Linguistic and Computational Morphology[1]

Agata Savary

December 19, 2008

[1] Master in Information Systems and Decision Support, Faculty of Computer Science, Blois

Morphology

Linguistic discipline interested in the internal structure of (written) words.

What is a word ?

In **linguistics** a **word** has two senses:

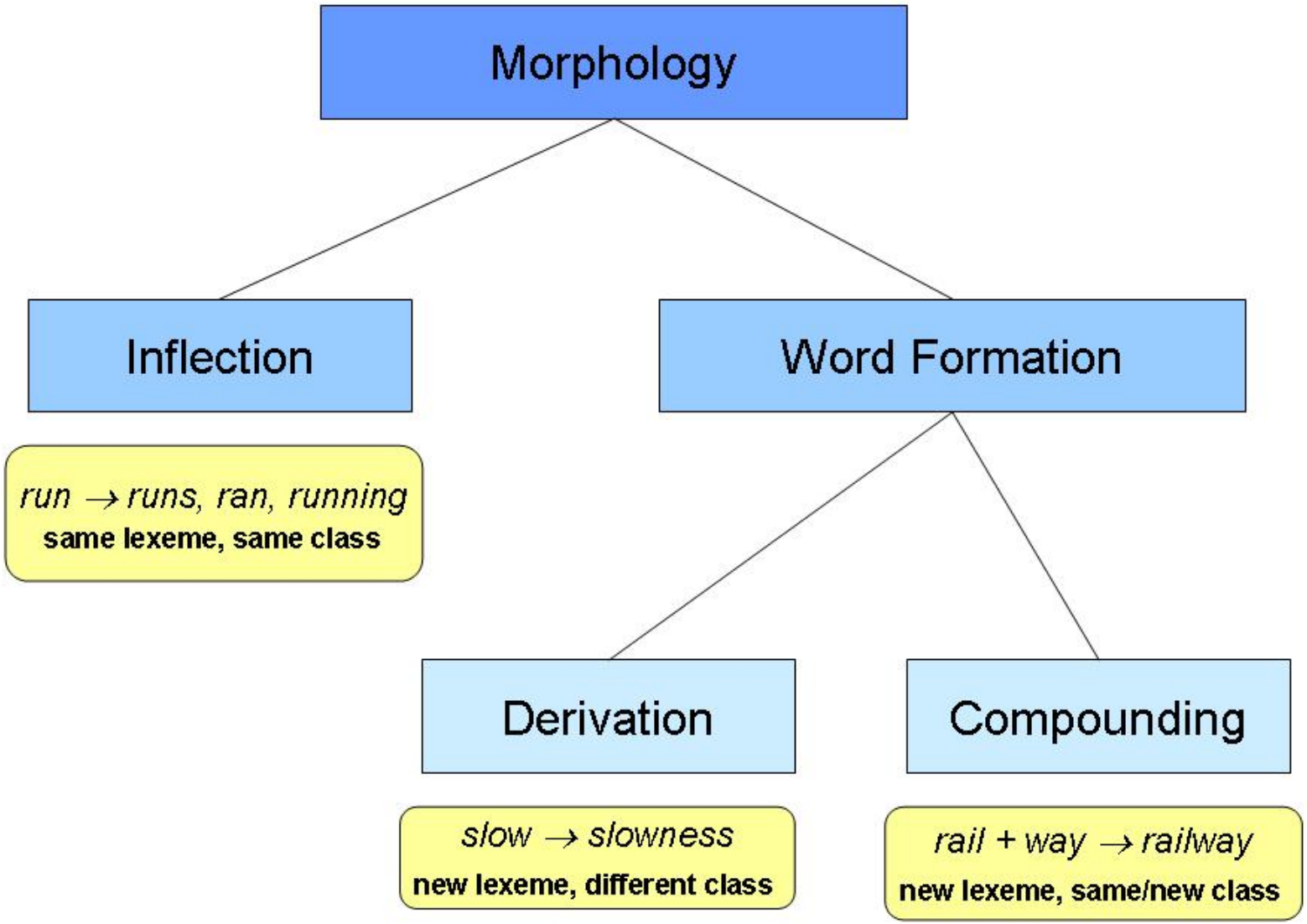
- **Lexeme** = abstract ("deep") unit having a certain **meaning**, and belonging to a certain **class**; **lexicon** = set of lexemes
- **Word form** = different textual ("surface") realizations of a lexeme

The **inflection paradigm** = all word forms of a lexeme.

A **lemma** = a canonical word form chosen to represent the lexeme.

	Word forms	Lemma
French	{porte, portes}	porte
	{porter, porte, portes, portiez, . . . }	porter
	{à}	à
Your language		

Types of morphological rules



Inflectional categories and values (1/4)

	English (Germanic)	French (Latin)	Serbian (Slavic)	Your language (...)
Number (*Nb*)	singular (*s*) plural (*p*)	singular (*s*) plural (*p*)	singular (*s*) plural (*p*) paukal (*w*)	
Gender (*Gen*)		masculine (*m*) feminine (*f*)	masculine (*m*) feminine (*f*) neuter (*n*)	
Case			nominative (*1*) genitive (*2*) dative (*3*) accusative (*4*) instrumental (*5*) locative (*6*) vocative (*7*)	

Inflectional categories and values (2/4)

	English	French	Serbian	Your language
Degree (*Deg*)	positive (*<E>*) comparative (*C*) superlative (*S*)		positive (*a*) comparative (*b*) superlative (*c*)	
Person (*Pers*)	first (*1*) second (*2*) third (*3*)	first (*1*) second (*2*) third (*3*)	first (*x*) second (*y*) third (*z*)	
Animate-ness (*Anim*)			animate (*v*) inanimate (*q*) no-care (*g*)	

Inflectional categories and values (3/4)

	English	French
Tense and mood (*TM*)	infinitive (*W*): *do* present indicative (*P*): *does* imperfect indicative (*I*): *did* past participle (*K*): *done* gerund (*G*): *doing*	infinitive (*W*): *faire* present indicative (*P*): *faisons* imperfect indicative (*I*): *faisait* present subjunctive (*S*): *fasse* imperfect subjunctive (*T*): *fisse* present imperative (*Y*): *faites* present conditional (*C*): *ferait* simple past (*J*): *fit* past participle (*K*): *faite* gerund (*G*): *faisant* future (*F*): *fera*

Inflectional categories and values (4/4)

	Your language
Tense and mood (*TM*)	

Inflectional classes ≈ parts of speech (POS) (1/6)

	Noun	Max. forms
English	inflects in number: *dog*, *dogs*	2
French	inflects in number *toile*, *toiles* has gender *toile* OR inflects in gender *cousin*, *cousine*	4
Serbian	inflects in number has gender OR inflects in gender inflects in case has animateness	28
Your language		

Inflectional classes ≈ parts of speech (POS) (2/6)

	Adjective		Max forms
English	uninflected OR inflects in	*famous* *big, bigger*	3
French	inflects in inflects in	*bleu, bleus* *bleue, bleues*	4
Serbian	inflects in number has gender OR inflects in gender inflects in case inflects in animateness inflects in degree inflects in determinedness		77
Your language			

Inflectional classes ≈ parts of speech (POS) (3/6)

	Verb		Max. forms
English	inflects in	*go, went, going*	9
	inflects in	*go, goes*	
	inflects in	*am, are*	
French	inflects in	*être, suis, été*	51
	inflects in	*suis, es, est*	
	inflects in	*suis, sommes*	
	inflects in	*aimés, aimées*	
Serbian	inflects in tense-mood		dozens
	inflects in person		
	inflects in number		
	inflects in gender		
Your language			

Inflectional classes ≈ parts of speech (POS) (4/6)

	Pronoun		Max forms
English	inflects in inflects in inflects in	*I, you, he* *I, we* *he, she*	8
French	inflects in inflects in inflects in	*je, tu, il* *tu, vous* *il, elle*	8
Serbian	inflects in person inflects in number inflects in gender . . .		10
Your language			

Inflectional classes ≈ parts of speech (POS) (5/6)

	Adverb	Max. forms
English	uninflected *yesterday* OR inflects in *early, earlier*	3
French	uninflected *hier, facilement*	1
Serbian	uninflected	1
Your language		

Inflectional classes ≈ parts of speech (POS) (6/7)

	Determiner		Max. forms
English	has	*a, this, those, the*	1
French	inflects in	*le, les*	4
	inflects in	*le, la*	
Serbian	inexistent		0
Your language			

Inflectional classes ≈ parts of speech (POS) (6/6)

	Preposition	**Conjunction**	**Interjection**
English	uninflected: *to*	uninflected: *and*	uninflected: *hurray*
French	uninflected: *de*	uninflected: *mais*	uninflected: *adieu*
Serbian	uninflected	uninflected	uninflected
Your language			

Inflectional paradigm (verb lemma *aimer*)

Word form	Features	Word form	Features	Word form	Features
aimer	W	aimais	I2s	aimais	I1s
aimait	I3s	aimions	I1p	aimiez	I2p
aimaient	I3p	aimassent	T3p	aimassiez	T2p
aimassions	T1p	aimât	T3s	aimasses	T2s
aimasse	T1s	aimai	J1s	aima	J3s
aimâmes	J1p	aimâtes	J2p	aimèrent	J3p
aimas	J2s	aimant	G	aimés	Kmp
aimé	Kms	aimées	Kfp	aimée	Kfs
aimons	Y1p	aimons	P1p	aimions	S1p
aimiez	S2p	aimerais	C2s	aimerais	C1s
aimerait	C3s	aimerions	C1p	aimeriez	C2p
aimeraient	C3p	aimerai	F1s	aimeras	F2s
aimera	F3s	aimerons	F1p	aimerez	F2p
aimeront	F3p	aime	Y2s	aime	S3s
aime	S1s	aime	P3s	aime	P1s
aiment	S3p	aiment	P3p	aimes	S2s
aimes	P2s	aimez	Y2p	aimez	P2p

Derivational morphology (1/2)

- Source word:
  - a lemma: *small* → *smallness*
  - an inflected form (in French): *normale* → *normalement*
- Derivational affix:
  - prefix: *ir* + *regular*(adj.) → *irregular*
  - infix: e.g. in Arabic
  - suffix: *small*+*ness* → *smallness*
  - no affix: *to enter* → *an enter*
- Target word: different lexeme and/or different class (inflects differently)
  - *small*(adj.) → *smallness*(noun)
  - *astonish*(verb) → *astonishment*(noun)
  - *count*(verb) → *countable*(adj.)
  - *courage*(noun) → *encourage*(verb)
  - *forest*(noun) → *forestry*(noun)
- Stem modification:
  - *regulate* → *regulation*
- Multiple affixes:
  - *un* + *forget* + *able* → *unforgettable*

Compounding

- Several lexemes form a **new lexeme**.
- The new lexeme shows some degree of **non-compositionality**
  - morphological: *un peau rouge*(masc.), unlike *peau*(fem.)
  - syntactic: *un moulin à vent*, but not **un moulin à brise*
  - distributional: *un cordon bleu*(human), unlike *cordon*(inanimate)
  - semantic: *pomme de terre* is not an apple from earth

Headword

- **Headword**: component from which the compound inherits its features

  fire<u>man</u> - noun in singular like *man*
  <u>cheval</u> à bascules - noun in singular masculine, like *cheval*

- Types of compounds:
  - **endocentric** (has a headword): *fire<u>man</u>*
  - **exocentric** (no headword): (EN) *forget-me-not*, (FR) *porte-serviettes*
  - **apposition** (two heads): *man servant* $\longrightarrow$ *men-servants*

Examples of compounds

	Noun	**Adjective**	**Verb**
English	*air brake* *forget-me-not* *man-of-war*	*bittersweet* *easy-going* *as busy as a bee*	*cut off* *co-occur* *make up for*
French	*rouge-gorge* *stylo à bille* *porte-monnaie*	*à pied* *anglo-saxon* *sans domicile fixe*	*sous-entendre* *faire avec* *contre-attaquer*
Your language			

Examples of compounds

	Adverb	**Preposition**	**Conjunction**
English	*all of a sudden* *as soon as possible* *on and on*	*instead of* *contrary to* *in front of*	*as well as* *if and only if* *neither . . . nor*
French	*trop bien* *un peu* *à l'envers*	*à propos de* *de façon à* *en cas de*	*alors que* *parce que* *au moment où*
Your language			

Ambiguity of compounds

- Non-ambiguous compound: each occurrence of its components is always a compound.

  Je suis venu parce que je le voulais.

- Ambiguous compound: an occurrence of its components may or may not be a compound.

  Je suis venu alors que je ne le voulais pas.
  Il m'a dit alors que l'affaire était close.

Natural language $\neq$ formal language

- Linguistic definitions are circular (. . . . .)
- Basic elements are not clearly defined (. . . . .)
- Many notions are based on human intuition, and remain formally undescribed

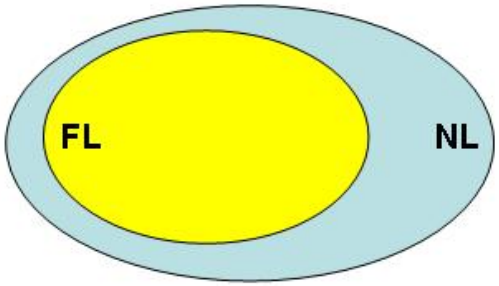
But:

- Computer programs cannot deal with implicit knowledge
- They can only treat formal languages

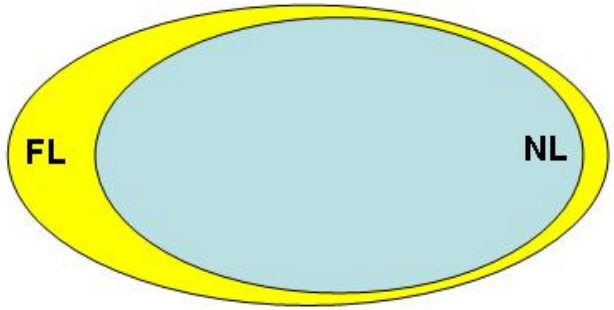
Solution:

- Define a formal language as close as possible to the natural language

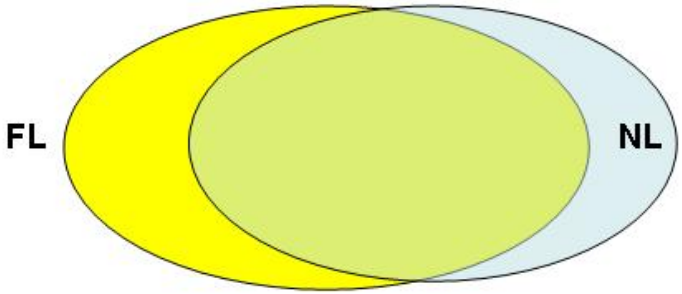
Natural language vs. formal language



Overlooking of exceptions



Overgeneralization



Both overgeneralization and overlooking of exceptions

What is a word ? What is an alphabet ?

In a **formal language**:

- **alphabet** Σ is a finite set of symbols
- a **word** over Σ is a (finite or infinite) sequence of elements in Σ: $\omega \in \Sigma^*$
- a **language** is a (finite or infinite) subset of Σ^* given by a **grammar**

Example:

- $\Sigma = \{a, b\}$
- $L = \{a, aba, aabaa, aaabaaa, \ldots\}$
- Grammar $= \ldots$

What is a word ? What is an alphabet ?

In a **natural language** - on the **morphological level**

- an **alphabet** = list of (lowercase, uppercase, accented,. . .) letters of the language

  In English: {A, a, B, b, C, c, . . . }
  In French: {A, a, Â, â, À, à, B, b, C, c, . . . }
  In your language:

- a **language** = list of all correct (*grammatical*) words of the language

  In English: {a, the, dog, dogs, make, making, example, . . . }
  In French: {un, à, cher, chères, exemple . . . }
  In your language:

- a **grammar**
  - A set of correct word forms
  - Grammar rules (over)generating sets of words

    In English: NOUN $\rightarrow$ ADJ *ness*
    *madness, emptiness, *irregularness, . . .*

What is a word ? What is an alphabet ?

In a **natural language** - on the **syntactic level**

- an **alphabet** = list of valid *morphological words*

  In English: {a, the, dog, dogs, make, making, example, . . . }
  In French: {un, à, cher, chères, exemple . . . }
  In your language:

- a **language** = list of all correct (*grammatical*) sentences of the language:

  In English: {Dogs like cats., Do cats like dogs?, We will see example 5., . . . }
  In French: {Ces maisons sont-elles chères?, Tais-toi!, . . . }
  In your language:

- a **grammar**

  Many formalisms were proposed (DCG, TAG, . . . , see lecture on syntax)
  An **complete** and **efficient** grammar remains a challenge

Non-alphabet characters

- They help to separate morphological words in a sentence
- They separate sentences
- They may be parts of words (*aujourd'hui*)
- They may miss between words (*Schul|errinnerung*)
- They may have a semantic content : λ-*calculus*, γ-*rays*

The English paradox

- It is the dominating language in the Natural Language Processing (NLP) community
- It is one of the least inflected occidental languages

Computational morphology

- **Tokenization** = dividing text into elementary graphical units (word forms, separators, . . .)
- **Morphological analysis** = assigning all possible morphological interpretations to a word form (out of context)
- **Morphological disambiguation (tagging)** = choosing the correct interpretation in the given context
- **Morphological generation** = for a given lemma and annotation, produce the corresponding word form(s)

Morphological analysis and generation

Morphological analysis: from "surface" form to an (several) annotation(s)

avions $\rightarrow \{\langle$ Lemma=**avion**, Class=**N**, Nb=**p**$\rangle$,
$\langle$ Lemma=**avoir**, Class=**V**, Nb=**p**, TM=**I**, Pers=**1**$\rangle\}$

Morphological generation: from and annotation to a (several) surface forms

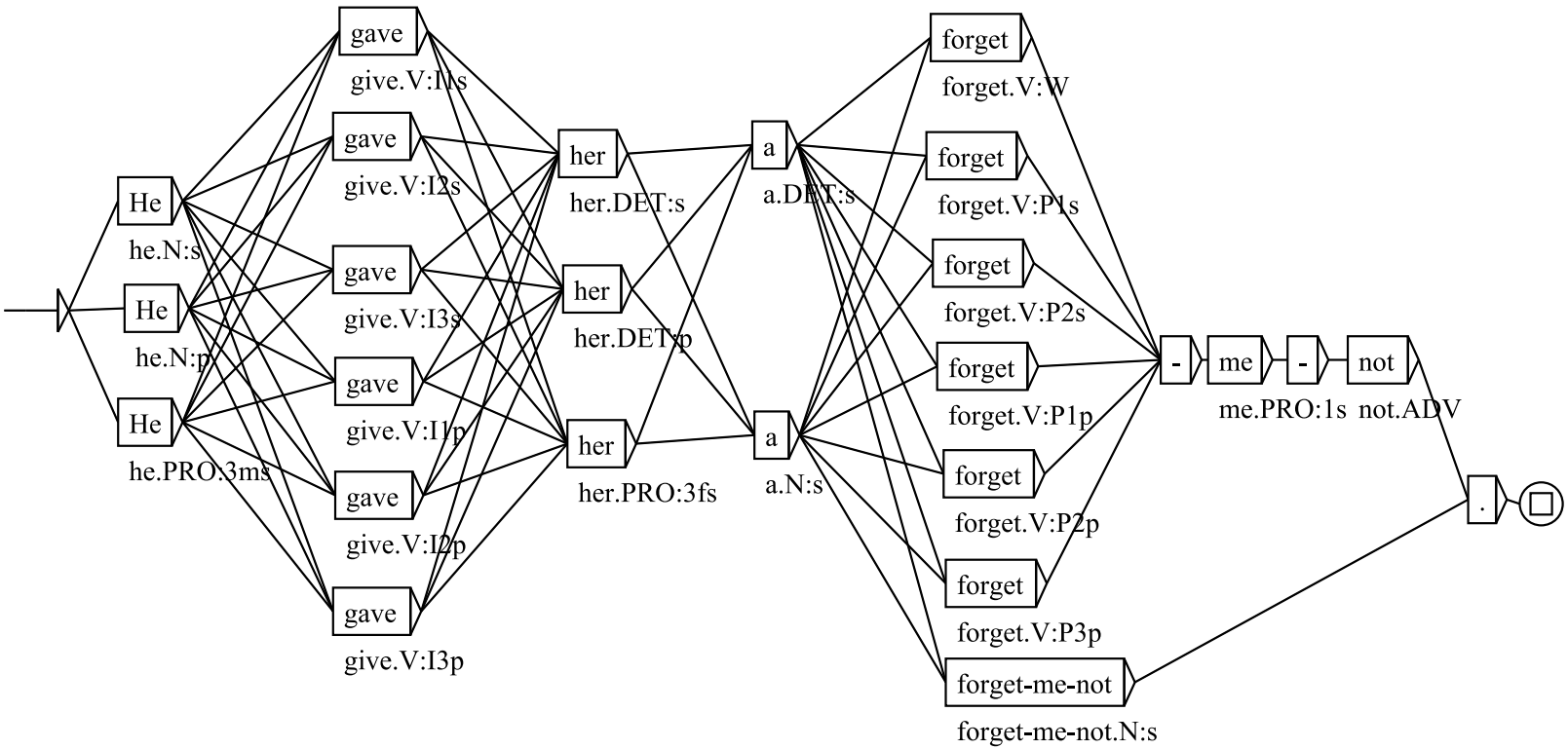
$\langle$ Lemma=**avoir**, Class=**V**, Nb=**p**, TM=**I**, Pers=**1**$\rangle \rightarrow$ **avions**

Tokenization and morphological analysis of a sentence

He gave her a forget-me-not.

He	gave	her	a	forget	-	me	-	not
he.N:s	give.V:I1s	her.DET:s	a.DET:s	forget.V:P1s	-	me.PRO:1s	-	not,.A
he.N:p	give.V:I2s	her.DET:p		forget.V:P2s	-		-	
he.PRO:3ms	give.V:I3s	her.PRO:3fs		forget.V:P1p	-		-	
	give.V:I1p			forget.V:P2p	-		-	
	give.V:I2p			forget.V:P3p	-		-	
	give.V:I3p			forget.V:W	-		-	
				forget-me-not.N:s				

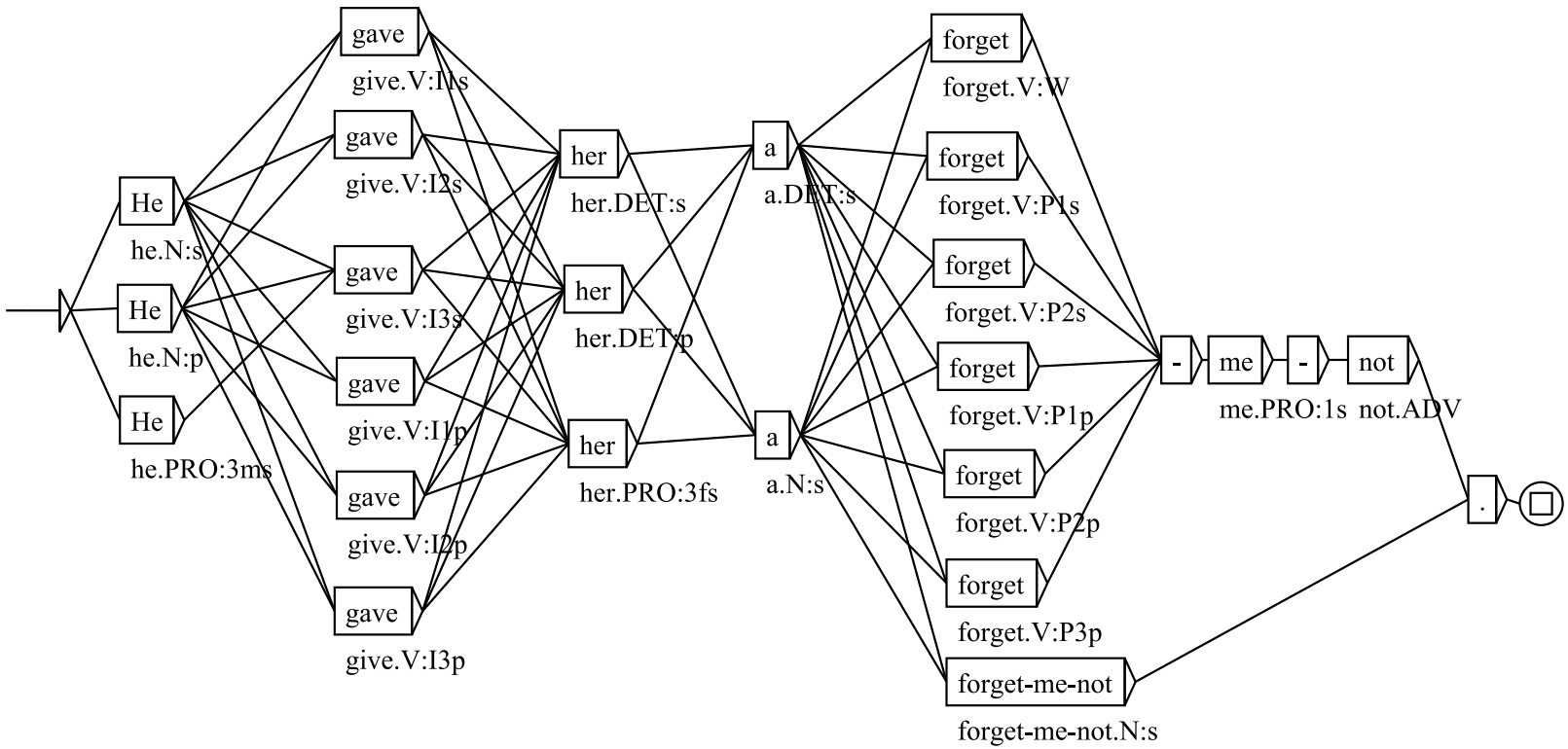
A sentence becomes a graph



How many possible interpretations of the sentence ?

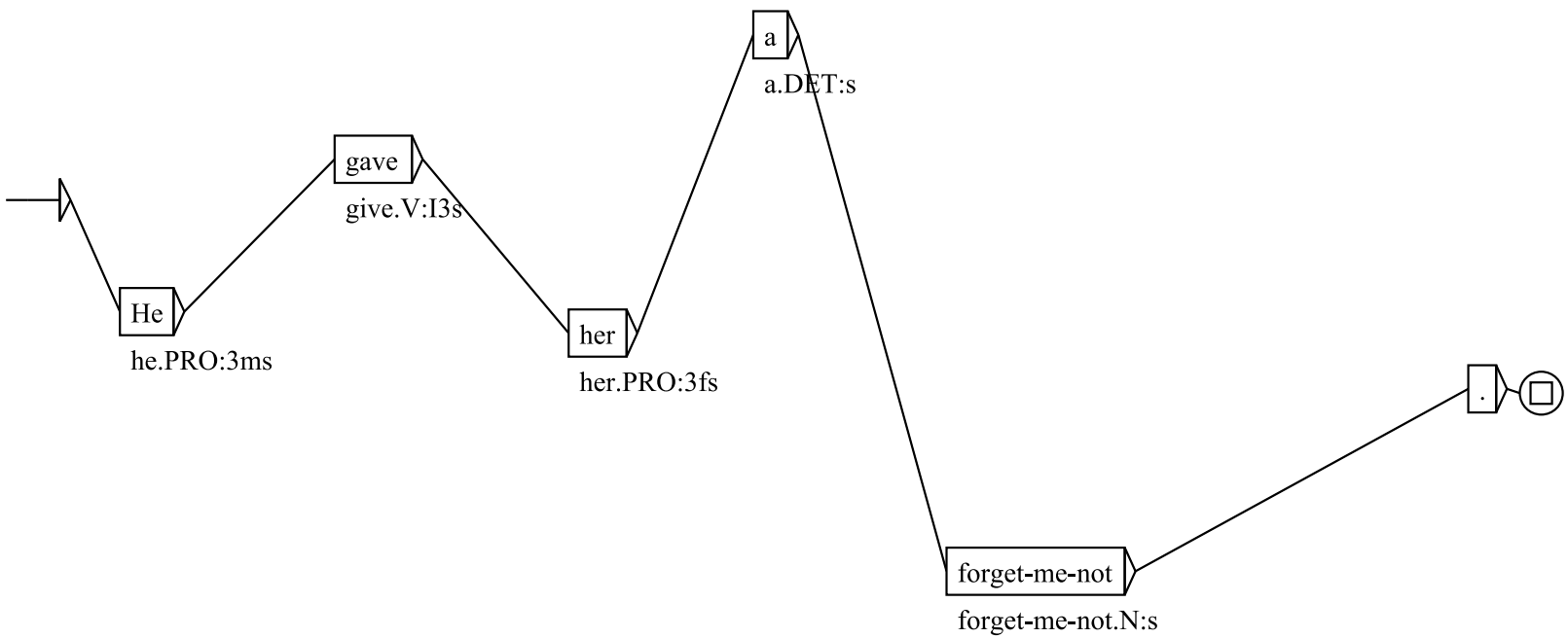
Disambiguation: cutting off forbidden paths

Disambiguating rule - example: *If a personal pronoun is followed by a verb, both must agree in number and person.*



How many possible interpretations of the sentence were eliminated ?

Tagging = choosing the correct interpretation of the sentence



A perfect tagging is not always possible

Truly ambiguous sentences exist:

La petite brise la glace.