corpus Linguistics*.
Edinburgh Textbooks in Empirical Linguistics. Edinburgh University Press, 1998.

M. Palmer, D. Gildea, and P. Kingsbury.
The proposition bank: A corpus annotated with semantic roles.
*Computational Linguistics*, 31(1), 2005.

R. Prasad, E. Miltsakaki, A. Joshi, and B. Webber.
Annotation and data mining of the penn discourse treebank.
In *Proceedings of the ACL Workshop on Discourse Annotation*, Barcelona, Spain, 2004.

I. Pratt-Hartmann.
From timeml to interval temporal logic.
In *Proceedings of IWCS*, 2007.

W. Preyer.
*The Mind of a Child.*
Appeltown, NewYork, 1889.

J. Pustejovsky, G. Katz, and R. Gaizauskas.
Practical applications of temporal and event reasoning, 2003.
http://www.cs.brandeis.edu/~jamesp/arda/time/esslli.html.

J. Véronis.
A study of polysemy judgements and inter-annotator agreement.
In *Programme and advanced papers of the Senseval workshop*, 1998.
URL http://www.up.univ-mrs.fr/veronis/pdf/1998senseval.pdf.

B. Webber, A. Joshi, E. Miltsakaki, R. Prasad, N. Dinesh, A. Lee, and K. Forbes.
A short introduction to the penn discourse treebank.
In *Copenhagen Working Papers in Language and Speech Processing*, 2005.
http://www.seas.upenn.edu/~pdtb/pdtb-corpus-1.0/papers/webber-etal-nodalida05.pdf.