Les fichiers ***results_phrases_DIST.ods*** et ***results_phrases_PROBA.ods*** contiennent les résultats de obtenus par le script lorsque j'ai lancé le script pour 43 phrases (une pour chaque langue du projet). Ces exemples de langues ont été extraits de la bible. Il s'agit de la dernière phrase du premier chapitre de la bible. Ce choix a été pris comme une forme d'assurer que toutes les phrases aient une longueur cohérente.

Calcul de Distance

- **1 Ngram** 1.100.000 p=1 s=0 d=0 (97% langues trouvées correctement - (marge moyenne 26,8%)
- **3 Ngram** 500.000 p=1 s=0 d=0 (100% langues trouvées correctement - (marge moyenne 778%)
- **5 Ngram** 500.000 p=0 s=0 d=0 (100% langues trouvées correctement - (marge moyenne 91,4%)

Somme de Probabilités

- **1 Ngram** 500.000 p=0 s=0 d=0 (62% langues trouvées correctement - (marge moyenne 34,3%)
- **3 Ngram** 1.100.000 p=0 s=0 d=0 (97% langues trouvées correctement - (marge moyenne 97,6%)
- **5 Ngram** 1.100.000 p=0 s=1 d=0 (100% langues trouvées correctement - (marge moyenne 75,9%)

La marge est calculée comme de la manière suivante :

$$\frac{\text{Scores des langues position 1 - Scores de langue position 2}}{\text{Scores de langues position 1}}$$

Comme nous pouvons constater les meilleurs résultats obtenus par le système utilisent la configuration suivante :

Méthode Calcul de Distance - 3 Ngram 500.000 p=1 s=0 d=0

Commentaires sur les résultats :

Les résultats trop élevés par le système peuvent être expliqués pour deux raisons différentes. La première est le plus évident est dû au fait que la phrase choisie pour réaliser les tests est inadéquate. Car le texte biblique possède des caractéristiques textuelles très marquées, et cela est vrai pour l'intégralité du texte. En outre, la phrase choisie contient deux entités nommées qui sont très utilisées parmi les autres chapitre. Voici la phrase utilisée par le test :

"***Joseph*** mourut, âgé de cent dix ans. On l'embauma, et on le mit dans un cercueil en ***Égypte.***"

Pour obtenir de résultat plus précis, il faudrait refaire les tests avec une autre phrase extraite du corpus Wikipedia, par exemple.

Une explication pour les bons résultats en termes de marge est dû au fait de la diversité de langues qui composent les corpus de travail. Dans le corpus nous avons de langues qui sont les seules à utiliser certains alphabets (comme le hébreu ou le thaï, par exemple). Ainsi, lorsque nous effectuons les calculs le score de langue en deuxième position dans rang possède une différence trop grande par rapport au score de la première.

Ex. :

Langues Position 1	Score	Langues Position 2	Score	Différence
Hébreu	245.865	afrikaner	9.725.262	9.479.397
Thaï	2.031.233	Danois	10.835.908	8.804.675

Alors que pour les langues qui utilisent le même alphabet, on obtient un marge plus baisse :

Langues Position 1	Score	Langues Position 2	Score	Différence
Néerlandais	58.119	Danois	1.113.565	1.055.446
Anglais	66.660	Français	335.992	269.332