L'exemple de Nicolas Pécheux

Je propose les mots 'la', 'Présidente' et 'voudrais' :

Sur les 100,000 premiers discours prononcés au parlement Européen on a :

'la' < 'Présidente' [98%] ; 'Présidente' < 'voudrais' [91%] ; 'voudrais' < 'la' [78%]

on comprend intuitivement pourquoi:

- "Madame la Présidente, je voudrais" commence très souvent une intervention;

- mais "Je voudrais" est encore plus fréquent en début de phrase, sans avoir à parler ensuite de la Présidente, mais en ayant recours au déterminant.

Si on fait une recherche exhaustive sur les 100,000 premiers discours, en ne conservant que les mots de fréquence supérieure à 1000, les couples de mots co-occurrent au moins 100 fois dans un même discours, et en imposant que mot1 < mot2 ssi mot 1 apparaît avant mot2 dans au moins 75% des cas, alors on a seulement deux cas. Le précédent et celui-ci :

'nous' < 'nos' [77%] ; 'nos' < 'collègues' [77%] ; 'collègues' < 'nous' [83%]

Si on est un peu plus laxiste sur le seuil (par exemple 51% des cas), on trouve 11093 triplets possibles.

Je n'ai pas testé sur l'ensemble du corpus (2M discours) mais j'ai fait petit script qui permet de calculer tout ça pour un corpus de textes quelconque si cela intéresse quelqu'un.

Nicolas Pécheux