Propriétés principales de l'analogie en relation avec les chaînes de la langue

Yves Lepage

17ème Séminaire sur le Raisonnement à partir de cas,
29 juin 2009

Deux questions

Pourquoi sait-on que l'accusatif de *puella* est *puellam* dès que l'on sait que celui de *rosa* est *rosam* ?

Peut-on tirer de la réponse à la question précédente des langages formels qui aient de bonnes propriétés pour la description des langues ?

Préoccupation initiale

Reproduction sur machine de certains mécanismes en langue ou de certains processus cognitifs, voire de l'acquisition des langues.

> It is hard to imagine our brain using auxiliary intermediate sentences of a nonterminal type. Instead, it looks more natural [...] to start with a collection of well-formed sentences, maybe acquired from experience, and to produce new well-formed ones by adding further words, in pairs that can observe dependencies and agreements [...]

[Marcus, Martín-Vide et Păun, 1998]

rosa	*puella*
rosa	?
rosam	?
rosae	?
rosae	?
rosā	?

rosa	*puella*
rosa	*puella*
rosam	*?*
rosae	*?*
rosae	*?*
rosā	*?*

rosa	*puella*
rosa	*puella*
rosam	*puellam*
rosae	?
rosae	?
rosā	?

rosa	*puella*
rosa	*puella*
rosam	*puellam*
rosae	*puellae*
rosae	?
rosā	?

rosa	*puella*
rosa	*puella*
rosam	*puellam*
rosae	*puellae*
rosae	*puellae*
rosā	*?*

rosa	*puella*
rosa	*puella*
rosam	*puellam*
rosae	*puellae*
rosae	*puellae*
rosā	*puellā*

rosa	*puella*
rosa	*puella*
rosam	?
rosae	?
rosae	?
rosā	?

rosa : *rosam* :: *puella* : ?

rosa	*puella*
rosa	*puella*
rosam	*?*
rosae	*?*
rosae	*?*
rosā	*?*

rosa : *rosam* :: *puella* : *puellam*

rosa	*puella*	
rosa	*puella*	
rosam	*puellam*	*rosa* : *rosam* :: *puella* : *puellam*
rosae	?	
rosae	?	
rosā	?	

rosa	*puella*		
rosa	*puella*		
rosam	*puellam*		*rosa* : *rosae* :: *puella* : ?
rosae	?		
rosae	?		
rosā	?		

rosa	*puella*	
rosa	*puella*	
rosam	*puellam*	*rosa* : *rosae* :: *puella* : *puellae*
rosae	?	
rosae	?	
rosā	?	

rosa	*puella*		
rosa	*puella*		
rosam	*puellam*		*rosa* : *rosae* :: *puella* : *puellae*
rosae	*puellae*		
rosae	*puellae*		
rosā	?		

rosa	*puella*		
rosa	*puella*		
rosam	*puellam*		*rosa* : *rosā* :: *puella* : *puellā*
rosae	*puellae*		
rosae	*puellae*		
rosā	*puellā*		

rosa	*puella*
rosa	*?*
rosam	*?*
rosae	*?*
rosae	*?*
rosā	*?*

rosa : *rosa* :: *puella* : *puella*

rosa	*puella*
rosa	*puella*
rosam	?
rosae	?
rosae	?
rosā	?

rosa : *rosa* :: *puella* : *puella*

Le terme technique d'analogie désigne une relation entre quatre objets A, B, C et D qui énonce que A est à B ce que C est à D. On la note : $A : B :: C : D$.

Mathématique et philosophie grecques : Eudoxe, Euclide, Aristote

Grammaire grecque et latine : Apollonius Dyscole, Quintilien, Jules César (si, si...), Varron

Linguistique comparatiste : Osthoff, Brugmann, Paul

Structuralismes européen et américain : Saussure, Bloomfield

Analogies comme relations

$$\frac{2}{3} = \frac{4}{6}$$

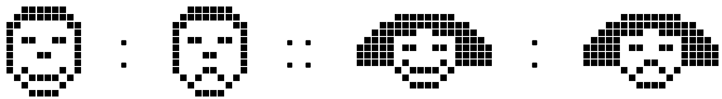
$$rosa : rosam :: puella : puellam$$

Équations analogiques

$$\frac{2}{3} = \frac{4}{x} \quad \Rightarrow \quad x = 6$$

$$\textit{rosa} : \textit{rosam} :: \textit{puella} : x \qquad \Rightarrow \quad x = \textit{puellam}$$

Exemples d'analogies



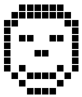
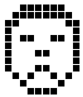
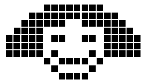
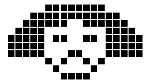
嫁 : 妙 :: 稼 : 秒

prendrai : *prendre* :: *viendrai* : *viendre*

aslama : *muslimun* :: *arsala* : *mursilun*

meja : *meja-meja* :: *bangsa* : *bangsa-bangsa*

Exemples d'analogies

 :  ::  : x $\Rightarrow$ $x =$ 

嫁 : 妙 :: 稼 : x $\Rightarrow$ $x =$ 秒

prendrai : *prendre* :: *viendrai* : x $\Rightarrow$ $x =$ *viendre*

aslama : *muslimun* :: *arsala* : x $\Rightarrow$ $x =$ *mursilun*

meja : *meja-meja* :: *bangsa* : x $\Rightarrow$ $x =$ *bangsa-bangsa*

Analogies en syntaxe et en corpus

I'd like a non-smoking table.
Where is the conference center ?
Turn left at the second corner.
I'd prefer a non-smoking table.
I prefer French food.
I like French food.

Analogies en syntaxe et en corpus

I'd like a non-smoking table.
Where is the conference center?
Turn left at the second corner.
I'd prefer a non-smoking table.
I prefer French food.
I like French food.

Analogies en syntaxe et en corpus

I'd like a non-smoking table.
I like French food.
I'd prefer a non-smoking table.
I prefer French food.

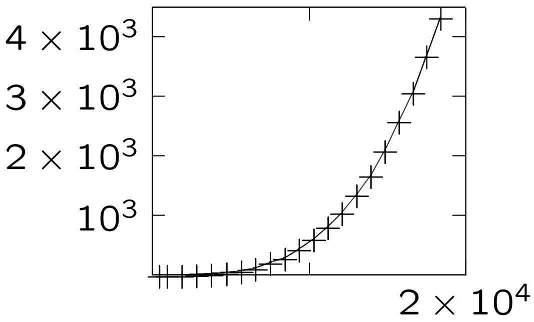
Analogies en syntaxe et en corpus

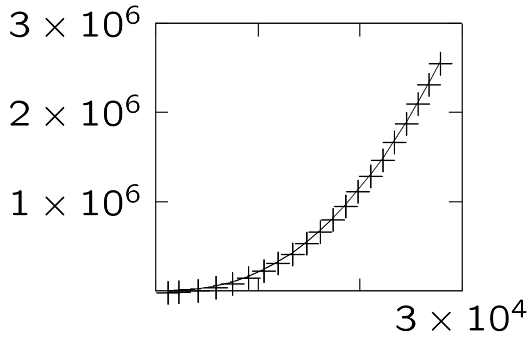
I'd like a non-smoking table. : *I like French food.* :: *I'd prefer a non-smoking table.* : *I prefer French food.*

Analogies en syntaxe et en corpus

$$\textit{I'd like a non-smoking table.} : \textit{I like French food.} :: \textit{I'd prefer a non-smoking table.} : x \quad \Rightarrow \quad x = \textit{I prefer French food.}$$

Estimation du nombre d'analogies





À gauche, nombre d'analogies entre phrases d'un même corpus, à droite entre *chunks* extraits de ces phrases.

Estimation du nombre de vraies analogies

Dans un corpus d'environ 100 000 phrases courtes en trois langues (anglais, japonais, chinois) avec correspondances de traduction.

	vraies analogies		seuil de
	nombre	pourcentage	rejet
intersection	68 164	99,04%	0,1%
inflation + intersection	1 507 380	97,05%	0,1%
échantillon (666 analogies)	2 384 202	96,18%	0,1%

Propriétés générales

Les huit formes équivalentes

fable : *fabuleux* :: *miracle* : *miraculeux*
fable : *miracle* :: *fabuleux* : *miraculeux*
fabuleux : *fable* :: *miraculeux* : *miracle*
fabuleux : *miraculeux* :: *fable* : *miracle*
miracle : *fable* :: *miraculeux* : *fabuleux*
miracle : *miraculeux* :: *fable* : *fabuleux*
miraculeux : *fabuleux* :: *miracle* : *fable*
miraculeux : *miracle* :: *fabuleux* : *fable*

Les huit formes équivalentes

– échange des moyens

$$A : B :: C : D \quad \Leftrightarrow \quad A : C :: B : D$$

– échange des extrêmes

$$A : B :: C : D \quad \Leftrightarrow \quad D : B :: C : A$$

– symétrie de la conformité

$$A : B :: C : D \quad \Leftrightarrow \quad C : D :: A : B$$

– inversion des rapports

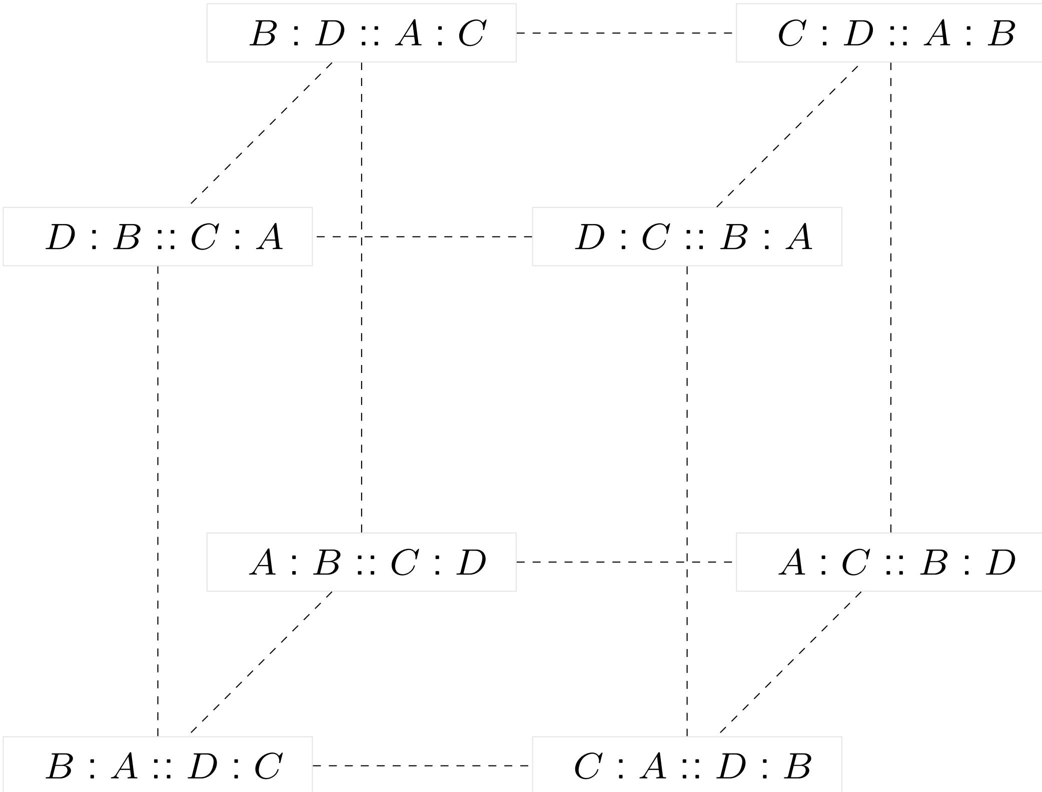
$$A : B :: C : D \quad \Leftrightarrow \quad B : A :: D : C$$

Les huit formes équivalentes

Les huit analogies suivantes sont équivalentes :

$A : B :: C : D$	
$A : C :: B : D$	(échange des moyens)
$B : A :: D : C$	(inversion des rapports)
$B : D :: A : C$	
$C : A :: D : B$	
$C : D :: A : B$	(symétrie de la conformité)
$D : B :: C : A$	(échange des extrêmes)
$D : C :: B : A$	(symétrie de lecture)

Le cube des huit analogies équivalentes



Les huit formes équivalentes

Soient quatre objets A, B, C et D, il existe au plus trois analogies entre ces quatre objets : $A : B :: C : D$, $A : B :: D : C$ et $A : D :: C : B$

Spécialisation du problème

Forme seulement

On s'intéresse maintenant aux chaînes de symboles.

On ne considère pas les analogies du type suivant car elles font appel à la connaissance de la langue ou au sens.

oiseau : *ailes* :: *poisson* : *nageoires*

prendrai : *prendre* :: *viendrai* : *venir*

Forme seulement

On s'intéresse maintenant aux chaînes de symboles.

On ne considère pas les analogies du type suivant car elles font appel à la connaissance de la langue ou au sens.

oiseau : *ailes* :: *poisson* : *nageoires*

prendrai : *prendre* :: *viendrai* : *venir*

On ne considérera donc que des analogies de forme.

Commutation seulement

Reduplication :	*guru* : *guru-guru* :: *pelajar* : *pelajar-pelajar*
	pago : *pepigi* :: *cado* : *cecidi*
Commutation :	*aslama* : *muslim* :: *arsala* : *mursil*
Miroir :	*abcd* : *dcba* :: *xyz* : *zyx*

Commutation seulement

Reduplication :	*guru* : *guru-guru* :: *pelajar* : *pelajar-pelajar*
Miroir :	*abcd* : *dcba* :: *xyz* : *zyx*
Commutation :	*aslama* : *muslim* :: *arsala* : *mursil*
Reduplication :	*pago* : *pepigi* :: *cado* : *cecidi*

Commutation seulement

Reduplication :	$A : AA :: B : BB$
Miroir :	$A : \bar{A} :: B : \bar{B}$
Commutation :	*aslama* : *muslim* :: *arsala* : *mursil*
Reduplication :	*pago* : *pepigi* :: *cado* : *cecidi*

Commutation seulement

Reduplication :	$A : AA :: B : x \quad \Rightarrow \quad x = BB$
Miroir :	$A : \bar{A} :: B : x \quad \Rightarrow \quad x = \bar{B}$
Commutation :	*aslama* : *muslim* :: *arsala* : *mursil*
Reduplication :	*pago* : *pepigi* :: *cado* : *cecidi*

Commutation seulement

Reduplication :	*guru* : *guru-guru* :: *pelajar* : *pelajar-pelajar*
Miroir :	*abcd* : *dcba* :: *xyz* : *zyx*
Commutation :	*aslama* : *muslim* :: *arsala* : *mursil*
Reduplication :	*pago* : *pepigi* :: *cado* : *cecidi*

Commutation seulement

Reduplication :	*guru* : *guru-guru* :: *pelajar* : *pelajar-pelajar*
Miroir :	*abcd* : *dcba* :: *xyz* : *zyx*
Commutation :	*aslama* : *muslim* :: *arsala* : *mursil*
Reduplication :	*pago* : *pepigi* :: *cado* : *cecidi*

On ne considérera que les analogies de commutation.

Difficulté

oratorem : *orator* :: *honorem* : *honor*

chodzić : *przechodzić* :: *rzucić* : *przerzucić*

SLM : *muSLiMun* :: *RSL* : *muRSiLun*

Difficulté

oratorem : *orator* :: *honorem* : *honor*

chodzić : *przechodzić* :: *rzucić* : *przerzucić*

SLM : *muSLiMun* :: *RSL* : *muRSiLun*

Difficulté

oratorem : *oratoris* :: *honorem* : *honoris*

wychodzić : *przechodzić* :: *wyrzucić* : *przerzucić*

aslama : *muslimun* :: *arsala* : *mursilun*

Difficulté

je marche : *tu marches* :: *je décompose* : *tu décomposes*

prendre un obstacle : *prendre un verre* :: *boire un obstacle* : *boire un verre*

une petite table : *une grande table* :: *une petite décision* : *une grande décision*

Difficulté

Est-ce que ces fenêtres-là, je peux les ouvrir ? : *Est-ce que vous pouvez m'ouvrir une fenêtre ?* :: *Ces chèques de voyage-là, je peux les échanger ?* : *Vous pouvez m'échanger un chèque de voyage ?*

La formalisation d'Yvon et Stroppa

$$A : B :: C : D \quad \Leftrightarrow \quad A \bullet D \cap B \bullet C \neq \phi$$

avec $\bullet$ pour le mélange (*shuffle*) de deux chaînes :

$$abc \bullet de = \{abcde, abdec, adebc, \ldots, adbec, dabec, \ldots\}$$

Déterminisme

L'analogie $C : C :: D_1 : D_2$, avec $D_1 \neq D_2$, est-elle possible ?

On appellera hypothèse du déterminisme le refus de la possibilité d'une telle analogie.

Non transitivité du signe ::

Il existe des équations analogiques ayant plusieurs solutions. Par exemple :

$$a : aa :: b : x \quad \Rightarrow \quad x = ba \text{ ou } ab$$

Non transitivité du signe ::

Il existe des équations analogiques ayant plusieurs solutions. Par exemple :

$$a : aa :: b : ba$$

$$a : aa :: b : ab$$

Non transitivité du signe ::

S'il y avait transitivité du signe ::, alors on aurait contradiction avec l'hypothèse de déterminisme. Considérons en effet une équation analogique avec deux solutions D_1 et D_2 différentes :

$$\left\{ \begin{array}{l} A : B :: C : D_1 \\ A : B :: C : D_2 \end{array} \right. \Rightarrow \left\{ \begin{array}{l} C : D_1 :: A : B \\ A : B :: C : D_2 \end{array} \right.$$

$$\Rightarrow \quad C : D_1 :: C : D_2$$

$$\Rightarrow \quad C : C :: D_1 : D_2$$

Non transitivité du signe ::

La formalisation par le mélange n'est pas compatible avec l'hypothèse du déterminisme et autorise les analogies $A : A :: C_1 : C_2$ avec $C_1 \neq C_2$.

Exemple : $aba : aba :: ab : ba$

$$ab \neq ba \text{ et } ababa \in (aba \bullet ba) \cap (aba \bullet ab)$$

Une autre formalisation

Première observation

abc : *def* :: *ghi* : x

abc : *abd* :: *ghi* : x

abc : *abd* :: *ghc* : x

prendre : *prenable* :: *alunir* : x

prendre : *prenable* :: *vendre* : x

Première observation

$abc : def :: ghi : x$ insoluble !

$abc : abd :: ghi : x$ insoluble !

$abc : abd :: ghc : x \quad \Rightarrow \quad x = ghd$

$prendre : prenable :: alunir : x$

$prendre : prenable :: vendre : x$

Première observation

abc : *def* :: *ghi* : x insoluble !

abc : *abd* :: *ghi* : x insoluble !

abc : *abd* :: *ghc* : x $\Rightarrow$ $x =$ *ghd*

prendre : *prenable* :: *alunir* : x insoluble !

prendre : *prenable* :: *vendre* : x $\Rightarrow$ $x =$ *venable*

Première observation

abc : *def* :: *ghi* : x insoluble !

abc : *abd* :: *ghi* : x insoluble !

abc : *abd* :: *ghc* : x $\Rightarrow$ $x =$ *ghd*

prendre : *prenable* :: *alunir* : x insoluble !

prendre : *prenable* :: *vendre* : x $\Rightarrow$ $x =$ *venable*

Formulation

Une condition suffisante pour que l'analogie $A : B :: C : D$ ne soit pas vérifiée est que des symboles de A n'apparaissent ni dans B ni dans C.

Contraposée :

Une condition nécessaire pour que l'analogie $A : B :: C : D$ soit vérifiée est que tous les symboles de A apparaissent soit dans B, soit dans C, soit dans les deux à la fois (et dans le même ordre).

Vérification

oratorem : *orator* :: *honorem* : *honor*

huzila : *huzāl* :: *ṣudi'a* : *ṣudā'*

setzen : *setzte* :: *lachen* : *lachte*

inné : *nées* :: *indu* : *dues*

biorąc : *bierzesz* :: *piorąc* : *pierzesz*

tinggal : *ketinggalan* :: *duduk* : *kedudukan*

Vérification

$aa : ab :: ba : bb$

$ab : aabb :: aaaaaaabbbbbbb : aaaaaaaabbbbbbbb$

$abc : aabbcc :: aaabbbccc : aaaabbbbcccc$

$aab : aaaabb :: aaaaaabbb : aaaaaaaabbbb$

$aaaabbbbcccc : aaabbbccc :: aabbcc : abc$

Inclusion des ensembles de symboles

Soit $\mathcal{V}$ un alphabet, $\forall (A, B, C, D) \in (\mathcal{V}^*)^4$,

$$A : B :: C : D \quad \Rightarrow \quad \overline{A} \subset \overline{B} \cup \overline{C}$$

Soit $\mathcal{V}$ un alphabet, $\forall (A, B, C, D) \in (\mathcal{V}^*)^4$,

$$A : B :: C : D \quad \Rightarrow \quad \overline{A} \cup \overline{D} = \overline{B} \cup \overline{C}$$

Similitude

La similitude de deux chaînes est la longueur de leur plus longue sous-séquence commune.

Exemple :

$$s(\textit{angélique}, \textit{angulaire}) = |\textit{angle}| = 5$$

$$s(\textit{angélique}, \textit{angulaire}) = |\textit{angie}| = 5$$

$$s(\textit{angélique}, \textit{angulaire}) = |\textit{angue}| = 5$$

Similitude

Soit $\mathcal{V}$ un alphabet, soit s la similitude entre deux chaînes.
$\forall (A, B, C, D) \in (\mathcal{V}^*)^4$,

$$A : B :: C : D \quad \Rightarrow$$

$$|A| \;\leq\; s(A,B) + s(A,C)$$

Similitude

Soit $\mathcal{V}$ un alphabet, soit s la similitude entre deux chaînes.
$\forall (A, B, C, D) \in (\mathcal{V}^*)^4$,

$$A : B :: C : D \quad \Rightarrow$$

$$|A| = s(A, B) + s(A, C) - \gamma(A, B, C)$$

Hypothèse

Par échange des moyens, on a :

$$\gamma(A, B, C) = \gamma(A, C, B)$$

Si l'analogie existe, alors on pose que :

$$\gamma(A, B, C) = \gamma(B, A, D) = \gamma(C, A, D) = \gamma(D, B, C)$$

De la similitude à la distance

$$\forall (A, B, C, D) \in (\mathcal{V}^*)^4, \quad A : B :: C : D \quad \Rightarrow$$

$$\left\{ \begin{array}{l} |A| = s(A, B) + s(A, C) - \gamma(A, B, C) \\ |B| = s(B, A) + s(B, D) - \gamma(B, A, D) \\ |C| = s(C, A) + s(C, D) - \gamma(C, A, D) \\ |D| = s(D, B) + s(D, C) - \gamma(D, B, C) \end{array} \right.$$

De la similitude à la distance

$$\forall (A,B,C,D) \in (\mathcal{V}^*)^4, \quad A : B :: C : D \quad \Rightarrow$$

$$\left\{ \begin{array}{ll} |A| - s(A,B) = |C| - s(C,D) & (1) \\ |B| - s(B,D) = |A| - s(A,C) & (2) \\ |C| - s(C,A) = |D| - s(D,B) & (3) \\ |D| - s(D,C) = |B| - s(B,A) & (4) \end{array} \right.$$

De la similitude à la distance

$$\forall (A,B,C,D) \in (\mathcal{V}^*)^4, \quad A : B :: C : D \quad \Rightarrow$$

$$\begin{cases} |A| + |B| - 2 \times s(A,B) = |C| + |D| - 2 \times s(C,D) \\ |A| + |C| - 2 \times s(A,C) = |B| + |D| - 2 \times s(B,D) \end{cases}$$

Or,

$$\forall (A,B) \in (\mathcal{V}^*)^2, \;\; d(A,B) = |A| + |B| - 2 \times s(A,B)$$

avec d la distance d'édition canonique (insertion et suppression seulement).

Egalité des distances

Soit $\mathcal{V}$ un alphabet, $\forall(A, B, C, D) \in (\mathcal{V}^*)^4$,

$$A : B :: C : D \quad \Rightarrow \begin{cases} d(A,B) = d(C,D) \\ d(A,C) = d(B,D) \end{cases}$$

Egalité des distances

Soit $\mathcal{V}$ un alphabet, $\forall (A, B, C, D) \in (\mathcal{V}^*)^4$,

$$A : B :: C : D \quad \Rightarrow \left\{ \begin{array}{l} d(A,B) = d(C,D) \\ d(A,C) = d(B,D) \end{array} \right.$$

Il existe des algorithmes rapides pour calculer la similarité entre deux chaînes et donc leur distance d'édition canonique : Hunt & Szymanski (1977), Allison & Dix (1986).

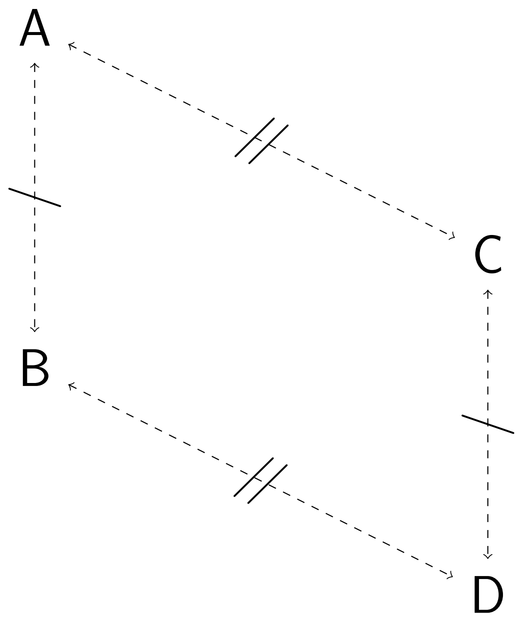
Egalité des distances

L'égalité des distances implique l'hypothèse du déterminisme :-) :

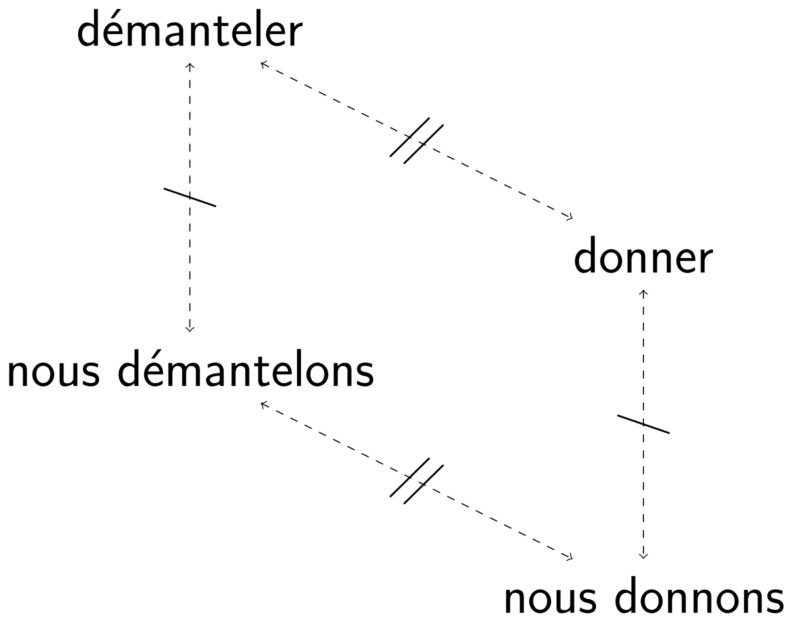
$$A : A :: C : x \qquad \Rightarrow \quad x = C \quad \text{et } C \text{ seule solution.}$$

En effet, $d(A, A) = 0 \Rightarrow d(C, x) = 0 \Rightarrow x = C$.

Visualisation



Visualisation

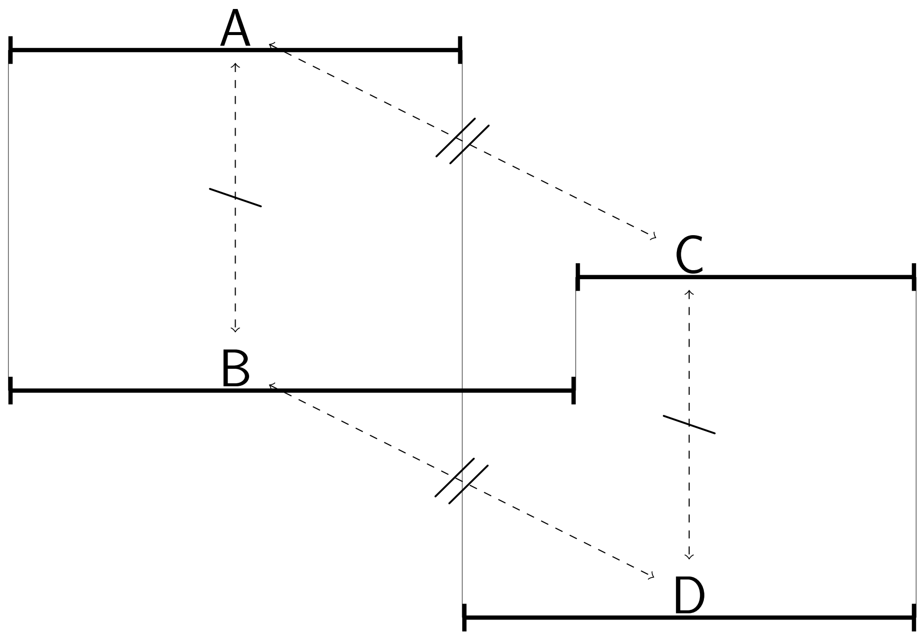


Egalité des sommes des longueurs des extrêmes et des moyens

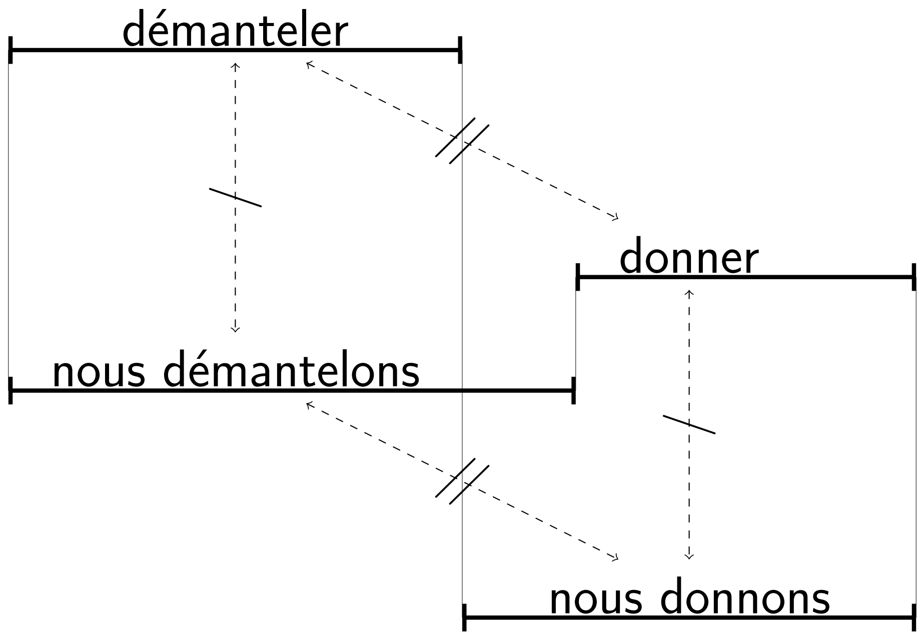
$$\forall (A, B, C, D) \in (\mathcal{V}^*)^4,$$

$$A : B :: C : D \quad \Rightarrow \quad |A| + |D| = |B| + |C|$$

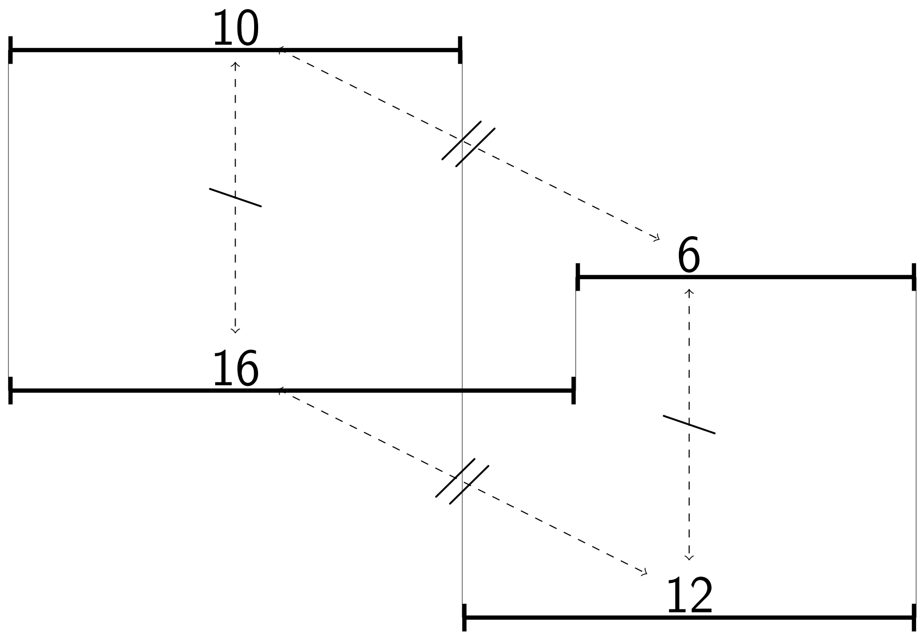
Visualisation



Visualisation



Visualisation



Deuxième observation

oratorem : *orator* :: *honorem* : *honor*

huzila : *huzāl* :: *sudi'a* : *sudā'*

setzen : *setzte* :: *lachen* : *lachte*

inné : *nées* :: *indu* : *dues*

biorąc : *bierzesz* :: *piorąc* : *pierzesz*

tinggal : *ketinggalan* :: *duduk* : *kedudukan*

Deuxième observation

$aa : ab :: ba : bb$

$ab : aabb :: aaaaaaabbbbbbb : aaaaaaaabbbbbbbb$

$abc : aabbcc :: aaabbbccc : aaaabbbbcccc$

$aab : aaaabb :: aaaaaabbb : aaaaaaaabbbb$

$aaaabbbbcccc : aaabbbccc :: aabbcc : abc$

Deuxième observation

Pour chaque symbole, on en a le même nombre dans A et D que dans B et C. En notant par $|A|_a$ l'effectif du symbole a dans A :

Soit $\mathcal{V}$ un alphabet, $\forall (A, B, C, D) \in (\mathcal{V}^*)^4$,

$$A : B :: C : D \quad \Rightarrow \quad |A|_a + |D|_a = |B|_a + |C|_a, \forall a$$

Deuxième observation

Pour chaque symbole, on en a le même nombre dans A et D que B et C. En notant par $|A|_a$ l'effectif du symboles a dans A :

Soit $\mathcal{V}$ un alphabet, $\forall (A, B, C, D) \in (\mathcal{V}^*)^4$,

$$A : B :: C : D \quad \Rightarrow \quad |A|_a - |B|_a = |C|_a - |D|_a, \forall a$$

Egalité des sommes des longueurs des extrêmes et des moyens

Trivialement, la propriété précédente implique l'égalité des sommes des longueurs des extrêmes et des moyens déjà vue plus haut.

Soit $\mathcal{V}$ un alphabet, $\forall (A, B, C, D) \in (\mathcal{V}^*)^4$,

$$A : B :: C : D \quad \Rightarrow \quad |A| + |D| = |B| + |C|$$

Formalisation (incomplète) de l'analogie de commutation

$$\forall (A, B, C, D) \in (\mathcal{V}^*)^4,$$

$$A : B :: C : D \quad \Rightarrow \begin{cases} d(A, B) = d(C, D) \\ |A|_a - |B|_a = |C|_a - |D|_a, \forall a \end{cases}$$

Les défauts des deux formalisations

$ab : ac :: db : cd$ (en plus de dc)

$$\begin{cases} d(ab, ac) = d(db, bd) = 2 \\ d(ab, db) = d(ac, cd) = 2 \end{cases}$$

$$adcb \in (ab \bullet dc) \cap (ac \bullet db)$$

Les défauts des deux formalisations

$ab : aabb :: aaabbb : aaababbb$ (en plus de $aaaabbbb$)

$$\begin{cases} d(ab, aabb) = d(aaabbb, aaababbb) = 2 \\ d(ab, aaabbb) = d(aabb, aaababbb) = 4 \end{cases}$$

$$aaaababbbb \in (ab \bullet aaababbb) \cap (aabb \bullet aaabbb)$$

Une piste de recherche : les morphismes

Quels morphismes conservent l'analogie de commutation (ex. : *aslama* : *muslim* :: *arsala* : *mursil*) :

- Miroir : *amalsa* : *milsum* :: *alasra* : *lisrum*
- Bijection sur l'alphabet : *btmbnb* : *nvtmjn* :: *bstbma* : *nvstjm*
- c $\longrightarrow$ cx : *axsxlxaxmxax* : *mxuxsxlxixmx* :: *axrxsxaxlxax* : *mxuxrxsxi*
- $B \longrightarrow C$: ? ? : *arsala* :: *muslim* : ? ?

Analogie et langages formels

Deux questions

Pourquoi sait-on que l'accusatif de *puella* est *puellam* dès que l'on sait que celui de *rosa* est *rosam* ?

Peut-on tirer de la réponse à la question précédente des langages formels qui aient de bonnes propriétés pour la description des langues ?

Classification de Chomsky-Schutzenberger

réguliers	$\subset$ hors-contexte	$\subset$ sous-contexte
rationnels	algébriques	
$\left\{\begin{array}{l} A \longrightarrow Ba \\ A \longrightarrow a \end{array}\right.$	$\left\{\begin{array}{l} A \longrightarrow BC \\ A \longrightarrow a \end{array}\right.$	$w \longrightarrow w' \ / \ \lvert w \rvert \leq \lvert w' \rvert$
$\{a^n / n > 0\}$	$\{a^n b^n / n > 0\}$	$\{a^n b^n c^n / n > 0\}$ $\{a^n b^n c^n d^n / n > 0\}$ $\{a^n b^m c^n d^m / n, m > 0\}$

Langues et hors-contexte

Les générativistes pensaient que la syntaxe des langues se décrit à l'aide de grammaires hors-contexte.

Théorie X-barre ($X = V$ ($S = \bar{\bar{V}}$), I, N) :

$$\begin{array}{lcl} \bar{\bar{X}} & \rightarrow & [\text{ spécificateur }]\ \bar{X} \\ \bar{X} & \rightarrow & \bar{X}\ [\text{ adjoint }] \\ \bar{X} & \rightarrow & X\ [\text{ complément }]^* \end{array}$$

Noam Chomsky, "Three models for the description of language". *IRE Transactions on Information Theory*, vol.2, 1956, pp.113-124.

Langues et hors-contexte

Exemple de structure en bambara :

wulu
chien

chien

Christopher Culy, The complexity of the vocabulary of Bambara, *Linguistics and Philosophy,* vol. 8, 1985, p. 345–351.

Langues et hors-contexte

Exemple de structure en bambara :

wulu	o wulu
chien	QUANTIF. UNIVERSEL

tout chien

Christopher Culy, The complexity of the vocabulary of Bambara, *Linguistics and Philosophy,* vol. 8, 1985, p. 345–351.

Langues et hors-contexte

Exemple de structure en bambara :

wulu	*(nyinina)*^{n}	*(filèla)*^{m}
chien	qqn cherchant	qqn regardant

(regarde quelqu'un qui)^{m} *(cherche quelqu'un qui)*$^{n-1}$ *cherche un chien*

Christopher Culy, The complexity of the vocabulary of Bambara, *Linguistics and Philosophy,* vol. 8, 1985, p. 345–351.

Langues et hors-contexte

Exemple de structure en bambara :

wulu	*(nyinina)*^{n}	*(filèla)*^{m}	*o wulu (nyinina)*^{n} *(filèla)*^{m}
chien	qqn cherchant	qqn regardant	QUANTIF. UNIVERSEL

quiconque (regarde quelqu'un qui)^m *(cherche quelqu'un qui)*$^{n-1}$ *cherche un chien*

Christopher Culy, The complexity of the vocabulary of Bambara, *Linguistics and Philosophy*, vol. 8, 1985, p. 345–351.

Langues et hors-contexte

Exemple de structure dans la variante zurichoise du suisse-allemand (!) :

(d'chind)^n	*(em Hans)*^{m}	*es huus haend wele*	*laa*^{n}	*hälfe*^{m}	*aastriche*
les enfants-ACC	Hans-DAT	la maison-ACC avoir voulu	laisser	aider	peindre

[...que nous] ayons voulu (laisser les enfants)^n (aider Hans à)^m peindre la maison

Stuart Shieber, Evidence against the context-freeness of natural language, *Linguistics and Philosophy*, vol. 8, 1985, p. 333–343.

Langues et hors-contexte

Une structure fondamentale qui n'est pas hors-contexte : $a^n.b^m.c^n.d^m$.

Une grammaire universelle devra donc pouvoir produire plus que du hors-contexte, mais pas tout le sous-contexte. Une grammaire universelle devra donc être strictement sous-contexte, mais pas trop.

Joshi propose le modérement sous-contexte

Le modérément sous-contexte

- Les langages modérément sous-contexte incluent tous les langages hors-contexte ;
- ils s'analysent en temps polynomial ;
- ils rendent seulement compte de certaines dépendances ;
- ils ont la propriété de croissance bornée des longueurs.

Joshi, Vijay-Shanker & Weir, *On the convergence of mildly context-sensitive formalisms*, 1991, p. 32

Analyse polynomiale

L'analyse doit s'effectuer en un temps polynomiale en la longueur de la phrase (en mots).

Pour mémoire, l'analyse s'effectue
– en $O(n^3)$ avec une grammaire hors-contexte ;
– en $O(n^6)$ avec une grammaire d'arbres adjoints (TAG).

Dépendances à représenter

– reduplication $w.a.w$;
– dépendance croisée $a^n.b^m.c^n.d^m$;
– accord multiple $a^n.b^n.c^n$.

Salomon Marcus, Gheorghe Paŭn, Carlos Martín-Vide, *Contextual grammars, a model for natural language,* 1998, p. 246

Croissance bornée des longueurs

Un langage vérifie la propriété de croissance bornée des longueurs si (et seulement si) quand on ordonne les chaînes du langage par longueurs croissantes, deux longueurs consécutives ne diffèrent pas par des différences devenant arbitrairement grandes.

Croissance bornée des longueurs

Un langage vérifie la propriété de croissance bornée des longueurs si (et seulement si) quand on ordonne les chaînes du langage par longueurs croissantes, deux longueurs consécutives ne diffèrent pas par des différences devenant arbitrairement grandes.

Exemple :	$a^n.b^n$
Contre-exemple :	a^{2^n}

Langages de chaînes analogiques

Exemple

$$\begin{array}{rcl} \mathcal{V} & = & \{\textit{A, \ldots, Z, a, \ldots, z, espace}\ \} \\ \mathcal{A} & = & \{\textit{il va à la campagne}\} \\ \mathcal{M} & = & \{\textit{il va} \longrightarrow \textit{ils vont, la campagne} \longrightarrow \textit{la ville}\} \end{array}$$

$$\textit{ils vont à la ville} \in \Lambda(\mathcal{A}, \mathcal{M})$$

Exemple

$$
\begin{aligned}
\mathcal{V} &= \{\textit{A, \dots, Z, a, \dots, z, espace}\,\} \\
\mathcal{A} &= \{\textit{il va à la campagne}\} \\
\mathcal{M} &= \{\textit{il va} \longrightarrow \textit{ils vont, la campagne} \longrightarrow \textit{la ville}\}
\end{aligned}
$$

$$\textit{ils vont à la ville} \in \Lambda(\mathcal{A}, \mathcal{M}) \ \ \textbf{?}$$

Exemple

$$\begin{aligned} \mathcal{V} &= \{A, \ldots, Z, a, \ldots, z, espace\ \} \\ \mathcal{A} &= \{\textit{il va à la campagne}\} \\ \mathcal{M} &= \{\textit{il va} \longrightarrow \textit{ils vont}, \textit{la campagne} \longrightarrow \textit{la ville}\} \end{aligned}$$

$$: \quad :: \quad \textit{il va à la campagne} \quad :$$

ils vont à la ville $\in \Lambda(\mathcal{A}, \mathcal{M})$ **?**

Exemple

$$\mathcal{V} = \{A, \ldots, Z, a, \ldots, z, \textit{espace}\}$$
$$\mathcal{A} = \{\textit{il va à la campagne}\}$$
$$\mathcal{M} = \{\textit{il va} \longrightarrow \textit{ils vont}, \textit{la campagne} \longrightarrow \textit{la ville}\}$$

$$: \quad :: \quad \textit{il va à la campagne} \quad :$$

$\textit{ils vont à la ville} \in \Lambda(\mathcal{A}, \mathcal{M})$ **?**

Exemple

$$\begin{array}{rcl} \mathcal{V} & = & \{\textit{A, ..., Z, a, ..., z, espace}\} \\ \mathcal{A} & = & \{\textit{il va à la campagne}\} \\ \mathcal{M} & = & \{\textit{il va} \longrightarrow \textit{ils vont}, \textit{la campagne} \longrightarrow \textit{la ville}\} \end{array}$$

$$\textit{il va} : \textit{ils vont} :: \textit{il va à la campagne} : x$$

$$\textit{ils vont à la ville} \in \Lambda(\mathcal{A}, \mathcal{M}) \ \textbf{?}$$

Exemple

$$
\begin{aligned}
\mathcal{V} &= \{\textit{A, \ldots, Z, a, \ldots, z, espace}\ \} \\
\mathcal{A} &= \{\textit{il va à la campagne}\} \\
\mathcal{M} &= \{\textit{il va} \longrightarrow \textit{ils vont}, \textit{la campagne} \longrightarrow \textit{la ville}\}
\end{aligned}
$$

il va : *ils vont* :: *il va à la campagne* : *ils vont à la campagne*

ils vont à la ville $\in \Lambda(\mathcal{A}, \mathcal{M})$ **?**

Exemple

$$\begin{aligned}\mathcal{V} &= \{A, \ldots, Z, a, \ldots, z, \textit{espace}\ \} \\ \mathcal{A} &= \{\textit{il va à la campagne}\} \\ \mathcal{M} &= \{\textit{il va} \longrightarrow \textit{ils vont, la campagne} \longrightarrow \textit{la ville}\}\end{aligned}$$

: :: : *ils vont à la campagne*

ils vont à la ville $\in \Lambda(\mathcal{A}, \mathcal{M})$ **?**

Exemple

$$\begin{array}{rcl} \mathcal{V} & = & \{\textit{A, \ldots, Z, a, \ldots, z, espace}\ \} \\ \mathcal{A} & = & \{\textit{il va à la campagne}\} \\ \mathcal{M} & = & \{\textit{il va} \longrightarrow \textit{ils vont, la campagne} \longrightarrow \textit{la ville}\} \end{array}$$

$$: \qquad\qquad :: \ \begin{array}{c}\textit{ils vont à la}\\ \textit{campagne}\end{array} :$$

$$\textit{ils vont à la ville} \in \Lambda(\mathcal{A}, \mathcal{M}) \ \ \textbf{?}$$

Exemple

$$
\begin{aligned}
\mathcal{V} &= \{A,\ \ldots,\ Z,\ a,\ \ldots,\ z,\ \textit{espace}\ \} \\
\mathcal{A} &= \{\textit{il va à la campagne}\} \\
\mathcal{M} &= \{\textit{il va} \longrightarrow \textit{ils vont},\ \textit{la campagne} \longrightarrow \textit{la ville}\}
\end{aligned}
$$

$$ \quad : \qquad\qquad :: \ \begin{array}{l}\textit{ils vont à la}\\ \textit{campagne}\end{array} \ : $$

$$\textit{ils vont à la ville} \in \Lambda(\mathcal{A}, \mathcal{M}) \ \ \textbf{?}$$

Exemple

$$\begin{aligned}\mathcal{V} &= \{\textit{A, \ldots, Z, a, \ldots, z, espace}\,\} \\ \mathcal{A} &= \{\textit{il va à la campagne}\} \\ \mathcal{M} &= \{\textit{il va} \longrightarrow \textit{ils vont, la campagne} \longrightarrow \textit{la ville}\}\end{aligned}$$

la campagne : *la ville* :: *ils vont à la campagne* : x

ils vont à la ville $\in \Lambda(\mathcal{A}, \mathcal{M})$ **?**

Exemple

$$
\begin{aligned}
\mathcal{V} &= \{\textit{A, \ldots, Z, a, \ldots, z, espace}\ \} \\
\mathcal{A} &= \{\textit{il va à la campagne}\} \\
\mathcal{M} &= \{\textit{il va} \longrightarrow \textit{ils vont}, \textit{la campagne} \longrightarrow \textit{la ville}\}
\end{aligned}
$$

la campagne : *la ville* :: *ils vont à la campagne* : *ils vont à la ville*

ils vont à la ville $\in \Lambda(\mathcal{A}, \mathcal{M})$ **?**

Exemple

$$
\begin{aligned}
\mathcal{V} &= \{\textit{A, \ldots, Z, a, \ldots, z, espace}\ \} \\
\mathcal{A} &= \{\textit{il va à la campagne}\} \\
\mathcal{M} &= \{\textit{il va} \longrightarrow \textit{ils vont, la campagne} \longrightarrow \textit{la ville}\}
\end{aligned}
$$

ils vont à la

ville

$\textit{ils vont à la ville} \in \Lambda(\mathcal{A}, \mathcal{M})$ **?**

Exemple

$$
\begin{aligned}
\mathcal{V} &= \{A, \ldots, Z, a, \ldots, z, \textit{espace}\ \} \\
\mathcal{A} &= \{\textit{il va à la campagne}\} \\
\mathcal{M} &= \{\textit{il va} \longrightarrow \textit{ils vont}, \textit{la campagne} \longrightarrow \textit{la ville}\}
\end{aligned}
$$

$$\textit{ils vont à la ville} \in \Lambda(\mathcal{A}, \mathcal{M}) \quad \text{:-)}$$

Langages de chaînes analogiques

$\mathcal{A}$ est appelé l'ensemble des chaînes attestées et $\mathcal{M}$ l'ensemble des modèles.

Définition par couches successives

Soit $\mathcal{V}$ un alphabet. Soient $\mathcal{A} \subset \mathcal{V}^*$ et $\mathcal{M} \subset \mathcal{V}^* \times \mathcal{V}^*$, tous deux finis et non vides. Alors, le langage de chaînes analogiques $\Lambda(\mathcal{A}, \mathcal{M})$ est défini de la façon suivante.

$$
\begin{aligned}
\Lambda^0 &= \mathcal{A} \\
\Lambda^{n+1} &= \{\ D \in \mathcal{V}^* \ /\ \exists (A, B, C) \in \mathcal{M} \times \Lambda^n, \\
&\qquad\qquad A : B :: C : D\ \ \} \\
\Lambda(\mathcal{A}, \mathcal{M}) &= \bigcup_{i=0}^{+\infty} \ \Lambda^i
\end{aligned}
$$

Dérivation immédiate

Soit $\mathcal{V}$ un alphabet. Soit $\mathcal{M} \subset \mathcal{V}^* \times \mathcal{V}^*$ dont les éléments (A, B) sont notés $A \rightarrow B$. La dérivation analogique immédiate modulo $\mathcal{M}$, notée $\vdash_{\overline{\mathcal{M}}}$, est définie de la façon suivante.

$$\forall (C, D) \in \mathcal{V}^* \times \mathcal{V}^*, \quad C \vdash_{\overline{\mathcal{M}}} D$$

$$\Leftrightarrow \quad \exists A \rightarrow B \in \mathcal{M} \; / \quad A : B :: C : D$$

Définition par système dérivationnel

Soit $\mathcal{V}$ un alphabet. Soient $\mathcal{A} \subset \mathcal{V}^*$ et $\mathcal{M} \subset \mathcal{V}^* \times \mathcal{V}^*$, tous deux finis et non vides. Alors, $\Lambda(\mathcal{A}, \mathcal{M}) = \, < \mathcal{A}, \{\vdash_{\overline{\mathcal{M}}}\} >$ est le langage de chaînes analogiques défini de la façon suivante

$$\Lambda(\mathcal{A}, \mathcal{M}) = \mathcal{A} \cup \{ \, D \in \mathcal{V}^* \, / \, \exists C \in \mathcal{A}, \; C \overset{+}{\vdash_{\overline{\mathcal{M}}}} D \, \}$$

avec $\overset{+}{\vdash_{\overline{\mathcal{M}}}}$, la fermeture transitive de la dérivation analogique immédiate $\vdash_{\overline{\mathcal{M}}}$.

Préoccupation initiale

Reproduction sur machine de certains mécanismes de la langue ou de certains processus cognitifs, voire de l'acquisition des langues.

> It is hard to imagine our brain using auxiliary intermediate sentences of a nonterminal type. Instead, it looks more natural [. . .] to start with a collection of well-formed sentences, maybe acquired from experience, and to produce new well-formed ones by adding further words, in pairs that can observe dependencies and agreements [. . .]

[Marcus, Martín-Vide et Păun, 1998]

Les langages de chaînes analogiques ne font pas intervenir de symboles non-terminaux. :-)

Dépendances

$\boxed{\{w.a.w\}}$

Les langages du type $\{w.a.w\}$ seraient des langages de chaînes analogiques si on considérait les analogies par reduplication : ;-(

$$A : A.C.A :: B : B.C.B$$

$$\{a^n \ / \ \ n > 0\}$$

$$\Lambda(\{a\}, \{a \rightarrow aa\}) = \{a^n / n > 0\}$$

Démonstration : par induction, avec la définition des langages de chaînes analogiques par couches successives : $\Lambda(\mathcal{A}, \mathcal{M}) = \bigcup_{i=0}^{+\infty} \ \Lambda^i$.
Base : $\Lambda^0 = \{a = a^{0+1}\}$.
Induction : on suppose $\Lambda^n = \{a^{n+1}\}$. Comme

$$\Lambda^{n+1} = \{ \ D \in \mathcal{V}^* \ / \ \exists (a, aa, C) \in \{a, aa\} \times \Lambda^n, \ \ a : aa :: C : D \ \ \}$$

on a trivialement : $\Lambda^{n+1} = \{a^{n+2}\}$. cqfd

$$\boxed{\{a^n.b^n \ / \ \ n > 0\}}$$

$$\Lambda(\{ab\}, \{ab \rightarrow aabb\}) = \{a^n b^n / n > 0\}$$

Démonstration : On voudrait faire un raisonnement similaire au précédent sauf que, rigoureusement, selon la définition incomplète de l'analogie, $a^{n+2}.b^{n+2}$ n'est pas l'unique solution à l'analogie $\quad ab : aabb :: a^{n+1}.b^{n+1} : x \quad$ On est obligé, pour les besoins de la cause d'introduire une hypothèse spéciale, celle de la concaténation des analogies disjointes. cqfd ? ? ?

Concaténation d'analogies disjointes

Soient $\mathcal{V}$ et $\mathcal{V}'$ tels que $\mathcal{V} \cap \mathcal{V}' = \phi$.
Soient $(A, B, C, D) \in (\mathcal{V}^*)^4$
et $(A', B', C', D') \in (\mathcal{V}'^*)^4$ tels que

$$A : B :: C : D \quad \text{et} \quad A' : B' :: C' : D'$$

Alors, l'équation analogique $A.A' : B.B' :: C.C' : x$ admet une et une seule solution, $D.D'$.

$$\boxed{\{a^n.b^n \ / \ \ n > 0\}}$$

$$\Lambda(\{ab\}, \{ab \to aabb\}) = \{a^n b^n / n > 0\}$$

Démonstration : Raisonnement similaire à celui pour $\{a^n \ / \ \ n > 0\}$. L'hypothèse de la concaténation des analogies disjointes étant admise, $a^{n+2}.b^{n+2}$ est l'unique solution de l'équation analogique $ab : aabb :: a^{n+1}.b^{n+1} : x$. cqfd

$$\boxed{\{a^n.b^n.c^n \ / \ \ n > 0\}}$$

$$\Lambda(\{abc\}, \{abc \rightarrow aabbcc\}) = \{a^n b^n c^n / n > 0\}$$

Démonstration : Raisonnement similaire à celui pour $\{a^n b^n \ / \ \ n > 0\}$. L'hypothèse de la concaténation des analogies disjointes admise, $a^{n+2}.b^{n+2}.c^{n+2}$ est l'unique solution de l'équation analogique $abc : aabbcc :: a^{n+1}.b^{n+1}.c^{n+1} : x$.

cqfd

$$\boxed{\{a_1^n.a_2^n.\dots.a_m^n \ / \ \ n > 0\}}$$

$$\Lambda(\{a_1.a_2\dots a_m\}, \{a_1.a_2\dots a_m \rightarrow a_1.\, a_1.a_2.a_2\dots a_m.a_m\})$$
$$= \{a_1^n.a_2^n.\dots.a_m^n/n > 0\}$$

Démonstration : Par induction sur m, étant admis l'hypothèse de la concaténation des analogies disjointes. cqfd

Dépendances : $a^n.b^m.c^n.d^m$

$$\Lambda(\{abcd\}, \{abcd \rightarrow abbcdd, abcd \rightarrow aabccd\})$$

$$= \{a^n b^m c^n d^m \ / \ n, m > 0\}$$

Dépendances

Les langages a^n (régulier), $a^n.b^n$ (hors-contexte), $a^n.b^n.c^n$ ou $a_1^n.\ldots.a_m^n$ (sous-contexte) sont tous des langages de chaînes analogiques et ils ont la même simplicité lorsqu'ils sont décrits ainsi. :-)

Croissance bornée des longueurs

Croissance bornée des longueurs

Soit $\mathcal{L}$ un langage. On note $L(\mathcal{L}) \subset \mathbb{N}$ l'ensemble des longueurs des chaînes de $\mathcal{L}$. Soit l un élément de $L(\mathcal{L})$. On note par $\mathsf{s}(l)$ le plus petit élément supérieur strictement à l dans $L(\mathcal{L})$ quand il existe. Un langage $\mathcal{L}$ vérifie la propriété de croissance bornée des longueurs si et seulement si

$$L(\mathcal{L}) \text{ est fini} \quad \vee \quad \exists k \in \mathbb{N} \;/\; \forall l \in L(\mathcal{L}),\ \mathsf{s}(l) - l \leq k$$

Croissance bornée des longueurs

Tout langage de chaînes analogiques a la propriété de croissance bornée des longueurs.

Croissance bornée des longueurs

Toute union finie de langages formels ayant la propriété de croissance bornée des longueurs a trivialement la propriété de croissance bornée des longueurs.

Tout langage de chaînes analogiques est trivialement l'union finie des langages de chaînes analogiques ayant même ensemble de modèles et comme ensemble de chaînes attestées l'une des chaînes attestées (donc un singleton).

$$\Lambda(\mathcal{A}, \mathcal{M}) = \bigcup_{a \in \mathcal{A}} \Lambda(\{a\}, \mathcal{M})$$

Croissance bornée des longueurs

Tout langage de chaînes analogiques dont l'ensemble des chaînes attestées est un singleton vérifie la propriété de croissance bornée des longueurs.

Démonstration : On pose $\mathcal{A} = \{C_0\}$ et $k = |\max_{A \to B \in \mathcal{M}} (|B| - |A|)|$.

$\forall D \in \Lambda^{n+1},\ \ \exists C \in \Lambda^n,\ \ \exists A \to B \in \mathcal{M}\ \ /\ \ A : B :: C : D$

$$\begin{array}{lll} A : B :: C : D & \Rightarrow & |A| + |D| = |B| + |C| \\ & \Rightarrow & |D| - |C| = |B| - |A| \leq k \end{array}$$

Par induction sur le n de Λ^n : $\forall l \in L(\Lambda(\mathcal{A}, \mathcal{M})), \quad s(l) - l \ \leq k$ cqfd

Conclusion

Deux propositions de formalisation de l'analogie de commutation. Elles sont inachevées au sens où elles ne rendent pas compte exactement de tous les exemples et contre-exemples possibles, langagiers comme formels.

Des problèmes d'apparence simple restent à résoudre. Comment avoir :

$$ab : ac :: db : x \quad \Rightarrow \quad x = dc \quad \text{et pas } cd\,?$$

Ou bien :

$$ab : aabb :: aaabbb : x \quad \Rightarrow \quad x = aaaabbbb \quad \text{seulement}\,?$$

Une proposition de langages formels, fondée sur l'analogie et ayant un certain nombre des propriétés requises pour le modérément sous-contexte.

Il n'y a pas de symboles non-terminaux dans les grammaires de chaînes analogiques Ceci est peu habituel dans les descriptions classiques des grammaires formelles.

Si l'analogie est bien un procédé d'apprentissage des langues, alors cette proposition pourrait constituer un bon candidat pour la description des langues.

Intérêt récent pour l'analogie en traitement automatique des langues : 3 articles à TALN 2009.

Nammer, Hathout, Claveau, L'Homme, Yvon, Stroppa, Langlais, Zweigenbaum, ...

Singh, Ford, Goldsmith, Turney, ...

Merci de votre attention.