Proportion analogique entre arbres : une étude bibliographique

Anouar Ben Hassena

Publication interne IRISA numéro 1912

Proportion analogique entre arbres : une étude bibliographique.

Anouar BEN HASSENA

Projet CORDIAL

Résumé : En intelligence artificielle, l'analogie est utilisée comme une technique de raisonnement non exact pour la résolution de problèmes, la compréhension du langage naturel et l'apprentissage de règles de classification, etc. Nous essayons de l'utiliser pour la prédiction de la prosodie en synthèse de la parole. Dans ce rapport, nous présentons les travaux liés à l'analogie dans la représentation que nous employons pour la prosodie : les arbres ordonnées et étiquetés. Pour ce type de structure, nous proposons notre propre définition de la proportion analogique et nous montrons ses enjeux par rapport aux autres définitions antérieures.

Mots clés : Synthèse de la parole, analogie, arbres étiquetés ordonnés

(Abstract: pto)

IRISA, Campus universitaire de Beaulieu, 35042 Rennes Cedex (France)

Téléphone : (33) 02 99 84 71 00 — Télécopie : (33) 02 99 84 71 71 — www.irisa.fr

Analogical proportion between trees: a bibliographic study

Abstract: In Artificial Intelligence, analogy is used as a non exact reasoning technique to solve problems, for natural language understanding, for learning classification rules, etc. We try to use it for prosody prediction in speech synthesis. In this report, we present the works that have been produced to study analogy on labeled ordered trees (which is the representation that we use for prosody). We propose a new definition of analogical proportion and we show its issues in comparison with the previous definitions.

Key-words: Speech synthesis, analogy, labeled ordered trees.

Table des matières

1 Introduction

Dans cette section, nous rappelons le principe de l'analogie, les notions de base de la proportion analogique et nous donnons la terminologie utilisée dans l'univers des arbres.

1.1 Principe de l'analogie

L'analogie, telle que nous l'utiliserons dans ce document, se fonde sur la similitude de termes dans un certain univers pouvant s'énoncer dans une proportion: "A est à B ce que C est à D". Elle peut donc être utilisée donc, par un mécanisme de mise en correspondance de structures ou de concepts, à transposer des connaissances d'un objet à un autre.

Soit une ensemble $\mathcal{S}$ d'objets connus, définis selon un ensemble de valeurs des attributs c_1,c_2,...,c_n (prenant leurs valeurs dans des domaines éventuellement différents) et x un nouvel objet dont seule une partie des attributs c_1,...,c_n est connue. Une inférence analogique détermine les caractéristiques inconnues de x en suivant un schéma en deux étapes :

1. toutes les proportions analogiques mettant en jeu les attributs connus de x et les attributs correspondants de trois objets de $\mathcal{S}$ sont identifiées ;

2. pour chacun des triplets d'objets de $\mathcal{S}$ identifiés, le transfert analogique est calculé pour obtenir les attributs inconnus de x. Si le transfert réussit, son résultat est ajouté aux solutions possibles.

1.2 Proportion analogique

Définition 1.1. *Une proportion analogique sur un ensemble X est une relation sur X^4 qui respecte trois axiomes. Nous notons quatre éléments $(A,B,C,D) \in X$, qui sont donc en proportion analogique, par : $A : B :: C : D$, ce qui se lit "A est à B ce que C est à D". Les axiomes sont les suivants :*

– *Symétrie de la relation "'est à"' : $A : B :: C : D \Leftrightarrow C : D :: A : B$.*

- *Échange des moyens :* $A : B :: C : D \Leftrightarrow A : C :: B : D$.
- *Déterminisme :* $A : B :: A : x \Rightarrow x = B$ *et* $A : A :: B : x \Rightarrow x = B$.

1.3 Définitions et terminologie

Soit un arbre T. Nous notons l'ensemble de ses nœuds (les feuilles sont inclues) par $\mathcal{N}(T)$.

Arbre enraciné. Un *arbre enraciné* est un ensemble de nœuds et un ensemble d'arcs satisfaisant trois propriétés :

- Il existe un nœud particulier appelé *racine*.
- Tout nœud n, autre que la racine, est relié par un arc à un unique autre nœud p appelé *père* de n. Similairement, n est appelé *fils* de p.
- On peut atteindre la racine à partir de n'importe quel nœud de l'arbre, en se déplaçant de père en père.

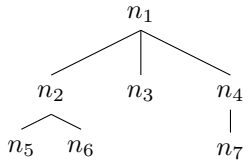
On note aussi un arbre enraciné par $T(r)$, où r est la racine.

Dans la suite, nous ne traiterons que d'arbres enracinés.

Arbre ordonné. Un arbre est *ordonné* s'il existe un ordre parmi les fils de chaque nœud. Par exemple soit l'arbre $T(n_1)$ ordonné de racine n_1 (dans cet exemple, l'ordre des indices des nœuds est compatible avec l'ordre des fils) :

Dans la suite, nous ne traiterons que d'arbres enracinés.

Arbre ordonné. Un arbre est *ordonné* s'il existe un ordre parmi les fils de chaque nœud. Par exemple soit l'arbre $T(n_1)$ ordonné de racine n_1 (dans cet exemple, l'ordre des indices des nœuds est compatible avec l'ordre des fils) :



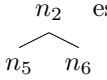
Ancêtre ou ascendant, descendant Les *ascendants* ou *ancêtres* d'un nœud n sont les nœuds par lesquels on passe pour atteindre la racine (racine et nœud n inclus). Si n_1 est ascendant de n_2, par définition n_2 est descendant de n_1.

Ancêtre propre Un *ancêtre propre* d'un nœud n est un ancêtre de n (n exclus). Par exemple, dans $T(n_1)$, n_2 est un ancêtre propre de n_5, mais pas n_2.

Étiquette. Une *étiquette* d'un nœud n_i (étiquette(n_i))est un élément d'un alphabet fini Σ étendu par le symbole *vide* λ que nous notons $\Sigma_\lambda = \Sigma \cup \{\lambda\}$.

Arbre étiqueté. Un arbre est *étiqueté* s'il existe une injection de l'ensemble des nœuds $\mathcal{N}(T)$ dans un ensemble d'étiquettes : on assigne à chaque nœud une étiquette de Σ_λ.

Sous-arbre. Un *sous-arbre* est constitué d'un nœud n_i ainsi que de tous ses descendants. On l'appelle sous arbre de T de racine n_i.

Par exemple,  est un sous arbre de $T(n_1)$

Hauteur. La hauteur d'un nœud n_i est la longueur du chemin à la feuille la plus distante dans le sous-arbre de de n_i (par exemple, la hauteur de n_1 est 2).

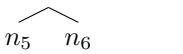
Profondeur. La profondeur (ou le niveau) d'un nœud n_i est la longueur du chemin qui mène à n_i à partir de la racine (par exemple, la profondeur de n_5 est 2).

Forêt. Une forêt est un ensemble fini ordonné d'arbres. Dans ce document nous associons une forêt à un arbre. Nous notons $F(n_1)$ la forêt obtenue à partir de $T(n_1)$ en supprimant n_1:

$F(n_1) =$ 

Cette forêt est constituée de tous les sous arbres enraciné par tous les fils de n_1, soit $T(n_2)$, $T(n_3)$ et $T(n_4)$. On la note aussi $F(n_2,n_4)$.

Forêt partielle. Une forêt partielle est une partie d'une forêt dont les arbres constituants sont adjacents. Par exemple, une forêt partielle de $F(n_1)$ est :

$F\ (n_2,\ n_3) =$ 

Cette forêt partielle de $F(n_1)$ est constituée par les sous arbres $T(n_2)$ et $T(n_3)$.

Parcours préfixé. Le parcours préfixé dans un arbre consiste à traiter la racine, puis à parcourir les sous arbres de gauche à droite. Exemple: Sur l'arbre $T(n_1$ les nœuds sont traités dans l'ordre : n_1,n_2,n_5,n_6,n_3,n_4,n_7.

Parcours postfixé. Le parcours postfixé consiste à parcourir les sous arbres de gauche à droite et puis traiter la racine. Exemple: Sur l'arbre $T(n_1)$ les nœuds sont traités dans l'ordre : n_5,n_6,n_2,n_3,n_7,n_4,n_1.

Position d'un nœud. Les nœuds d'un arbre T sont identifiés par leur position. Une position est une séquence d'entiers. ϵ est la position du racine, 1 le premier fils le plus à gauche (*leftmost*) de la racine, etc (voir la figure 1). L'ensemble des positions des nœuds de $\mathcal{N}(T)$ est noté $Pos(T)$.

Nœud à gauche d'un autre. Un nœud N de position $Np_1.....Np_n$ est *à gauche* d'un nœud M de position $Mp_1.....Mp_m$ *ssi*,

- N est ni un ancêtre ni un descendant de M, et
- $\exists\ i \in [1,min(n,m)]$ tel que $Np_1.....Np_{i-1} = Mp_1.....Mp_{i-1}$ et $Np_i < Mp_i$.

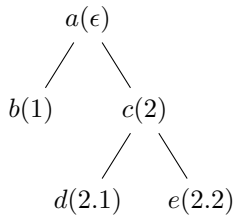


FIG. 1 – *Exemple d'un arbre T étiqueté. Les positions des nœuds sont entre parenthèses.*

Isomorphe. Deux arbres T_1 et T_2 sont dites isomorphes s'ils ont le même nombre n de nœuds, et si, pour tout $i \in Pos(T_1)$, le nœud de position i dans T_1 a le même nombre de sous-arbres que le nœud de position i dans l'arbre T_2.
Autrement dit, si deux arbres sont isomorphes, ils sont structurellement identiques.

2 Travaux antérieurs

Dans cette section, nous décrivons le seul (à notre connaissance) travail antérieur qui a traité le problème de l'analogie entre arbres, celui de Stroppa et Yvon [2].

2.1 Proportion analogique par factorisation

Dans [2], les auteurs ont proposé une définition générale de la proportion analogique, qui s'appuie sur un cadre algébrique générique et se décline pour un grand nombre de structures tels que les séquences et les arbres étiquetés.

L'idée est de pouvoir décomposer les objets en question sous forme d'une factorisation. Pour cela, et pour conserver cette idée générale, dans les arbres ils ont utilisé la notion de substitution.

Définition 2.1. ***(Arbre)***.
Les auteurs reprennent la définition d'un arbre de la section 1.3, mais ils l'étendent à un ensemble de nœuds $U = L \cup V$ où L un ensemble fini d'étiquettes (ou labels) et V un ensemble de variables. Autrement dit, un nœud peut être étiqueté par un label ou par une variable.

Dans la suite les auteurs ne considèrent que les arbres dont les variables n'apparaissent que dans les feuilles. De plus, ils imposent qu'une variable n'apparaisse qu'une seule fois dans un arbre.

Définition 2.2. ***(Substitution)***.
Une substitution est une paire $(v \leftarrow t)$ où v est une variable et t un sous arbre. Appliquer la substitution $(v \leftarrow t')$ à un arbre t consiste à remplacer la feuille de t étiquetée par v par le sous arbre t'.

Définition 2.3. ***(Factorisation)***.
Une factorisation de l'arbre t est une suite de substitutions dont le résultat est t (rappel : on a imposé que chaque substitution soit faite sur un seul nœud, qui est une feuille

Définition 2.4. ***(Proportion analogique entre arbres)***.
Les quatre arbres (x,y,z,t) sont en proportion analogique, notée $x : y :: z : t$ *si et seulement s'il existe quatre factorisations* $[s_x,v]$, $[s_y,v]$, $[s_z,v]$, $[s_t,v]$ *de x, y, z et t respectivement, telles que :*

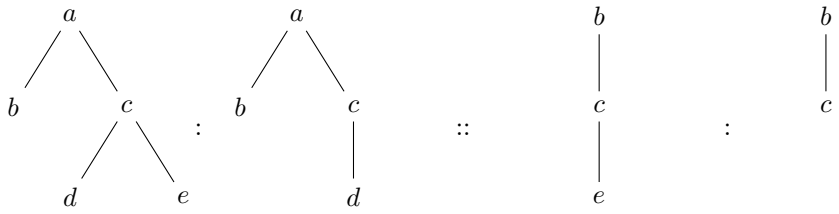
$$s_x = (x_1,...,x_n),$$
$$s_y = (y_1,...,y_n),$$
$$s_z = (z_1,...,z_n),$$
$$s_t = (t_1,...,t_n),$$

avec, $\forall\ i \in [1,n]$ *:*
- $(y_i,z_i) \in \{(x_i,t_i),(t_i,x_i)\}$, *et*
- $V(x_i) = V(y_i) = V(z_i) = V(t_i)$ *($V(x_i)$ est l'ensemble des variables de l'arbre x_i).*

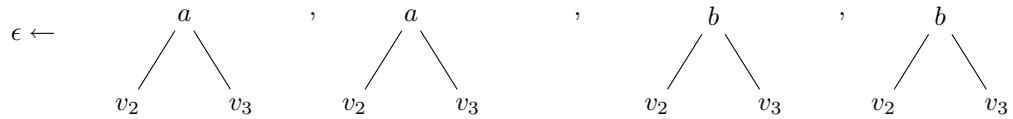
2.1.1 Exemples d'arbres en proportion analogique

Nous présentons dans ce qui suit des exemples d'arbres en proportion analogique par factorisation. Nous détaillons la construction de cette analogie en vérifiant à chaque étape les conditions imposées par les auteurs.

Exemple 1 :



Nous partons de quatre arbres vides ϵ et nous appliquons à chacun une première substitution : respectivement x_1, y_1, z_1, et t_1 :



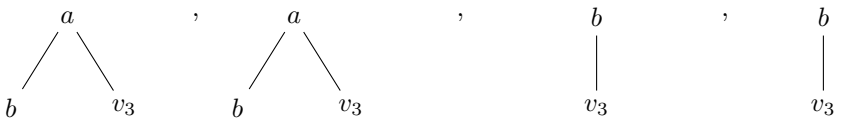
A chaque substitution nous vérifions bien les conditions de la définitions :
- $(y_1,z_1) \in \{(x_1,t_1),(t_1,x_1)\}$, et

– $V(x_1) = V(y_1) = V(z_1) = V(t_1)$ (rappelons que V est l'ensemble des variables).

Nous continuons la suite des substitutions jusqu'à construire les arbres en question.

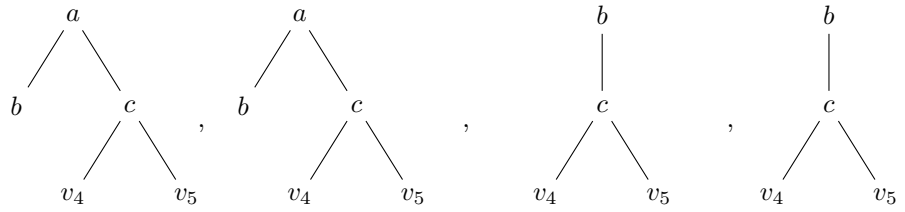
$v_2 \leftarrow$ b , b , ϵ , ϵ

$\hookrightarrow$



$v_3 \leftarrow$

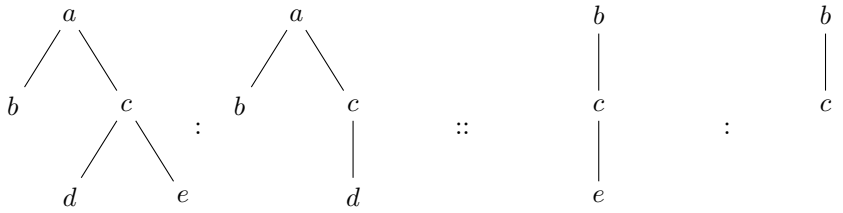
$\hookrightarrow$



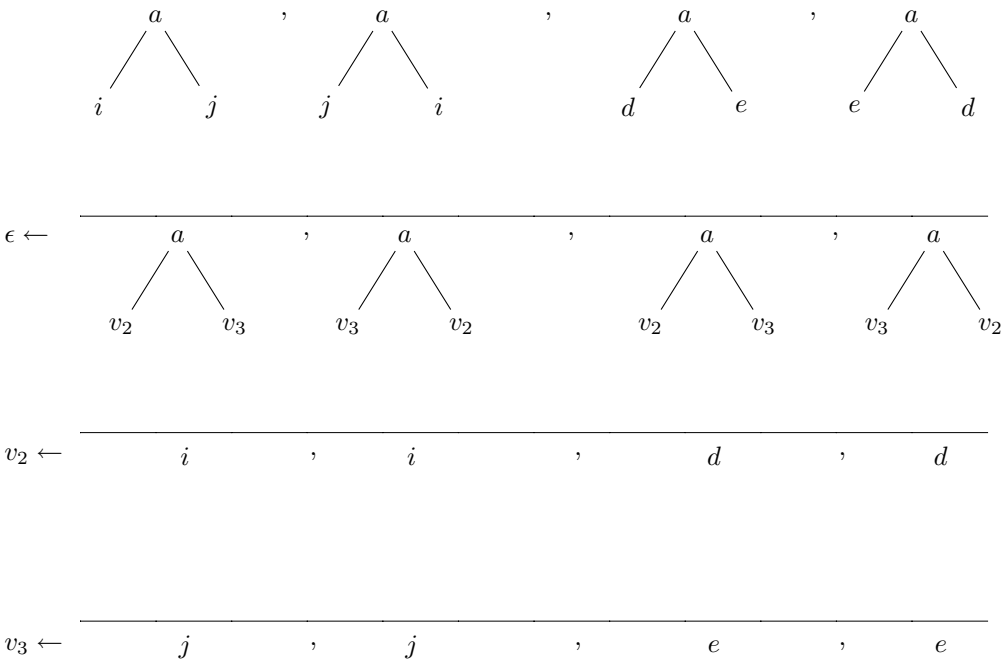
$v_4 \leftarrow$ d , d , ϵ , ϵ

$v_5 \leftarrow$ e , ϵ , e , ϵ

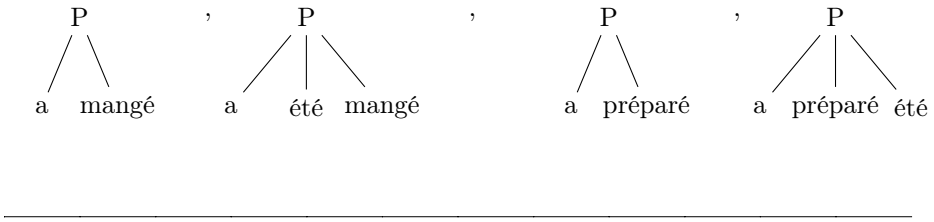
$\hookrightarrow$

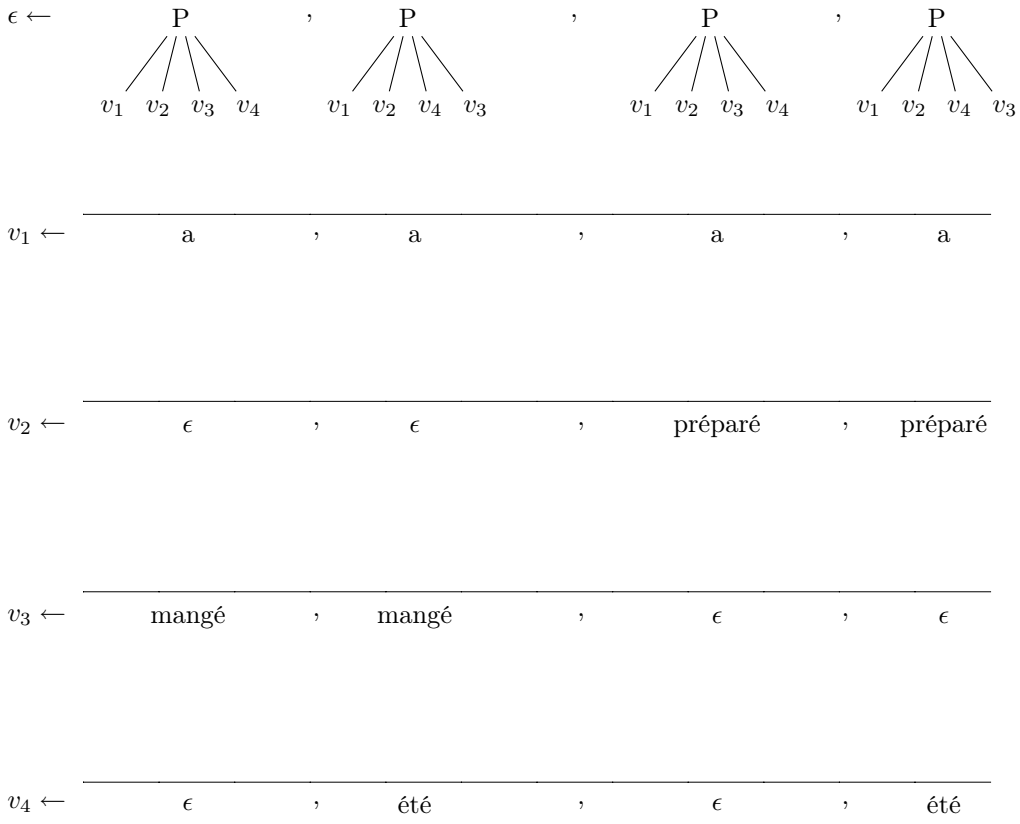


Exemple 2 :



Exemple 3 :





2.1.2 Calcul de la proportion analogique par factorisation

Nous décrivons rapidement la manière de calculer et valider la proportion analogique. Les auteurs présentent de manière détaillée un algorithme de complexité, dans le pire des cas, au moins exponentielle en le nombre des nœuds des arbres. Cet algorithme construit des factorisations des arbres selon la définition de la proportion analogique 2.1.4.
Dans une première étape, il construit pour chaque arbre $\alpha \in \{x,y,z,t\}$ un automate à états finis représentant toutes les factorisations possibles de α.
La deuxième étape consiste à chercher à partir des automates quatre factorisations (ou tous les quadruplets de factorisations) qui valident la définition 2.1.4.
Ces étapes sont détaillées dans [2].

Complexité Les auteurs montrent dans [2] que la complexité de construire un automate pour chaque arbre est (au moins) exponentielle en le nombre de ses nœuds.

La complexité de l'étape d'extraction et validation, dans le pire des cas, est donc exponentielle en le nombres des nœuds des arbres. Les auteurs conjecturent que le problème de décision (quatre arbres sont-ils en proportion analogique?) est *NP-complet*.

3 Analogie et alignement des arbres

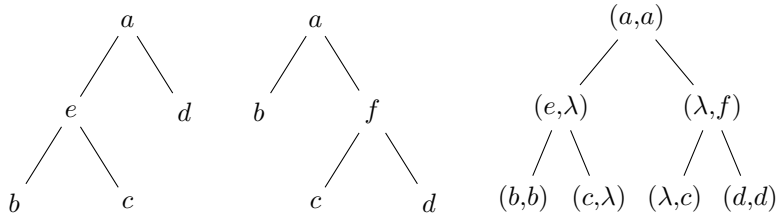
3.1 Proportion analogique par alignement

Nous présentons dans ce qui suit une définition différente de la proportion analogique entre quatre arbres.

Nous rappelons tout d'abord la définition de l'alignement entre deux arbres ordonnés et étiquetés sur un alphabet Σ :

Définition 3.1. *Un alignement entre deux arbres t_1 et t_2 de structure différentes est un arbre A obtenu, tout d'abord, par l'insertion des nœuds d'étiquette vide λ dans t_1 et t_2 de façon à ce qu'ils soient isomorphes, puis par leur superposition.*

De façon informelle, un alignement se fait par un appariement nœud à nœud de même position entre les deux arbres, dans lesquel un ou plusieurs nœuds d'étiquette vide ont d'abord été insérées. L'appariement (λ, λ) n'est pas permis.



(a) Arbre t_1 (b) Arbre t_2 (c) Alignement de t_1 et t_2

FIG. 2 – *Exemple d'un alignement entre arbres ordonnés*

Cette notion d'alignement entre deux arbres a été introduit par Jiang et al. dans [1] comme une alternative pour la distance d'édition entre arbres. Nous étendons cette notion d'alignement à un plus grand nombre d'arbres et en particulier à quatre arbres afin de définir une proportion analogique entre arbres comme suit.

Définition 3.2. .
Soit x, y, z et t quatre arbres d'étiquettes dans Σ. On suppose qu'une analogie dans Σ_λ est

définie. Nous disons que ces arbres sont en proportion analogique s'il existe un alignement des quatre arbres x', y', z' et t' d'étiquettes $\in \Sigma_\lambda$, vérifiant :

– $\forall\ i \in Pos$, *l'analogie* $x'_i : y'_i :: z'_i : t'_i$ *est vraie.*

Propriété 3.1. .
Pour tous les proportions analogiques $(x_{i'_1} : y_{i'_1} :: z_{i'_1} : t_{i'_1})$ *et* $(x_{i'_2} : y_{i'_2} :: z_{i'_2} : t_{i'_2})$ *dont les constituants* $\in \Sigma$,

1. $x_{i'_1} = x_{i'_2}$ *dans* $x \iff \begin{cases} y_{i'_1} = y_{i'_2} \text{ dans } y, \\ z_{i'_1} = z_{i'_2} \text{ dans } z, \\ t_{i'_1} = t_{i'_2} \text{ dans } t. \end{cases}$

2. $x_{i'_1}$ *est à gauche de* $x_{i'_2}$ *dans* $x \iff \begin{cases} y_{i'_1} \text{ est à gauche de } y_{i'_2} \text{ dans } y, \\ z_{i'_1} \text{ est à gauche de } z_{i'_2} \text{ dans } z, \\ t_{i'_1} \text{ est à gauche de } t_{i'_2} \text{ dans } t. \end{cases}$

3. $x_{i'_1}$ *est un ancêtre de* $x_{i'_2}$ *dans* $x \iff \begin{cases} y_{i'_1} \text{ est un ancêtre de } y_{i'_2} \text{ dans } y, \\ z_{i'_1} \text{ est un ancêtre de } z_{i'_2} \text{ dans } z, \\ t_{i'_1} \text{ est un ancêtre de } t_{i'_2} \text{ dans } t. \end{cases}$

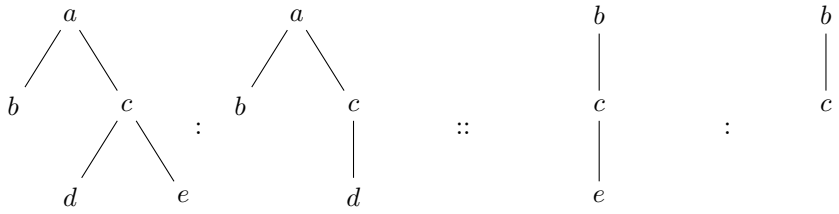
Ces conditions nécessaires non suffisantes [1] permettent la conservation de structures des arbres lors de l'alignement. En effet, le premier point indique qu'un nœud n'est concerné que par une et une seule proportion. Le point 2 détermine la conservation de l'ordre gauche-droite des nœuds lors de l'alignement des quatre arbres. Le point 3 traduit le maintien de l'ordre entre ancêtre-descendant.

3.1.1 Exemples d'arbres en proportion analogique par alignement

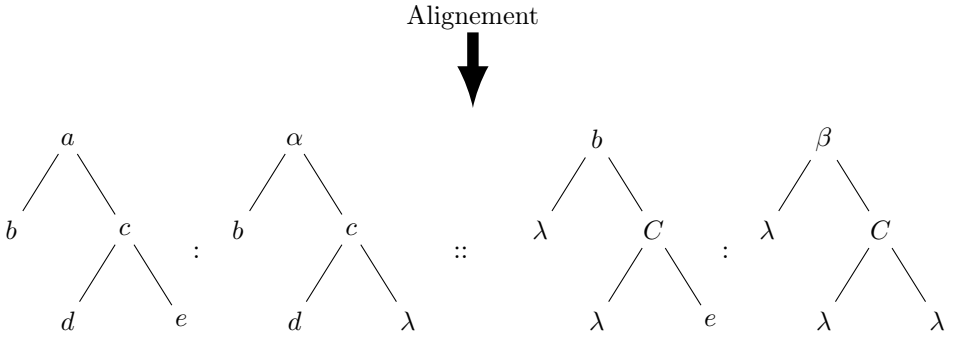
Nous reprenons ici les exemples d'arbres de la section précédente et nous appliquons notre définition d'analogie par alignement.

Exemple 1 :

Nous avons un alignement des quatre arbres tel que $\forall$i $\in Pos$, l'analogie $x'_i : y'_i :: z'_i : t'_i$ est vraie. Donc, les quatre arbres sont en proportion analogique par alignement.

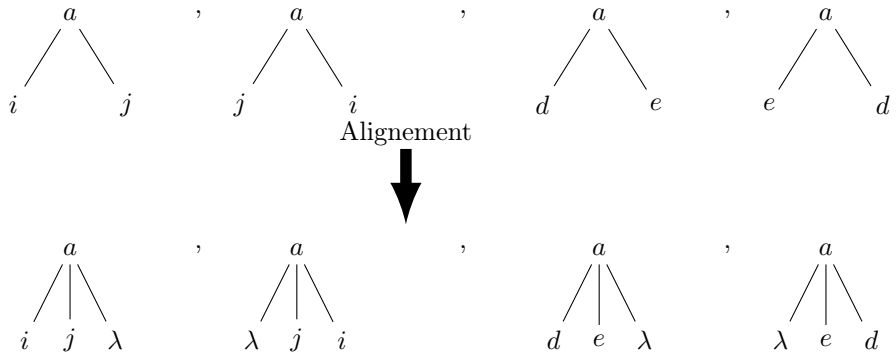


1. Une quatrième condition est en cours de vérification pour différencier l'alignement de la distance d'édition entre arbres.



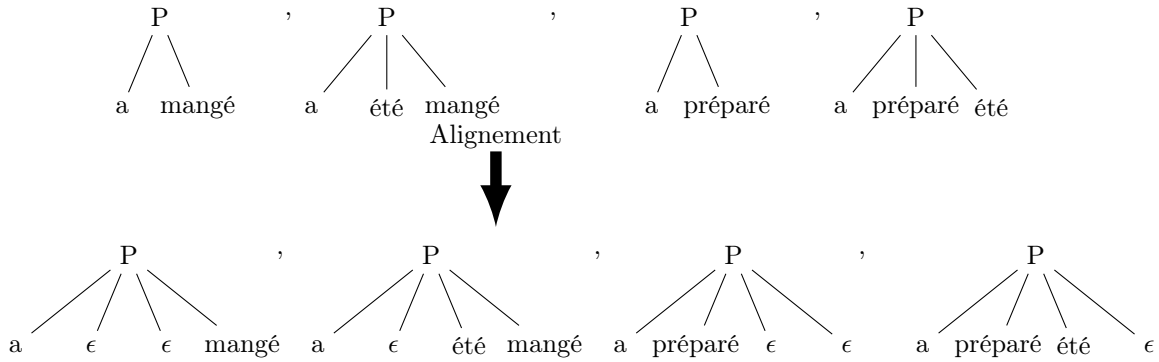
Exemple 2 :

Dans cet alignement de quatre arbres, pour certains nœuds, l'analogie $x'_i : y'_i :: z'_i : t'_i$ n'est pas vraie. Nous montrerons plus loin qu'il n'est pas possible de mettre ces quatre arbres en proportion analogique par alignement.



Exemple 3 :

Les quatre arbres sont en proportion analogique par alignement.



Retour sur l'exemple 2.

Pour l'exemple 2, les quatre arbres ne peuvent être en proportion analogique que si nous mettons en jeux les proportions suivantes sur les nœuds:

- $a : a :: a : a$,
- $i : i :: d : d$, et
- $j : j :: e : e$.

Pour notre définition, il n'est pas possible de définir ces proportions analogiques sur les nœuds parce que ils ne vérifient pas les conditions imposées par l'alignement. En effet, si i est à gauche de j dans le premier arbre donc il faut que i soit à gauche de j dans le deuxième arbre, d soit à gauche de e dans le troisième et d soit à gauche de e dans le quatrième arbre. Ce qui n'est pas le cas.

3.1.2 Calcul et validation de la proportion analogique

Pour mettre en œuvre notre définition nous proposons un algorithme polynomial, tout en introduisant une nouvelle notion : la *dissemblance analogique* entre arbres.

Dissemblance analogique Dans cette section, nous définissons une proportion analogique approximative, que pourrait traduire l'expression linguistique *"a est à b à peu prés ce que c est à d"*. Nous mesurons le terme "à peu prés" par une certaine valeur réelle positive, égale à 0 quand l'analogie est exacte, et croissante au fur et à mesure que les quatre objets sont de moins en moins en proportion. Nous appelons cette valeur *dissemblance analogique (DA)*.

Cette mesure a été introduit aux séquences par [], nous l'étendons ici aux arbres.

Dissemblance analogique entre arbres Nous supposons d'abord définie une dissemblance analogique DA sur l'alphabet Σ_λ, qui est une quantité réelle positive et telle que, $\forall\ a \in \Sigma$ et $b,c,d \in \Sigma_\lambda$:

- $(DA(a,a,b,b) = 0) \Leftrightarrow (a : b :: c : d)$,
- $DA(a,b,c,d) = DA(a,c,b,d)$,
-
- $DA(a,b,c,d) = DA(c,d,a,b)$,

Définissons maintenant la dissemblance analogique entre arbres ordonnés et étiquetés.

Définition 3.3. *Soit x, y, z et t quatre arbres d'étiquettes dans Σ.*

Le coût d'un alignement entre ces quatre arbres est défini comme la somme des dissemblances analogiques sur chacun de ses nœuds.

La dissemblance analogique $DA(x,y,z,t)$ est le coût minimal d'un alignement de ces quatre arbres.

Propriété 3.2. *Cette définition assure que les propriétés suivantes restent vraies :*
$\forall$ *x, y, z et t arbre,*

Cohérence avec l'analogie .
$DA(x,y,z,t) = 0 \Leftrightarrow x : y :: z : t.$

Symétrie de la relation *"à peu prés ce que"* et échange de moyens .
$DA(x,y,z,t) = DA(z,t,x,y) = DA(x,z,y,t).$

Non Symétrie de la relation *"est à"* .
$DA(x,y,z,t) = DA(y,x,z,t)$

Calcul de la *DA* entre arbres : algorithme *AnaTree* .
En nous inspirant de l'alignement entre deux arbres, nous proposons dans ce qui suit un algorithme de programmation dynamique appelé *AnaTree* pour construire un alignement optimal entre quatre arbres et par la suite calculer la DA entre ces arbres comme le coût minimum de cet alignement.

Nous présentons cet algorithme comme une récursion qui fait appel à des alignements de forêts partielles. Cette récursion peut s'appliquer à des forêts complètes ainsi qu'à des arbres, puisque une forêt est un ensemble ordonné d'arbres et que cet ensemble peut avoir un seul élément ($F[i_k,i_k] = T[i_k]$).

L'idée de comment aligner un nombre finis d'arbres et en particulier quatre, pour notre cas, est de raisonner sur l'étiquette de la racine la plus à droite (*rightmost*)de la forêt résultante et puis de traiter tous les cas que peut avoir cette étiquette(Figure3).

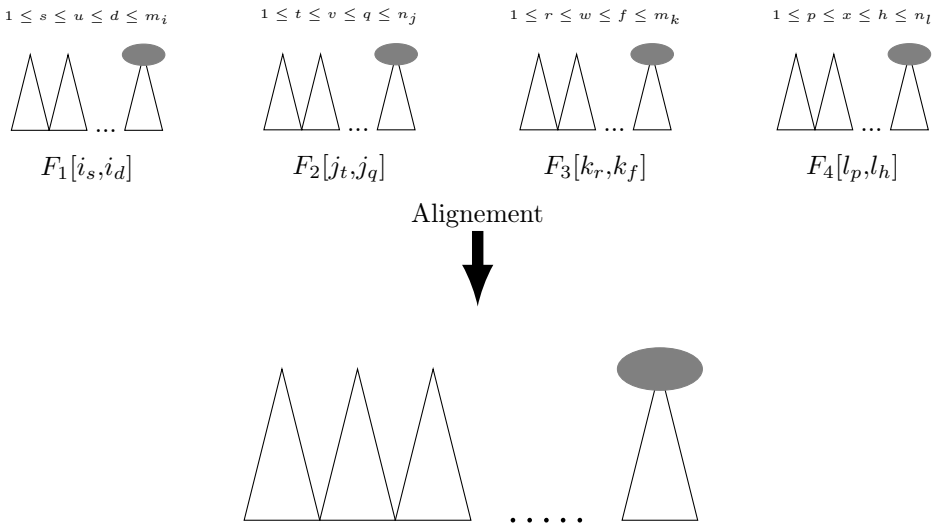


FIG. 3 – *Alignement de quatre forêts*

L'entrée de l'algorithme est l'alphabet Σ_λ dans lequel est définie une dissemblance analogique. La sortie de cet algorithme est la DA entre quatre arbres.

Le lemme suivant forme la base de notre algorithme.

Lemme 3.1.

$$\prod_{F_3[k_r,k_f]\ ,\ F_4[l_p,l_h]}^{F_1[i_s,i_d]\ ,\ F_2[j_t,j_q]} = min \begin{cases} DA(\lambda,\lambda,\lambda,l) + min_{\substack{s\le u\le d+1\\ t\le v\le q+1\\ r\le w\le f+1}} \{\prod_{F_3[k_r,k_{w-1}]\ ,\ F_4[l_p,l_{h-1}]}^{F_1[i_s,i_{u-1}]\ ,\ F_2[j_t,j_{v-1}]} + \prod_{F_3[k_w,k_f]\ ,\ F_4[l_h]}^{F_1[i_u,i_d]\ ,\ F_2[j_v,j_q]}\} \\ DA(\lambda,\lambda,k,l) + min_{\substack{s\le u\le d+1\\ t\le v\le q+1}} \{ \prod_{F_3[k_r,k_{f-1}]\ ,\ F_4[l_p,l_{h-1}]}^{F_1[i_s,i_{u-1}]\ ,\ F_2[j_t,j_{v-1}]} + \prod_{F_3[k_f]\ ,\ F_4[l_h]}^{F_1[i_u,i_d]\ ,\ F_2[j_v,j_q]}\} \\ DA(\lambda,j,k,l) + min_{s\le u\le d+1} \{ \prod_{F_3[k_r,k_{f-1}]\ ,\ F_4[l_p,l_{h-1}]}^{F_1[i_s,i_{u-1}]\ ,\ F_2[j_t,j_{q-1}]} + \prod_{F_3[k_f]\ ,\ F_4[l_h]}^{F_1[i_u,i_d]\ ,\ F_2[j_q]}\} \\ DA(i,j,k,l) + \prod_{F_3[k_r,k_{f-1}]\ ,\ F_4[l_p,l_{h-1}]}^{F_1[i_s,i_{d-1}]\ ,\ F_2[j_t,j_{q-1}]} + \prod_{F_3[k_f]\ ,\ F_4[l_h]}^{F_1[i_d]\ ,\ F_2[j_q]} \end{cases}$$

Ce lemme ne couvre pas toutes les étiquettes possibles que peut avoir le *rightmost*. Les autres cas sont similaires. Par exemple, dans le cas où nous avons traité l'étiquette (λ,λ,λ,l), il est similaire de traiter les cas des étiquettes suivantes :

– (λ,λ,k,λ), (λ,j,λ,λ), (i,λ,λ,λ).

Preuve .

Considérant un alignement optimale (forêt) A^4 de F_1, F_2, F_3 et F_4. Il y a quinze ($2^4 - 1$) étiquettes possibles du *rightmost* A^4, soit tous les combinaisons possibles à partir des *rightmost* des F_i sauf la combinaison (λ,λ,λ,λ).

Complexité Nous avons montré[2] que la complexité du calcul de la *dissemblance analogique* est de l'ordre $O(|T|^4 \times degré(T)^4)$, où |T| est la taille maximum des quatre arbres, en particulier si le *degré(T)* est borné par une constante la complexité devient de l'ordre $O(|T|^4)$.

4 Conclusion

Nous avons proposé dans ce rapport une nouvelle définition de l'analogie entre arbres qui sera à la base de l'apprentissage automatique pour la prédiction de la prosodie.

Dans un premier temps, nous avons étudié les travaux antérieurs faits sur l'analogie entre arbres, que nous avons appelée « analogie par factorisation ».

Dans un deuxième temps, nous avons proposé notre approche, une définition moins générale de l'analogie, appelée analogie par alignement, qui est naturellement adaptée à des proportions analogiques prédéfinies dans l'alphabet.

Le principe de notre démarche réside à étendre l'alignement entre deux arbres à un plus grand nombre d'arbres et de profiter des propriétés de l'alignement (la conservation

2. La complexité, la preuve du lemme et l'algorithme de l'alignement multi-arbres sont plus détaillés dans un autre document en rédaction.

de l'ordre des nœuds en verticale et en horizontale) lors de la propagation de l'analogie sur les quatre arbres en question. Cette idée a été inspiré des travaux faits sur les séquences. A la lumière de la comparaison de ces deux définitions, il apparaît pour chacune des enjeux particuliers.

- L'analogie par factorisation permet de définir des proportions analogiques entre arbres plus complexes, notamment pour prendre en compte des mouvement des nœuds d'une position à une autre comme il est décrit à l'exemple 2. Cela peut s'appliquer à des phénomènes linguistiques comme les transformations actif/passif. L'analogie par alignement est plus restrictive, à cause des conditions imposés par l'alignement sur l'ordre des nœuds.
- L'analogie par alignement s'adapte facilement à la notion de proportion analogique approximative, que pourrait traduire l'expression linguistique "a est à b à peu près ce que c est à d". Nous mesurons le terme "à peu prés" par une certaine valeur réelle positive, égale à 0 quand l'analogie est exacte, et croissante au fur et à mesure que les quatre objets sont de moins en moins en proportion. nous appelons cette valeur dissemblance analogique (DA).
- Le calcul de la dissemblance analogique avec l'analogie par alignement présente une complexité raisonnable, alors que la résolution dans l'analogie par factorisation est de nature exponentielle. C'est ce qui a amené les auteurs à en proposer des versions approchées.
- Les deux définitions peuvent s'étendre à des analogies non triviales, mais l'analogie par alignement se base naturellement sur des analogies prédéfinies sur les étiquettes.

Références

[1] Tao Jiang, Lusheng Wang, and Kaizhong Zhang. Alignment of trees - an alternative to tree edit. In *CPM '94: Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching*, pages 75–86, London, UK, 1994. Springer-Verlag.

[2] Nicolas Stroppa and François Yvon. Formal models of analogical proportions. *Technical report 2006D008, École Nationale Supérieure des Télécommunications, Paris, France.*, 2006.