

Manuel Patates

Catalina CHIRCU

Mars 2014

Table des matières

Introduction	3
1. Cas guirlande.....	5
2. Cas étoile.....	6
3. Cas des séquences multi-lingues de trois avec trois séquences multi-lingues de deux	7
4. Cas de séquences multi-lingues de trois avec une séquence multi-lingue de deux parmi les trois. .	8
5. Cas des séquences multi-lingues de trois avec deux séquences multi-lingues de deux parmi les trois.....	9
6. Cas des trois langues qui forment un chemin circulaire.....	10
Cas des séquences non représentées.....	11

Introduction

Les « patates » sont une représentation des occurrences multi-lingues qu'on peut trouver dans un ensemble de corpus. Cette représentation se trouve dans la page « caractérisation de corpus ». Pour y accéder, on va dans clapoty.vjf.cnrs.fr/contacts/outils/ et on choisit dans la liste un ou plusieurs corpus, ensuite on appuie sur le bouton « Visualiser ».

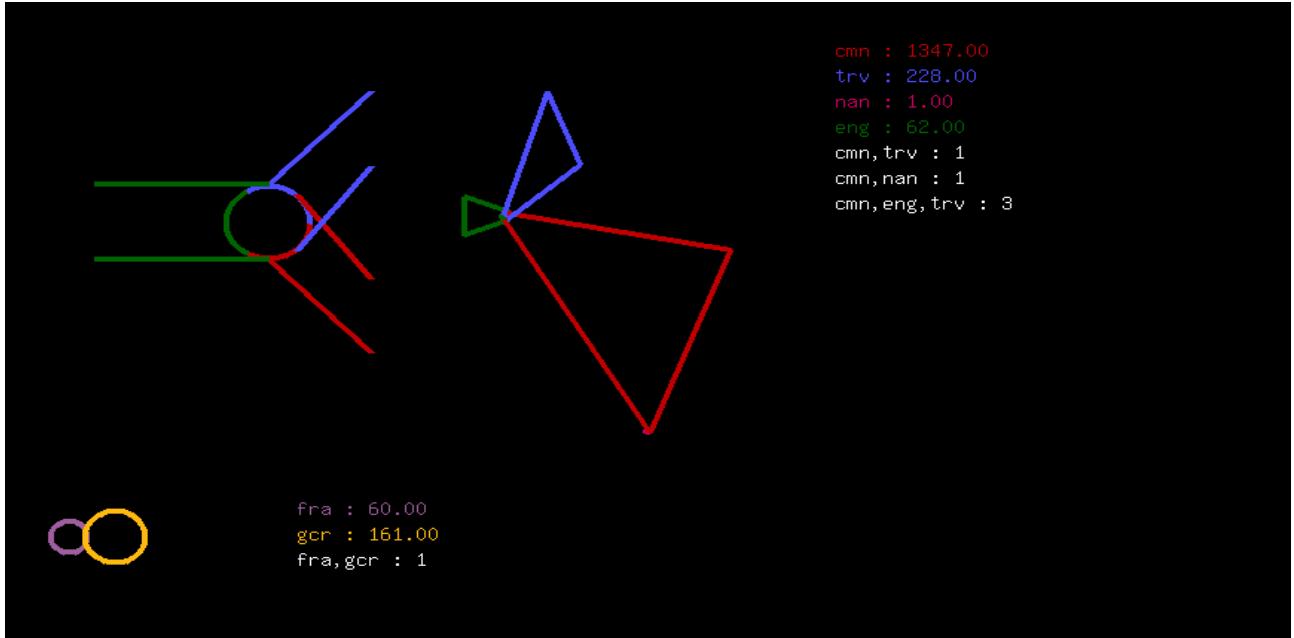
On a alors plusieurs informations sur l'ensemble de corpus choisi, et différentes statistiques, parmi lesquelles on a celle qu'on appelle « patates », qui est la dernière image et sous fond noir.

Les patates est une image qui représente les occurrences multi-lingues qu'on trouve dans l'ensemble de corpus choisis, ainsi que les langues qui composent ces séquences multi-lingues. Une langue est représentée par une forme géométrique (cercle ou trapèze) d'une couleur donnée : pour une recherche donnée, à chaque langue est associée une et une seule couleur unique et différente de la couleur des autres langues. L'aire de chaque forme (cercle ou trapèze) est proportionnelle au nombre total d'occurrences de cette langue dans l'ensemble de corpus choisis. Les cercles et les trapèzes s'intersectent, et leur intersection correspond aux occurrences multi-lingues. L'aire de l'intersection de deux ou plusieurs formes est proportionnelle au nombre d'occurrences multi-lingues pouvant appartenir à la fois à chaque langue parmi celles qui forment l'intersection (le mul).

Mis à part la représentation par des cercles ou trapèzes, dans l'image il y a une légende. Dans la légende on a, pour chaque image :

- les codes de chaque langue et le nombre total d'occurrences de la langue, dont la police est coloriée avec la couleur correspondant à la langue ;
- le type d'occurrences multi-lingues et le nombre d'occurrences multi-lingues correspondant, en police blanche ;

De manière générale, dans les patates sont représentées séparément des sous-ensembles de langues qui n'ont pas de lien entre eux. Par exemple, dans le cas suivant, l'ensemble [cmn,trv,nan,eng] n'a aucune langue commune avec l'ensemble [fra,gcr], donc on a deux groupes représentées séparément: le premier groupe est un groupe apparentant au type décrit dans le chapitre 4, tandis que le deuxième groupe appartient au type guirlande, décrit dans le chapitre 1.



En fonction des séquences multi-lingues qu'on trouve dans les corpus choisis, on peut avoir plusieurs cas de figure. Jusqu'au présent, pour les patates, sont représentés seulement une partie de ces cas. Pour les cas qui ne sont pas traités, est affichée une image avec les la légende en police blanche (voir dernier chapitre).

J'ai traité les ensembles de langues, sans lien entre eux, qui contiennent des séquences multi-lingues de deux langues seulement, ou des deux langues plus une séquence de trois langues.

Dans les chapitres suivants seront décrites les différents cas représentés.

1. Cas guirlande

Un premier cas est le cas que j'ai appelé « guirlande » car la représentation a une forme de guirlande ou de chaîne. Elle ne contient pas des chemins circulaires.

Il s'agit d'un ensemble de séquences multi-lingues de deux langues seulement, chaque langue faisant partie de 1 ou 2 types de séquence multi-lingue, maximum (et pas 3, 4, etc).

Une tel ensemble de muls sera du type :

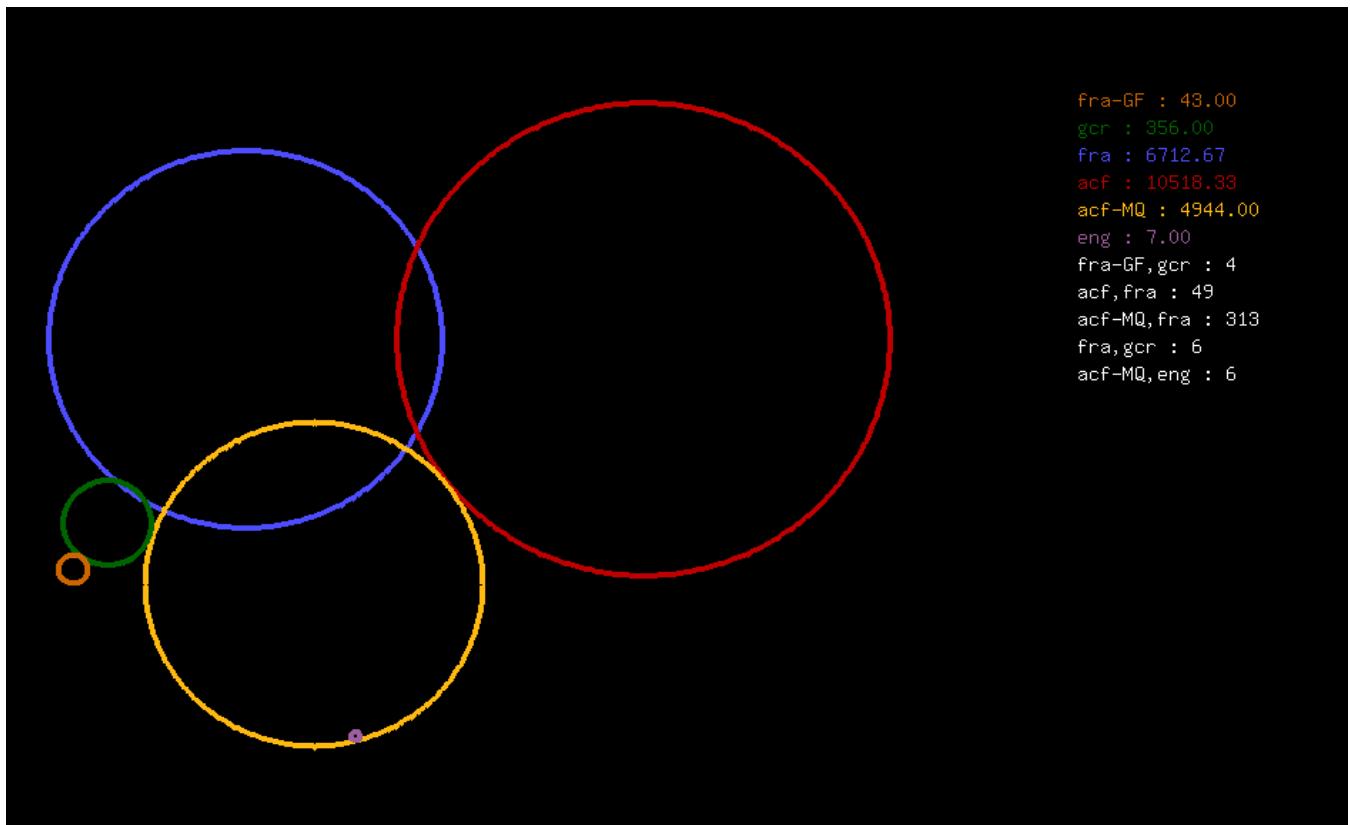
[l1 l2] [l2 l3] [l3 l4]

Dans l'image suivante, on voir que « fra » et « fra-GF » font partie chacune d'un seul type de séquence multi-lingue (fra,gcr et fra-GF,gcr) tandis que « gcr » fait partie de 2 types de séquences multi-lingues.



2. Cas étoile

Cette représentation est toujours une représentation des séquences multi-lingues de deux langues et pas plus, sans chemins circulaires, mais dont certaines langues peuvent apparaître dans plus de deux séquences multi-lingues (à la différence du cas précédent). On voit, en effet, dans l'exemple suivant, que cette image contient seulement des séquences multi-lingues de 2 langues mais que certaines langues apparaissent dans plus de deux muls (par exemple le français qui apparaît dans « acf,fra », « acf-MQ-fra » et « fra,gcr »).



3. Cas des séquences multi-lingues de trois avec trois séquences multi-lingues de deux

Il s'agit du cas :

[l1 l2 l3]
[l1 l2]
[l2 l3]
[l1 l3]

Dans ce cas, on aura un et un seul ensemble comme celui-ci, ainsi que, éventuellement, d'autres types de séquences multi-lingues composés de deux langues (sous-branches de type guirlande ou étoile).



4. Cas de séquences multi-lingues de trois avec une séquence multi-lingue de deux parmi les trois

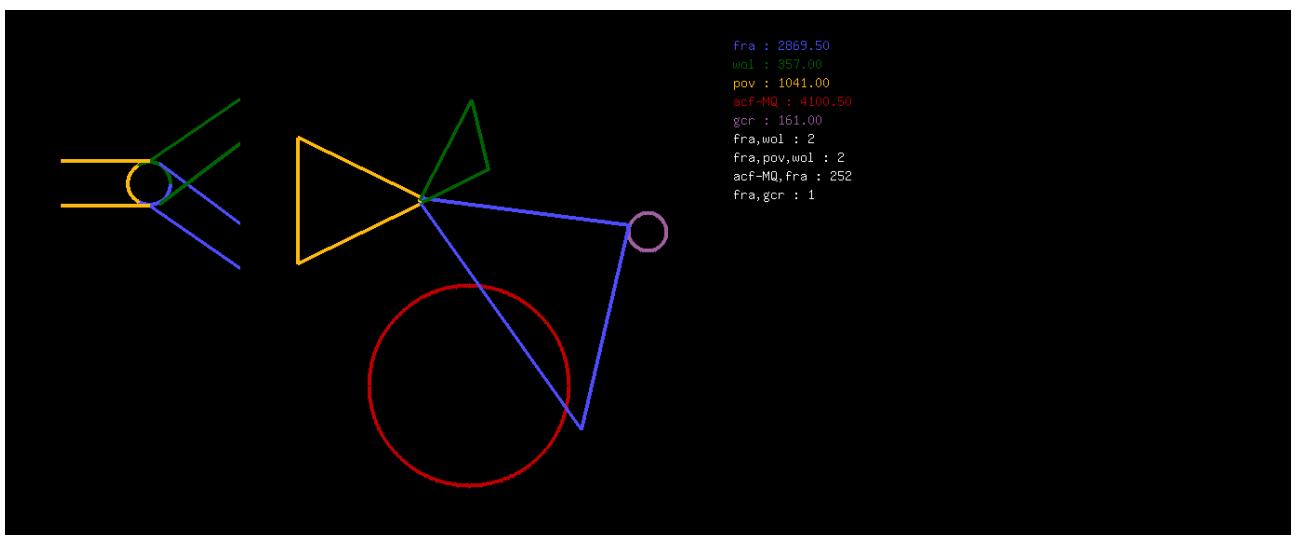
Il s'agit du cas :

[l1 l2 l3]
[l1 l2]

mais pas :

[l2 l3]
[l1 l3]

Dans ce cas, on aura un et un seul ensemble comme celui-ci, ainsi que, éventuellement, d'autres types de séquence multi-lingue composés de deux langues (sous-branches de type guirlande).



Dans ce cas, on a deux images : la deuxième image représente les langues l1, l2 et l3 sous forme de trapèzes, et les autres langues sous forme de cercles.

La première image représente les muls [l1 l2 l3] et [l1 l2], très agrandies. Le cercle représente le mul [l1 l2 l3], tandis que la forme délimité par une ligne bleue, une ligne verte et un arc de cercle, représente la séquence [l1 l2]. Dans la deuxième image, cette partie se trouve au niveau de la rencontre des trois trapèzes vert, bleu et jaune.

5. Cas des séquences multi-lingues de trois avec deux séquences multi-lingues de deux parmi les trois

Il s'agit du cas :

[l1 l2 l3]

[l1 l2]

[l2 l3]

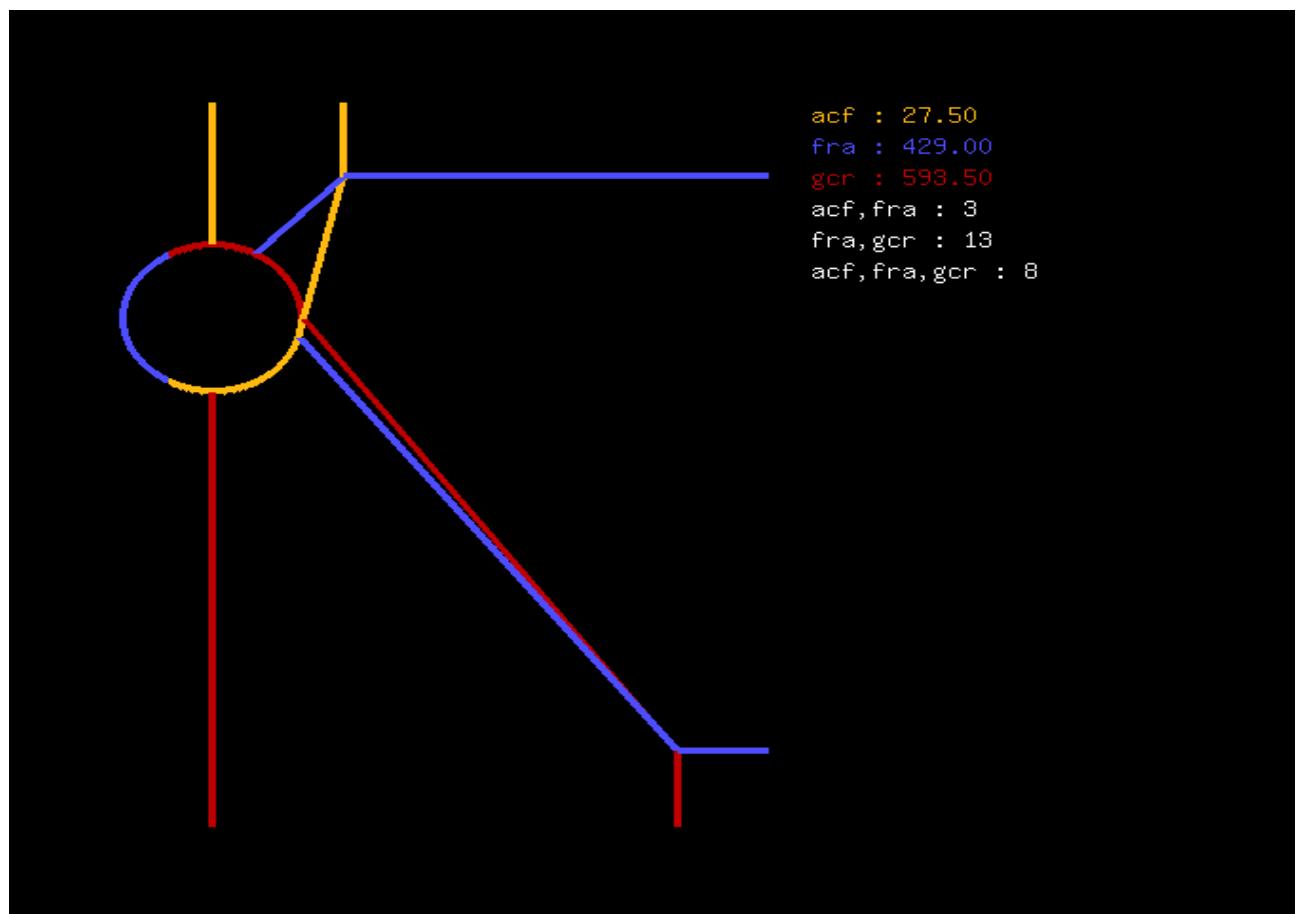
mais pas :

[l1 l3]

Dans ce cas, on aura un et un seul ensemble comme celui-ci, ainsi que, éventuellement, d'autres types de séquences multi-lingues de deux langues (sous-branches de type guirlande).

On a ici la même structure que précédemment.

En revanche, dans ce cas, les autres séquences multi-lingues ne seront pas représentées, et on aura affichée seulement l'image agrandies qui représente les séquences [l1 l2 l3], [l1 l2], [l2 l3], ainsi que la légende.



6. Cas des trois langues qui forment un chemin circulaire

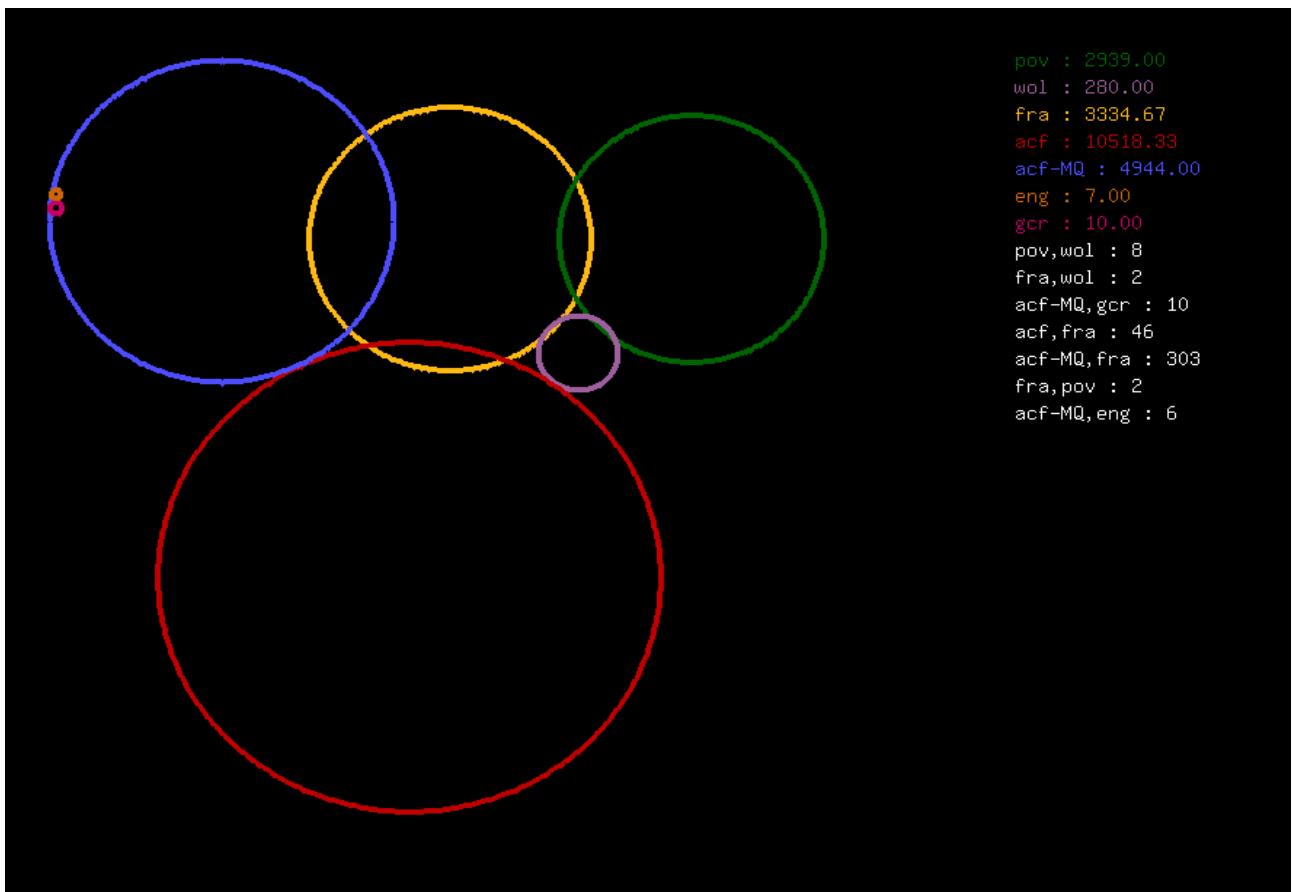
Il s'agit d'un cas où on a une structure centrale du type :

[l1 l2]
[l2 l3]
[l1 l3]

et en plus d'autres liens avec d'autres langues, mais toujours des séquences multi-lingues de deux langues et pas plus.

Dans ce cas il y aura un et un seul chemin circulaire. Il est représenté par trois cercles intersectés deux par deux.

Dans l'exemple suivant les trois langues qui forment un chemin circulaire sont fra, pov et wol (vert, violet et jaune). On remarque que à ce « noyau » de trois langues on peut attacher des « branches » du type « guirlande » (« fra,acf ») où « étoile ». (« fra,acf-MQ », « acf-MQ,gcr », « acf-MQ,eng »).



Cas des séquences non représentées

Dans les patates on a pas pour le moment la représentation des ensembles de corpus contenant des séquences multi-lingues avec plus de trois langues, ou avec deux ou plus des séquences multi-lingues de trois langues. On n'aura pas non plus la représentation des séquences multi-lingues de deux langues qui contiennent des chemins circulaires.

Je n'ai pas encore traité non plus le cas où on a : [l1 l2 l3] mais pas :

[l1 l2]
[l2 l3]
[l1 l3]

Dans le cas où pour un ensemble de corpus donné, on n'aura pas de représentation des patates, dans l'image on aura seulement la légende des séquences multi-lingues (ceci est l'image « Patates » pour le corpus d'Alice).

```
cgr,fra,fra-GF -- 1
cpf-019,fra,gcr -- 20
acf-MQ,fra -- 1
acf,fra -- 73
fra,srm -- 11
acf,cpf-019,fra,gcr,hat -- 2
fra,gcr -- 63
acf,cpf-019,fra,gcr -- 4
acf,cpf-019,fra,gcr,srm -- 6
acf-GP,cpf-019,fra,gcr -- 3
cpf-019,fra -- 123
```