LABORATOIRE D'INFORMATIQUE GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

June 2016
DICORA
Korea

Local Grammars

Background and Significance

Éric Laporte

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Local Grammars

Unitex/GramLab is an open-source corpus processor

It provides tools for creating and managing

- local grammars
- dictionaries

Which benefits do these tools provide to projects?

How do they combine with current linguistic approaches?

How do they contribute to modern natural language processing (NLP)?

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Outline

Accuracy

Attention to the lexicon

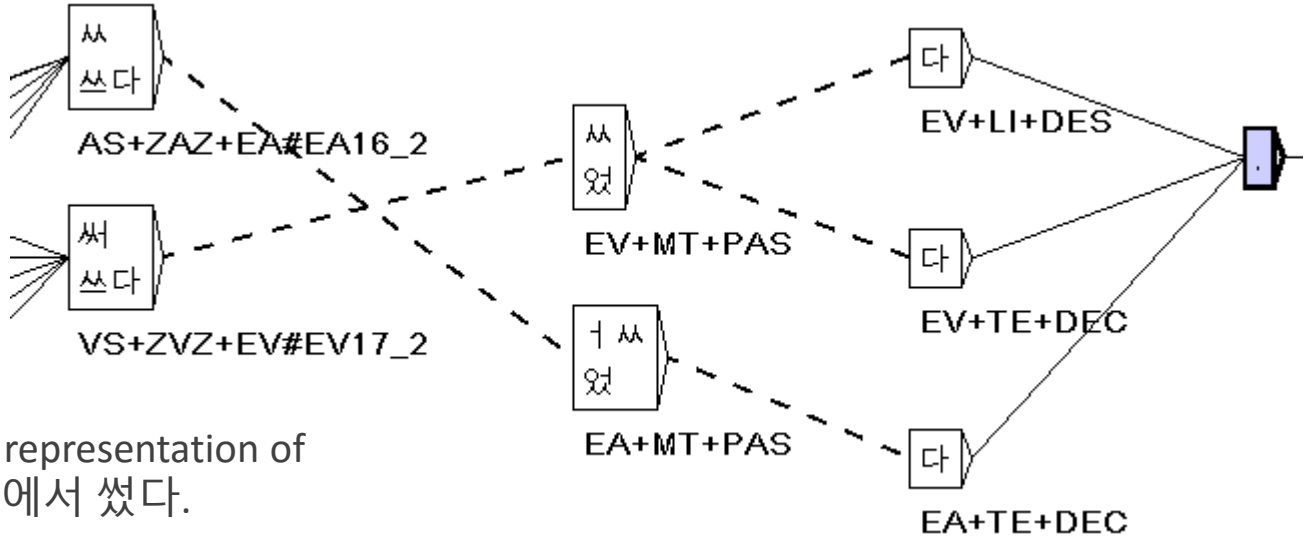
Readability

Formalization

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Accuracy



A Unitex representation of
(…) 목적에서 썼다.

Accurate and detailed linguistic description of

- **morphosyntax**
- sense distinctions
- phrases

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Accuracy

```
coréen,.A+Toponyme+Territoire+Pays:ms
Coréen,.N+Hum+Toponyme+Territoire+Pays:ms
coréen,.N+Langue:ms
coréenne,.A+Toponyme+Territoire+Pays:fs
Coréenne,.N+Hum+Toponyme+Territoire+Pays:fs
```

Adjective

Human noun

Language name

Sample of a Unitex/GramLab dictionary of French

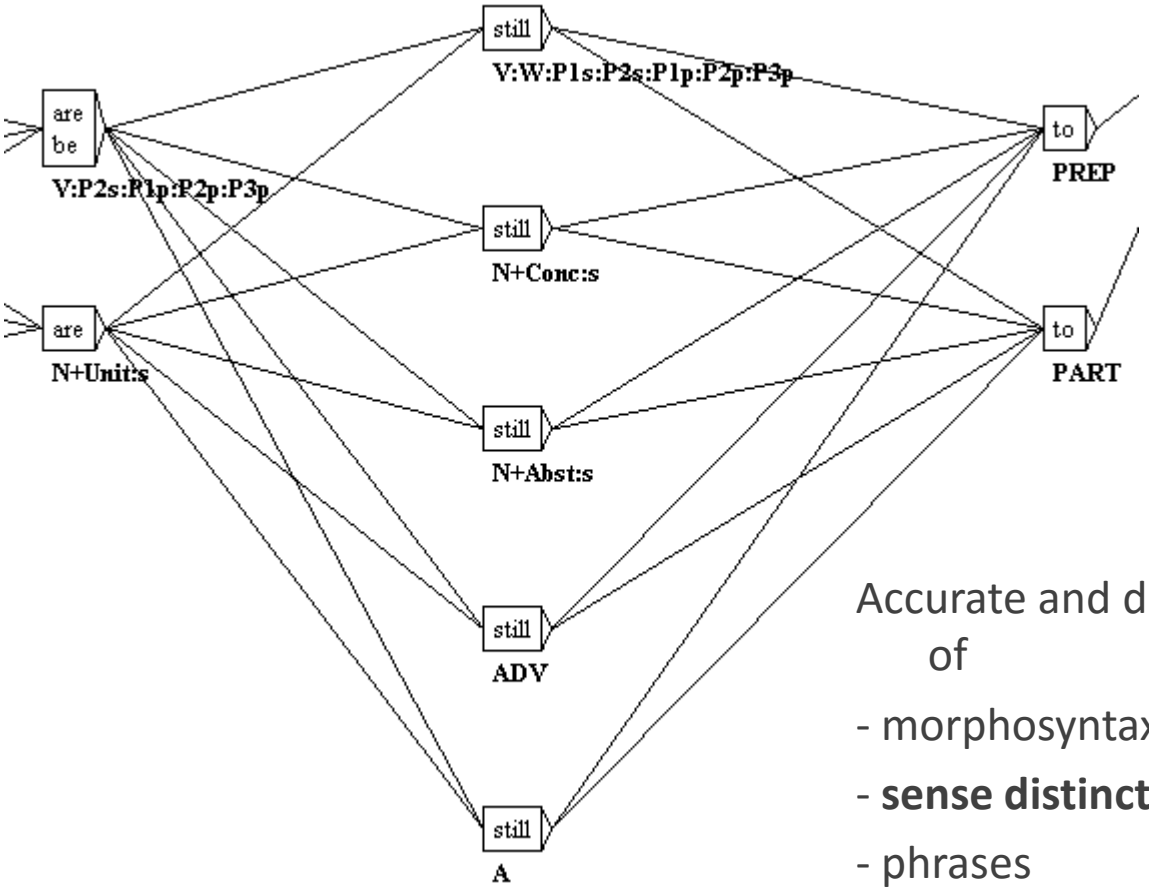
Accurate and detailed linguistic description of

- morphosyntax
- **sense distinctions**
- phrases

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Accuracy



A Unitex representation of
(...) *are still to be seen* (...)

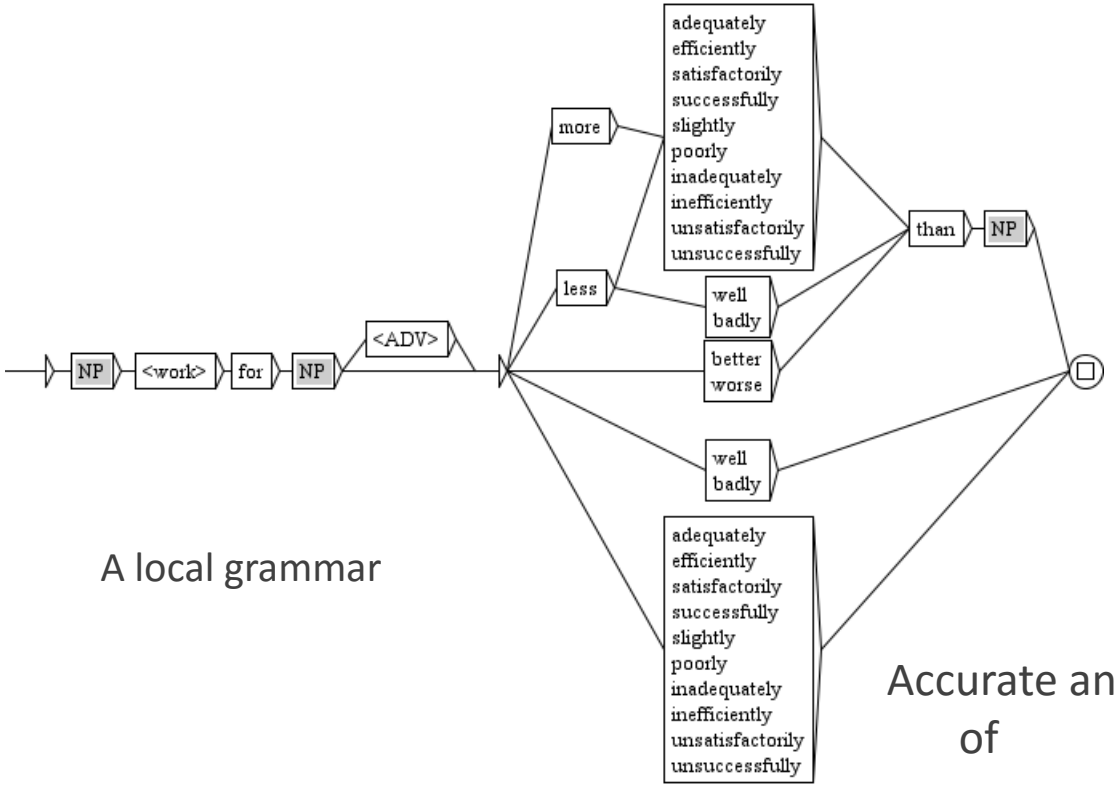
Accurate and detailed linguistic description of

- morphosyntax
- **sense distinctions**
- phrases

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Accuracy



A local grammar

Accurate and detailed linguistic description of

- morphosyntax
- sense distinctions
- **phrases** (sequences)

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

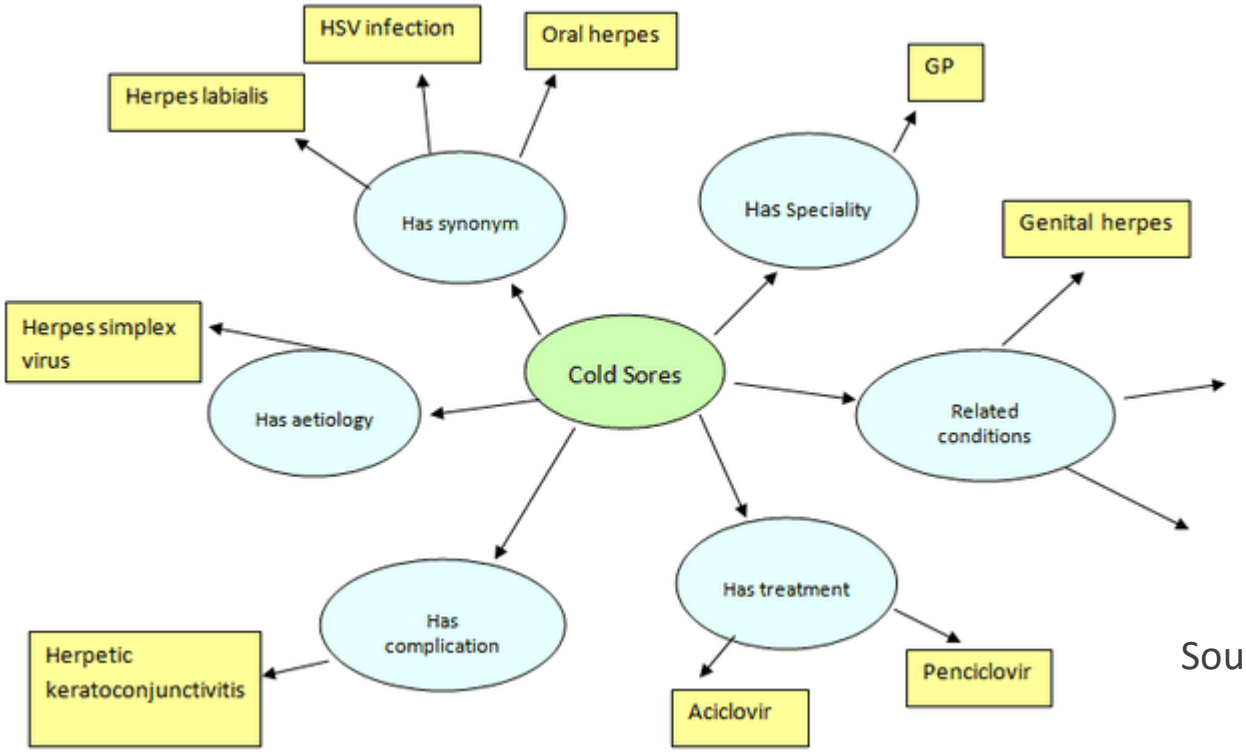
Accuracy

Can the Unitex/GramLab approach combine with others?

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Symbolic approaches



A medical ontology
Source: Merrall-Ross, 2012-2013

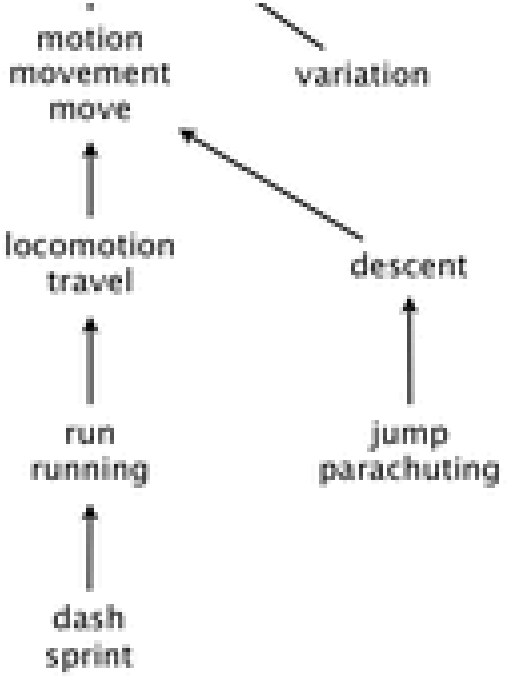
Unitex/GramLab resources are complementary to other resources for symbolic approaches

- ontologies: assume morphological analysis has been done before processing

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Symbolic approaches



A sample of the Princeton WordNet

Other resources for symbolic approaches

- WordNet and KorLex: assume morphological analysis has been done before processing; do not process sequences

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Computer-only approaches

The Unitex/GramLab approach is complementary to purely computational approaches

Supervised learning

Uses annotated or aligned corpora

Unitex/GramLab aids annotating and aligning corpora more accurately

Text categorization (Ko & Seo, 2011)

Computational semantic analysis (Kim et al., 2014)

Word sense induction (Li, 2013)

Unsupervised learning

Often uses available related information

Example: learning semantic roles is easier on the basis of delimited and labeled suffixes (Nam & Kim, 2016); text clustering probably too

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Computer-only approaches

Less labour-intensive than the Unitex/GramLab approach

Computers were designed to substitute human labour

The substitution may have a cost in quality

Example: accuracy

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Outline

Accuracy

Attention to the lexicon

Readability

Formalization

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

The lexicon

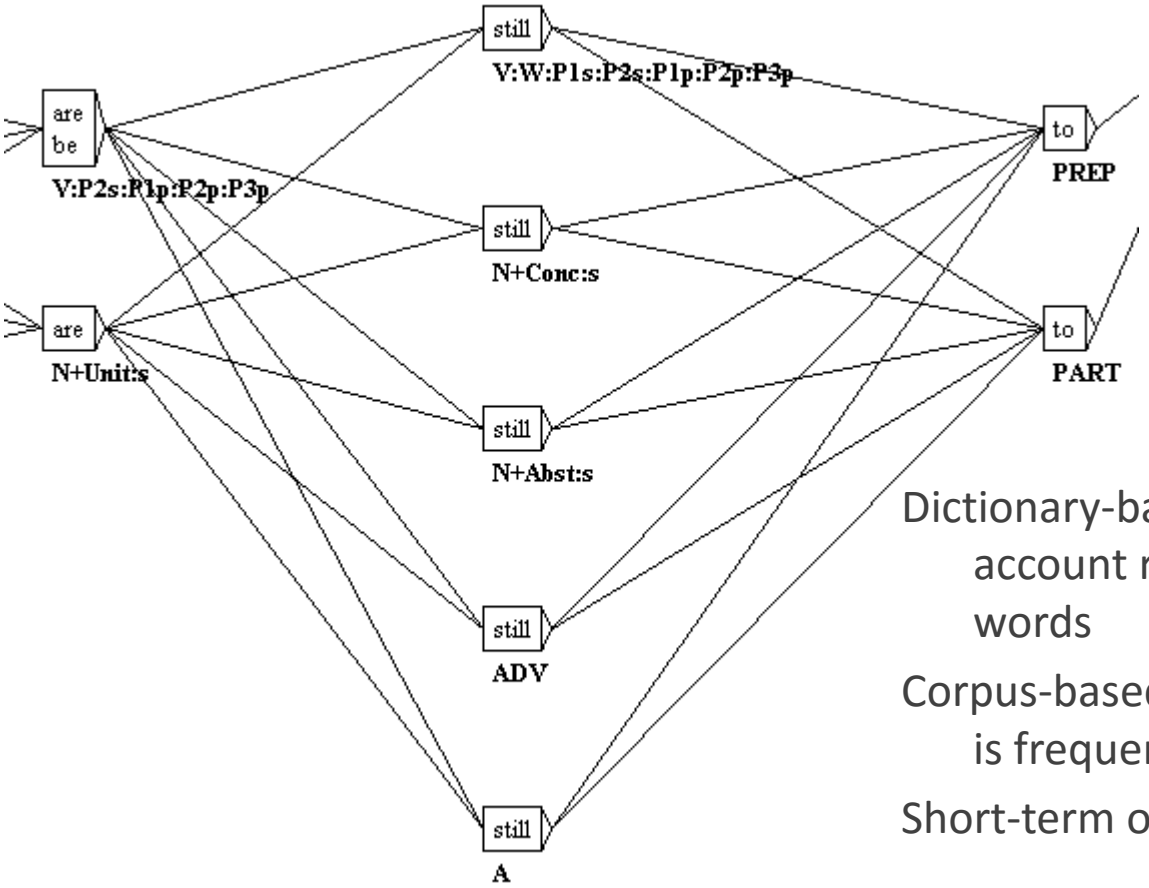
Approach	Unitex/GramLab	Supervised learning	Hybrid
Typical resource	dictionary	annotated corpora	both

Dictionary-based processing is complementary to corpus-based processing

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

The lexicon

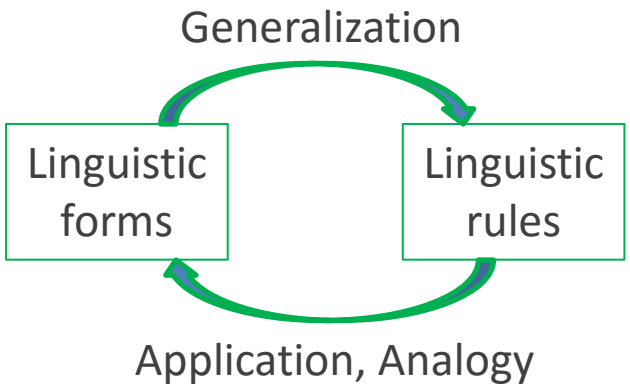


- Dictionary-based processing may take into account rare constructions and rare words
- Corpus-based processing focuses on what is frequent in the context of a project
- Short-term or long-term approach

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

The lexicon/corpus duality



Corpus linguistics focuses on linguistic forms

Generative linguistics focuses on linguistic rules

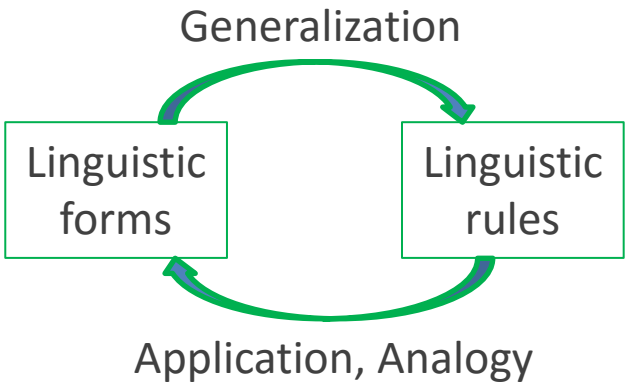
Both are relevant to linguistics

Both are relevant to NLP

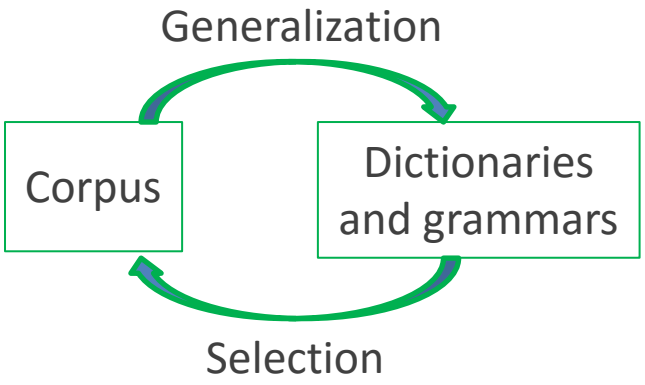
LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

The lexicon/corpus duality



In our minds



In linguistic description

Dictionaries and grammars may include rare words and constructions

A corpus is incomplete, but helps making dictionaries and grammars more complete

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Outline

Accuracy

Attention to the lexicon

Readability

Formalization

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

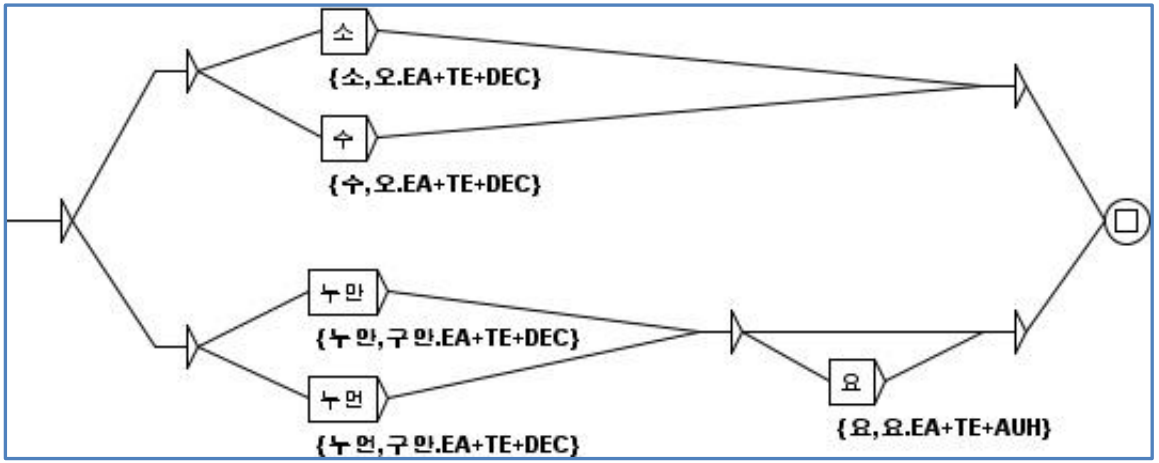
Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Readability

Readability of language resources is overlooked by most computational approaches

The descriptive linguist's point of view

Regular expressions are less readable than graphs



A Unitex/GramLab graph

소/{소,오.EA+TE+DEC}|수/{수,오.EA+TE+DEC}|(누만/{누만,구만.EA+TE+DEC}|누먼/{누먼,구만.EA+TE+DEC})(ε|요/{요,요.EA+TE+AUH})

An equivalent regular expression

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Readability and manageability of rules for morphosyntax

Approach	XFST	Unitex/GramLab
Rules apply to	all entries in a POS	specific lexical items

General rules for all items in a part-of-speech are less manageable than rules that apply to specific items

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Readability and manageability of rules for morphosyntax

듣+어 ⇒ 들어 "to hear"	자르+어 ⇒ 잘라 "to cut"	젓+어 ⇒ 저어 "to stir"	가+어 ⇒ 가 "to go"	rewrite rule
tuT.+E.	ca.Lu.+E.	ceS.+E.	ka.+E.	vowel harmony E -> a || [o|a] (CON|. CON u) . + _; E -> e;
tuT.+e.	ca.Lu.+a.	ceS.+e.	ka.+a.	u deletion u . + -> [..] || CON _ [e|a];
n/a	ca.La.	n/a	n/a	vowel merge a . + -> [..], e . + -> [..] || \w _ VOW
n/a	n/a	n/a	ka.	"T" irregular T -> l || _ . + E;
tul.+e.	n/a	n/a	n/a	

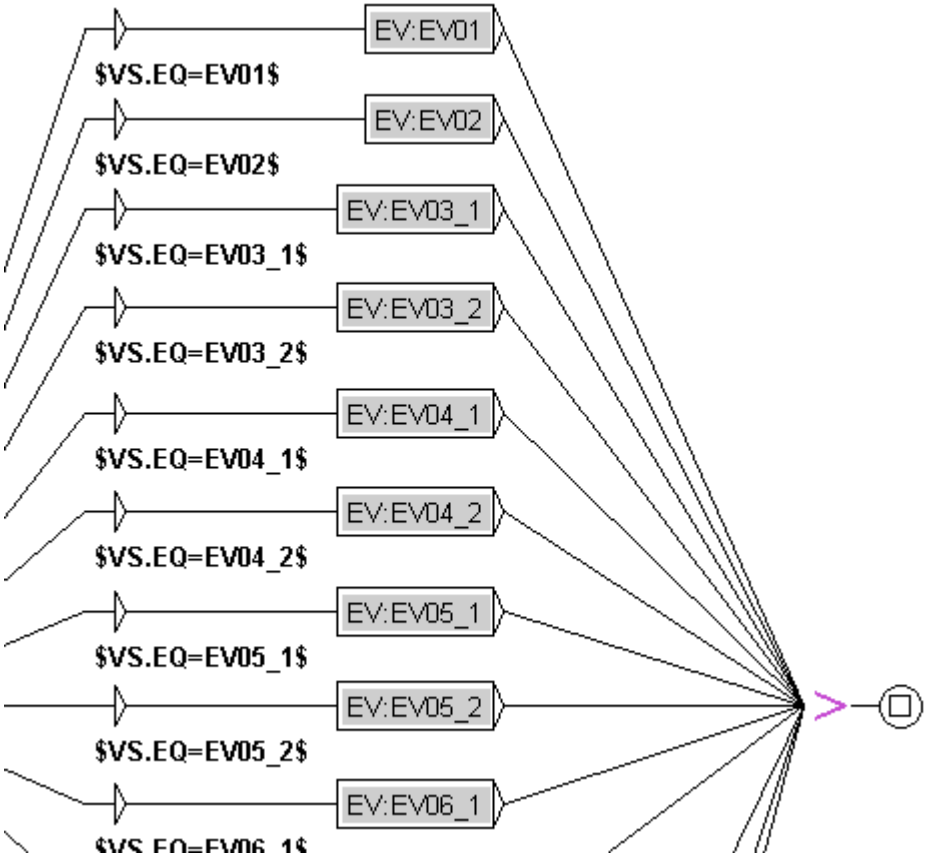
Source: Han, 2006

- XFST-style rules for morphosyntax
- General rules designed to work for all Korean verbs
- All must be checked for each new verb

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Readability and manageability of rules for morphosyntax



Source: Nam, 2009

- Unitex/GramLab rules for morphosyntax
- Each rule is designed for specifically marked verbs
- Each rule is independent

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Outline

Accuracy

Attention to the lexicon

Readability

Formalization

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Formalization

```
coréen,.A+Toponyme+Territoire+Pays:ms
Coréen,.N+Hum+Toponyme+Territoire+Pays:ms
coréen,.N+Langue:ms
coréenne,.A+Toponyme+Territoire+Pays:fs
Coréenne,.N+Hum+Toponyme+Territoire+Pays:fs
```

Sample of a Unitex/GramLab dictionary of French

Language resources for NLP require more formalization than other domains in linguistics

Fields

Categories

Delimiters

Codes

Formalization

Table 3. Frequency of *com* in Speech Levels

Speech Level	Mitigation Marker	Intensification Marker	Total
Polite	**35 (75%)**	12 (25%)	47
Plain	35 (40%)	**52 (60%)**	87
Total	70 (52%)	64 (48%)	134

Source: Ahn, 2009

A table of data

Categories: polite/plain, mitigation/intensification

A rare example of formalized data in discourse pragmatics

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Historical background

Unitex/GramLab is inspired by Maurice Gross' (1934-2001) approach to descriptive syntax and morphosyntax

Accuracy

The meaning of a text may depend on details

Attention to the lexicon

Each lexical item may be different (Gross, 1975)

Readability

Only readable data are manageable by linguists

Formalization

Only models are manageable by computers

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Thanks

CONTACT

ÉRIC LAPORTE

00 +33 (0)1 60 95 75 52

ERIC.LAPORTE@UNIV-PARIS-EST.FR