Détection de la cohésion lexicale par voisinage distributionnel : application à la segmentation thématique

Clémentine Adam & François Morlane-Hondère

CLLE - Université de Toulouse

25 juin 2009

Plan

Plan

Analyse du discours et cohésion lexicale

- L'analyse du discours bénéficie actuellement d'un regain d'intérêt dû à ses enjeux pour le TAL
- Elle repose sur l'observation selon laquelle un texte n'est pas une simple succession de phrases, mais un tout cohérent. Cette cohérence est reflétée par les marques de cohésion.
- Procédés cohésifs classiquement considérés (Halliday & Hasan, 1976) : référence, substitution, ellipse, conjonction et cohésion lexicale
- La cohésion lexicale est vue comme "the dominant mode of creating texture" (Hoey, 1991)
- Mais elle reste peu exploitée sur le plan applicatif car elle est difficile à appréhender

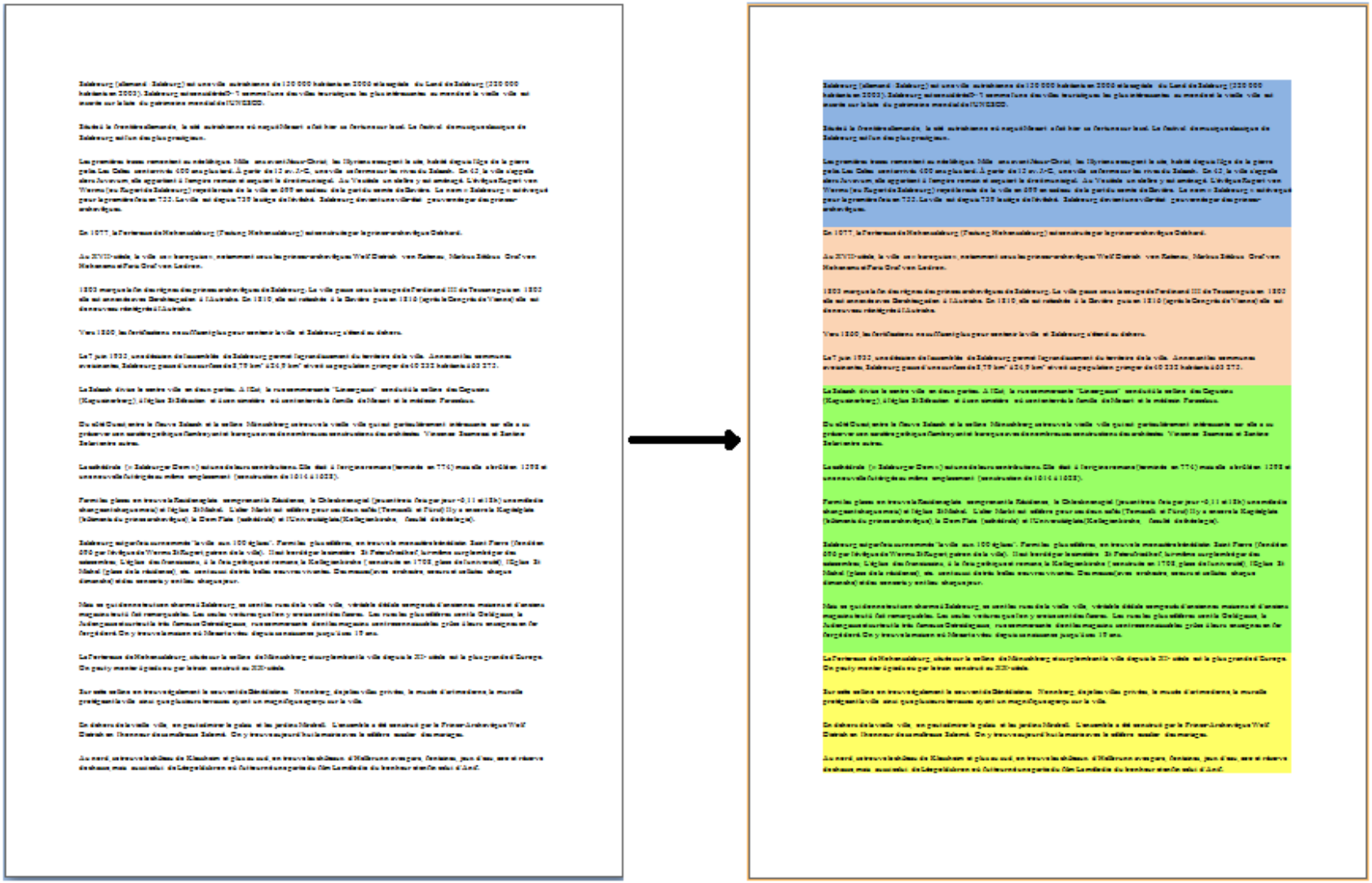
Le projet VOILADIS

- VOILADIS (VOIsinage Lexical pour l'Analyse du DIScours) : Projet du PRES Toulouse coordonné par C. Fabre impliquant des chercheurs des laboratoires CLLE-ERSS et IRIT
- Objectif : utiliser des indices lexicaux pour la mise au jour de phénomènes discursifs
- Ressource mobilisée : une base de *voisins distributionnels*
- Première étape : application à la segmentation thématique, qui a tout particulièrement exploité les indices de nature lexicale

Plan

Principes de la segmentation thématique

Objectif : segmenter un texte en blocs thématiques



Principes de la segmentation thématique

- Cette tâche est parmi les plus tributaires des phénomènes de cohésion
- De nombreux algorithmes, que l'on peut regrouper en deux familles (Hernandez, 2004) :
  (a) ceux qui parcourent linéairement le texte selon une fenêtre d'observation glissante, et procèdent donc de manière ascendante Ex. *TextTiling* de Hearst (1997)
  (b) ceux qui calculent une matrice de similarité pour l'ensemble des unités du texte avant de décider où placer les ruptures, procédant donc de manière descendante Ex. *C99* de Choi (2000)
- Pour mesurer la "force" de la cohésion lexicale entre deux pans de texte, on se base sur le nombre de liens qu'entretiennent leurs unités lexicales.

Panorama des ressources utilisées

Ressources utilisées pour repérer les liens lexicaux participant à la cohésion des textes :

- Pas de ressource : les seuls liens considérés sont les répétitions de formes, de formes tronquées ou de lemmes (Hearst, 1997; Choi, 2000), voire les répétitions de n-grammes (Beeferman *et al.*, 1997)
- Utilisation d'une ressource générique, construite à partir d'un dictionnaire ou d'un thésaurus (Kozima, 1993; Lin *et al.*, 2004; Morris & Hirst, 2004) : la synonymie, et éventuellement d'autres relations classiques comme l'hyperonymie ou l'antonymie, entrent alors en jeu.
- Utilisation d'une ressource construite en corpus, par l'extraction de collocations ou de cooccurences (Choi *et al.*, 2001; Ferret, 2002).

Les auteurs utilisant ce dernier type de ressource font généralement état de meilleures performances.

Relations mises en jeu dans la cohésion lexicale

- Les relations les plus pertinentes pour le repérage des structures discursives sont dans la plupart des cas des relations échappant aux typologies traditionnelles (Morris & Hirst, 2004)
- Lorsqu'il s'agit d'interpréter un texte, les relations comme la synonymie, l'antonymie, etc. cèdent le pas à des relations *non classiques*
- Notre ressource, construite en corpus par l'analyse distributionnelle, est à même de capter des relations échappant aux ressources traditionnelles, et repose sur des informations linguistiques plus riches qu'une simple extraction de collocations

Plan

Construction de la ressource

- La base de *voisins distributionnels* que nous utilisons a été construite à partir d'un corpus constitué de l'ensemble des articles de la version francophone de Wikipédia (version d'avril 2007), soit plus de 470 000 articles pour 194 millions de mots.
- La chaîne de traitement qui aboutit à cette base est la chaîne Syntex (analyse syntaxique) - Upéry (analyse distributionnelle), développée par D. Bourigault et adaptée à Wikipédia par F. Sajous.

Construction de la ressource

Le Kilimandjaro ou Kilimanjaro est une montagne située au nord-est de la Tanzanie et composée de trois volcans éteints : le Shira à l'ouest, culminant à 3962 mètres d'altitude, le Mawenzi à l'est, s'élevant à 5149 mètres d'altitude et le Kibo, le plus récent géologiquement, situé entre les deux autres et dont le pic Uhuru à 5891,8 mètres d'altitude constitue le point culminant de l'Afrique. Outre cette caractéristique, le Kilimanjaro est connu pour sa calotte glaciaire sommitale en phase de retrait accéléré depuis le début du XXe siècle et qui devrait disparaître totalement d'ici 2020 à 2050. La baisse des précipitations neigeuses qui en est responsable est souvent attribuée au réchauffement climatique mais la déforestation est également un facteur majeur. Ainsi, malgré la création du parc national en 1973 et alors même qu'elle joue un rôle essentiel dans la régulation bioclimatique du cycle de l'eau, la ceinture forestière continue à se resserrer.

En effet, la montagne est notamment le berceau des pasteurs masaï au nord et à l'ouest qui ont besoin de prairies d'altitude pour faire paître leurs troupeaux et des cultivateurs wachagga au sud t à l'est qui cultivent des parcelles toujours plus étendues sur les piémonts, malgré uneprise de conscience depuis le début du XXIe siècle.

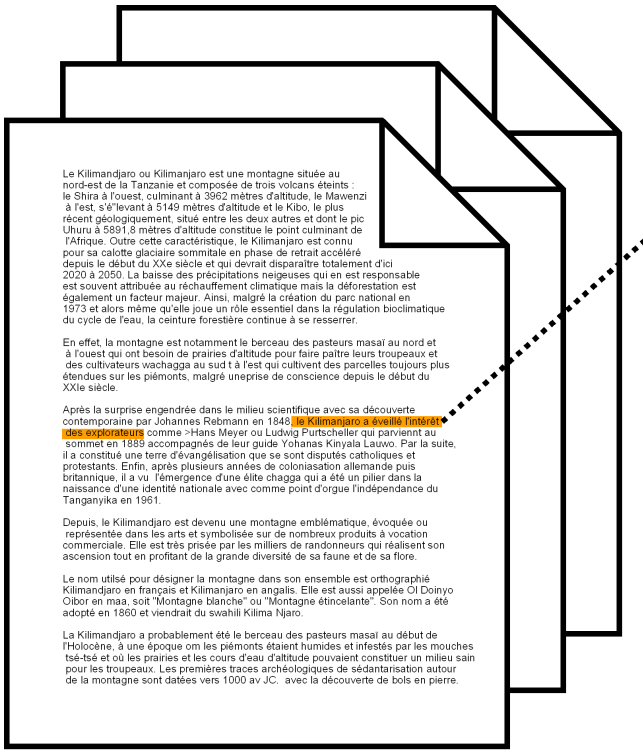
Après la surprise engendrée dans le milieu scientifique avec sa découverte contemporaine par Johannes Rebmann en 1848, le Kilimanjaro a éveillé l'intérêt des explorateurs comme >Hans Meyer ou Ludwig Purtscheller qui parviennt au sommet en 1889 accompagnés de leur guide Yohanas Kinyala Lauwo. Par la suite, il a constitué une terre d'évangélisation que se sont disputés catholiques et protestants. Enfin, après plusieurs années de colonisation allemande puis britannique, il a vu l'émergence d'une élite chagga qui a été un pilier dans la naissance d'une identité nationale avec comme point d'orgue l'indépendance du Tanganyika en 1961.

Depuis, le Kilimandjaro est devenu une montagne emblématique, évoquée ou représentée dans les arts et symbolisée sur de nombreux produits à vocation commerciale. Elle est très prisée par les milliers de randonneurs qui réalisent son ascension tout en profitant de la grande diversité de sa faune et de sa flore.

Le nom utilisé pour désigner la montagne dans son ensemble est orthographié Kilimandjaro en français et Kilimanjaro en angalis. Elle est aussi appelée Ol Doinyo Oibor en maa, soit "Montagne blanche" ou "Montagne étincelante". Son nom a été adopté en 1860 et viendrait du swahili Kilima Njaro.

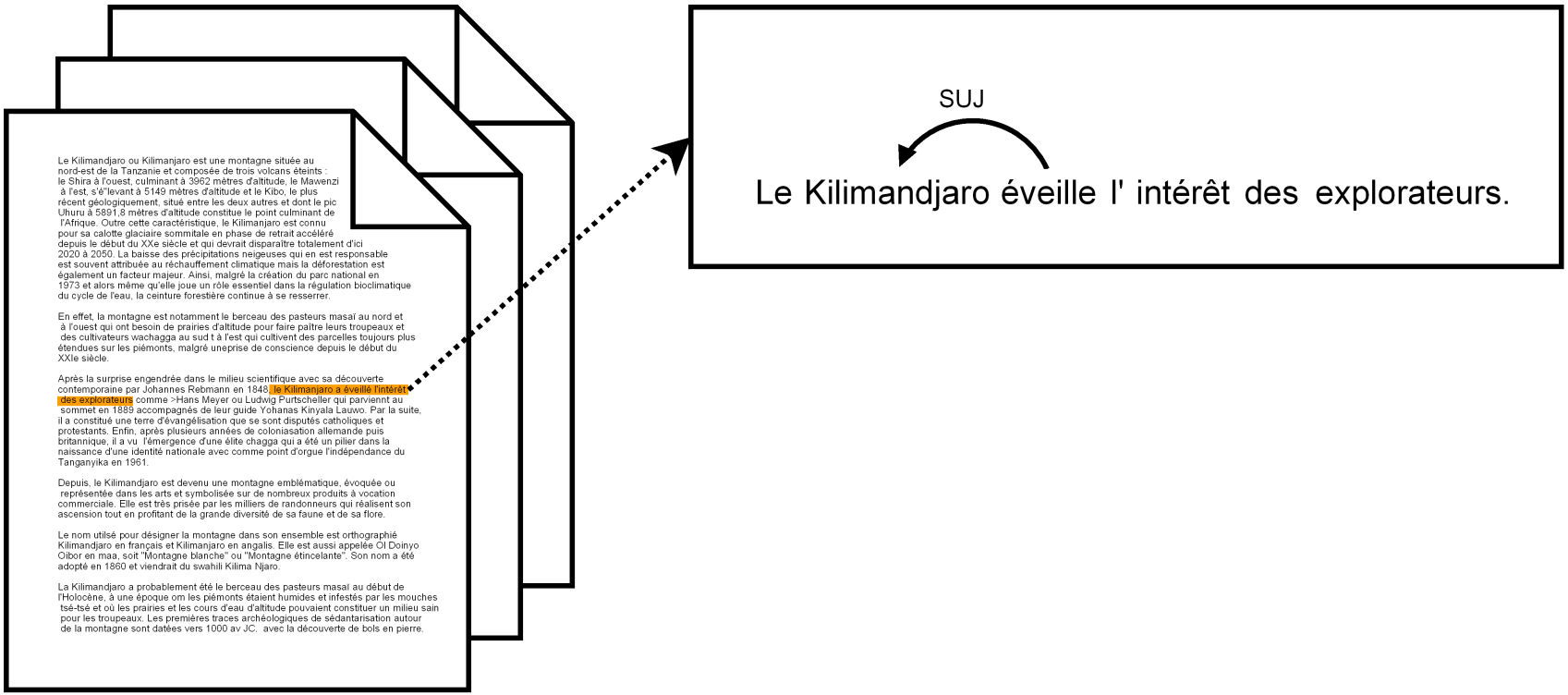
La Kilimandjaro a probablement été le berceau des pasteurs masaï au début de l'Holocène, à une époque om les piémonts étaient humides et infestés par les mouches tsé-tsé et où les prairies et les cours d'eau d'altitude pouvaient constituer un milieu sain pour les troupeaux. Les premières traces archéologiques de sédantarisation autour de la montagne sont datées vers 1000 av JC. avec la découverte de bols en pierre.

Construction de la ressource

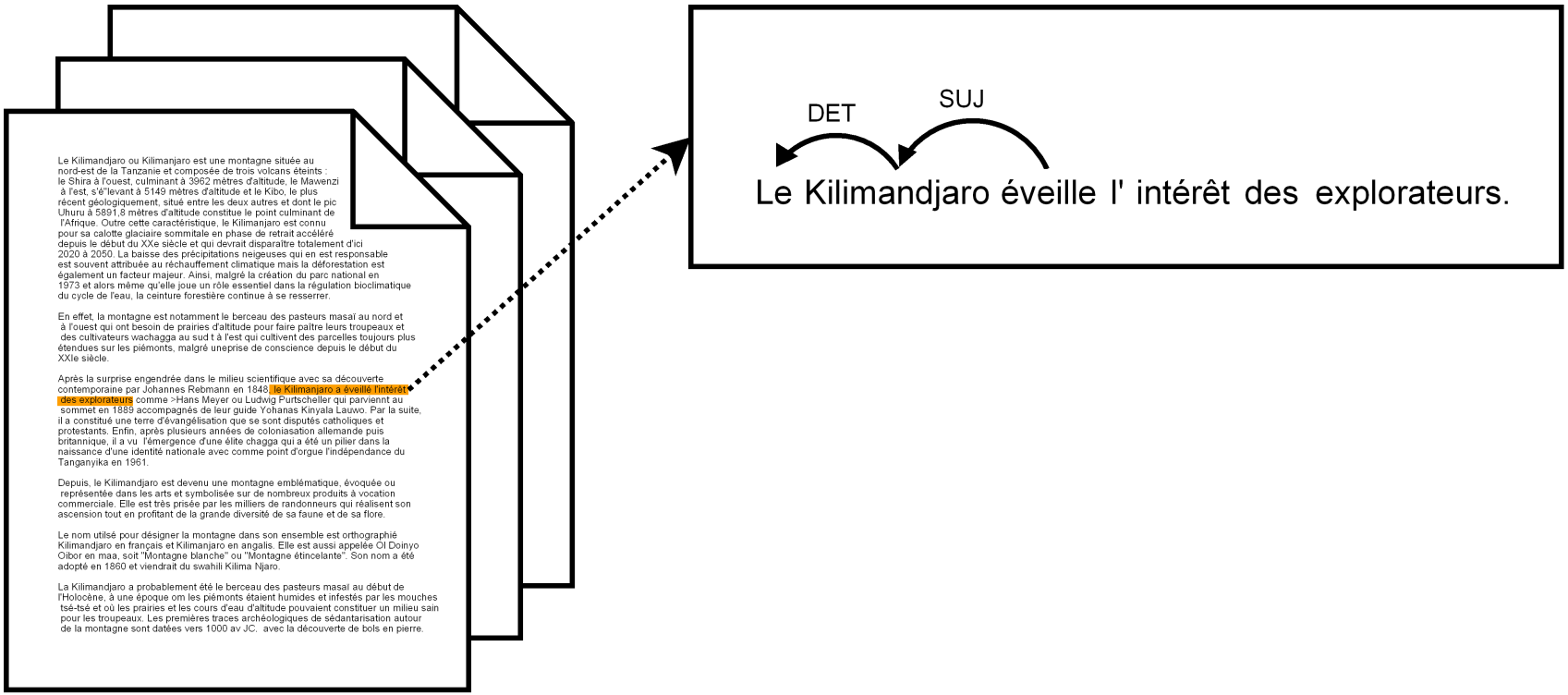


Le Kilimandjaro éveille l' intérêt des explorateurs.

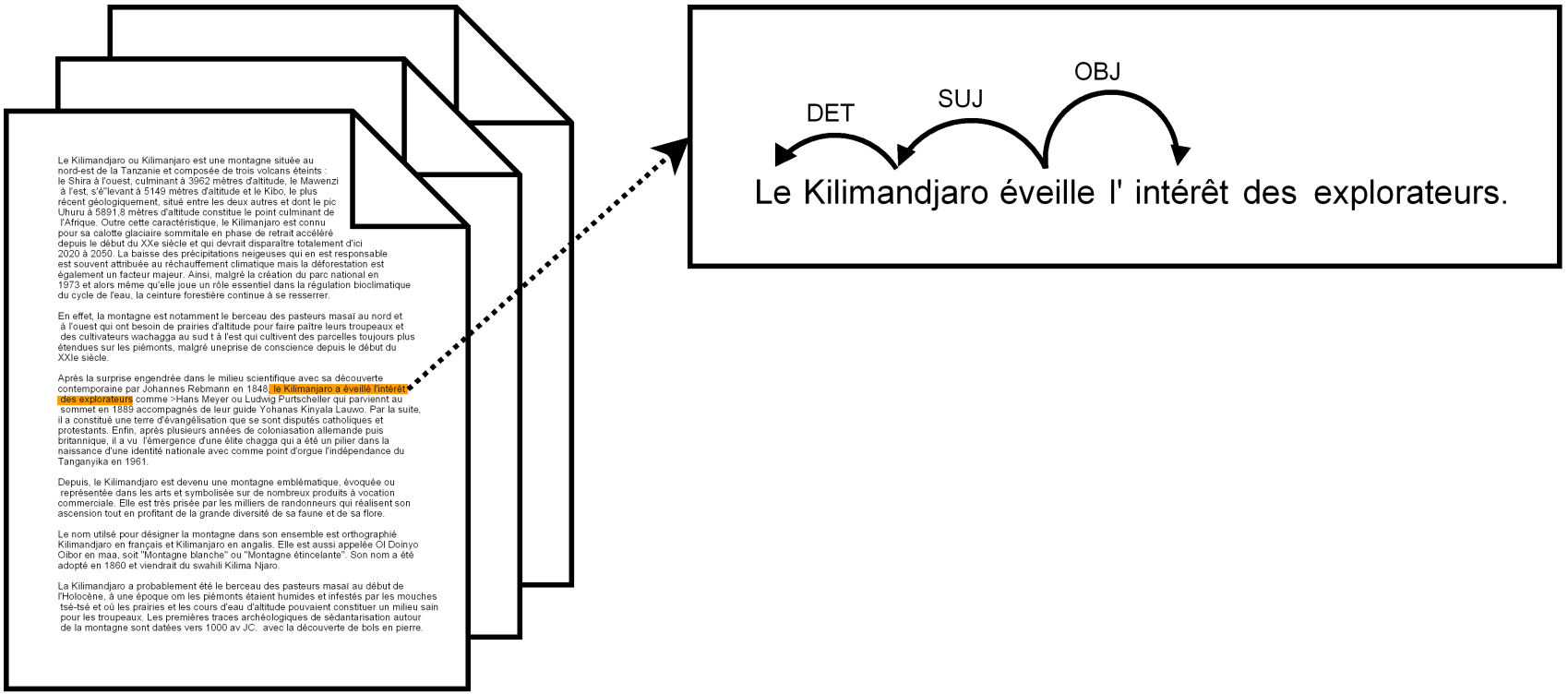
Construction de la ressource



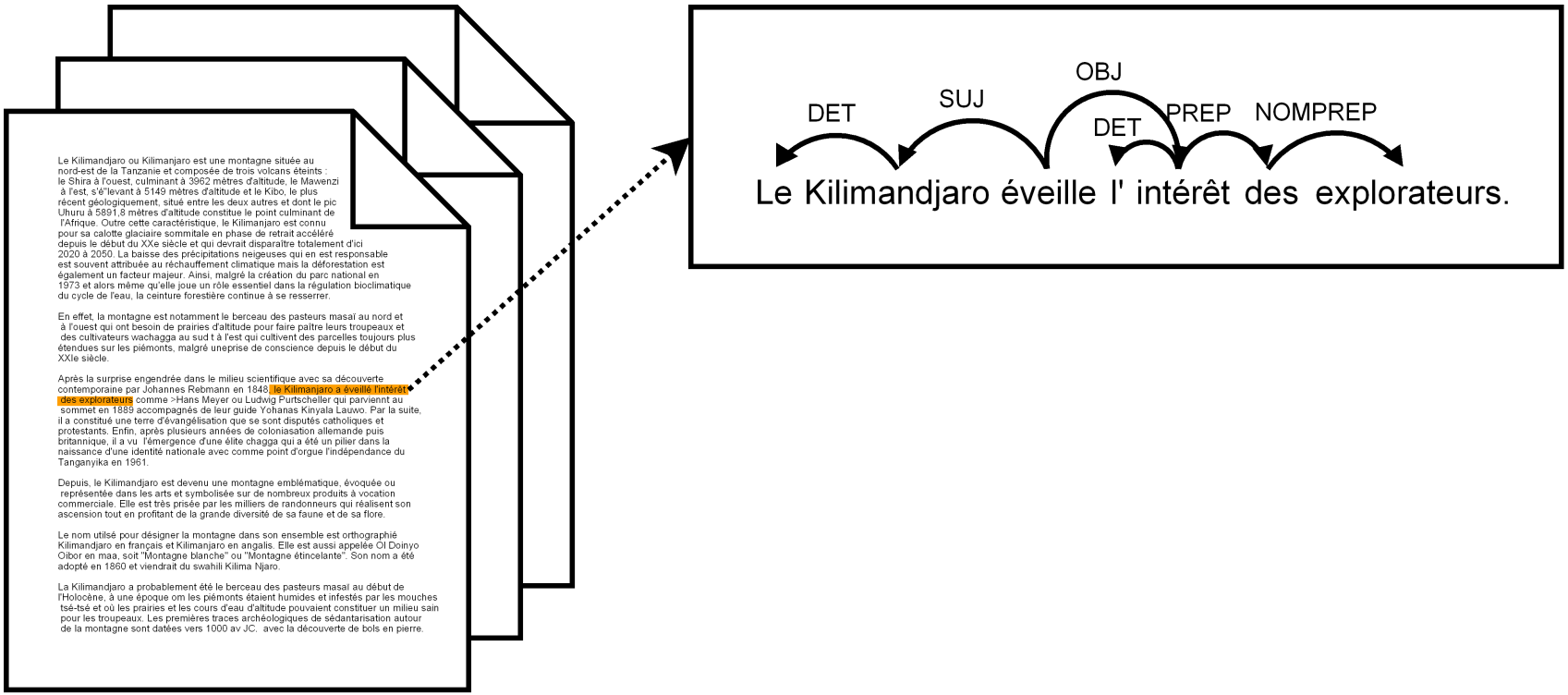
Construction de la ressource



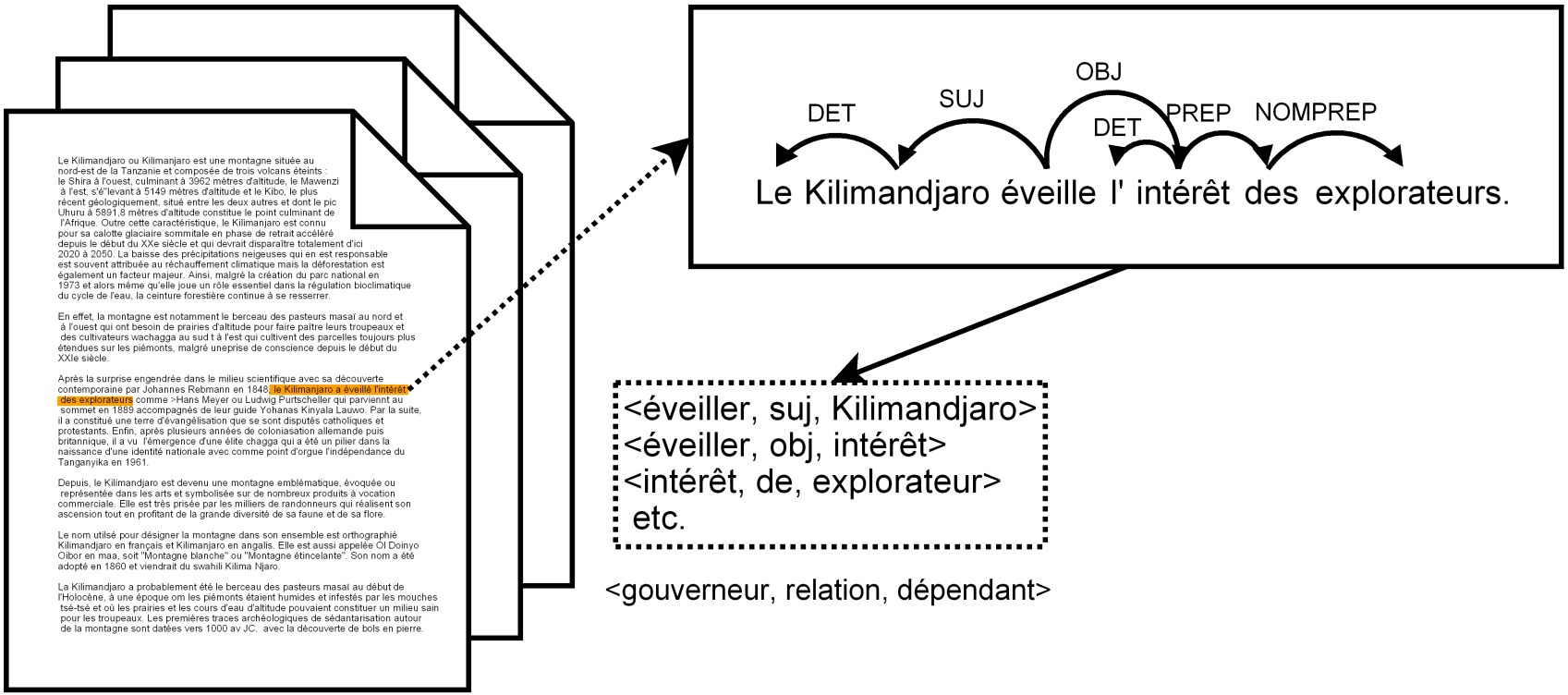
Construction de la ressource



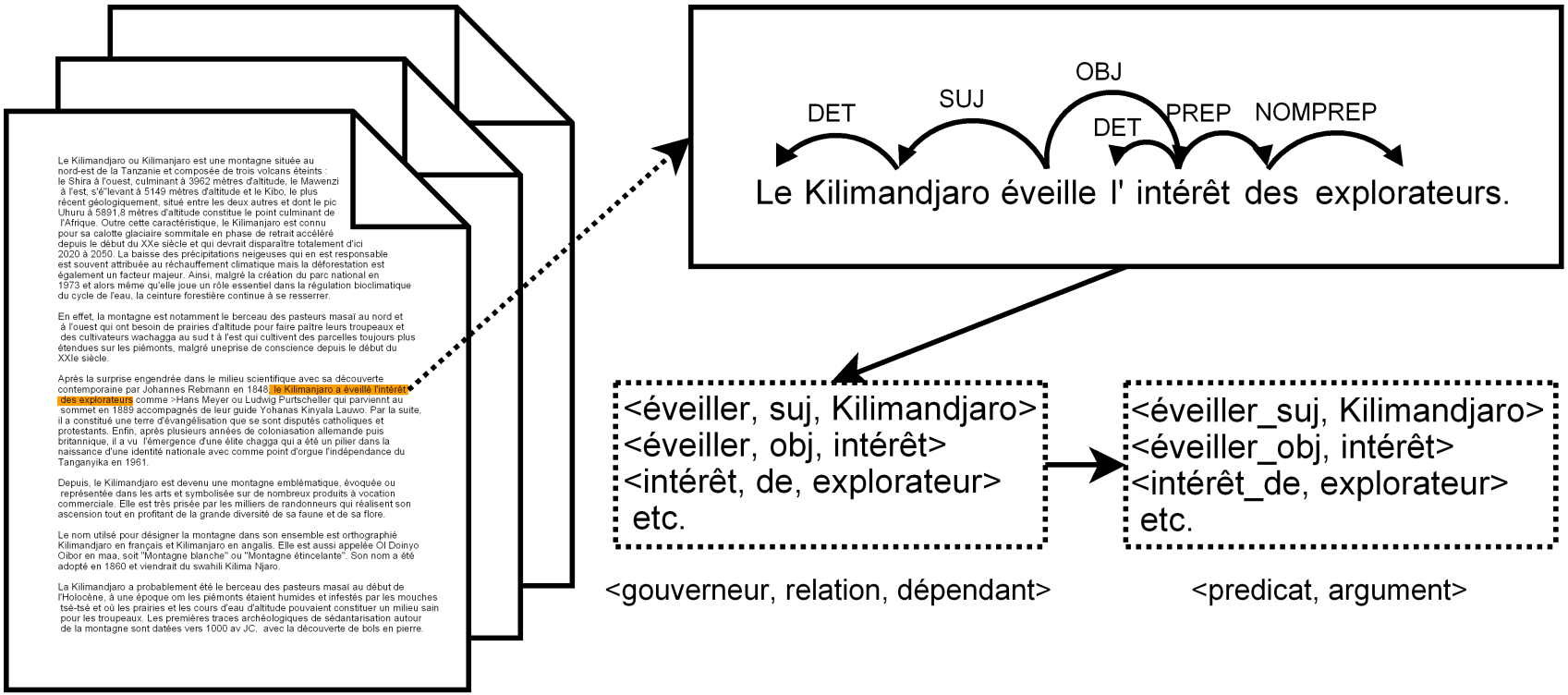
Construction de la ressource



Construction de la ressource



Construction de la ressource



Construction de la ressource

Autres couples <prédicat, argument> extraits dans le corpus :

Construction de la ressource

Autres couples <prédicat, argument> extraits dans le corpus :

<éveiller_obj, soupçon>
<éveiller_obj, curiosité>
<éveiller_obj, conscience>
<éveiller_obj, attention>
etc.

Construction de la ressource

Autres couples <prédicat, argument> extraits dans le corpus :

<éveiller_obj, soupçon>
<éveiller_obj, curiosité>
<éveiller_obj, conscience>
<éveiller_obj, attention>
etc.
<taux_de, intérêt>
<défendre_obj, intérêt>
<groupement_de, intérêt>
<servir_obj, intérêt>
<porter_obj, intérêt>
etc.

Construction de la ressource

Autres couples <prédicat, argument> extraits dans le corpus :

<éveiller_obj, soupçon>
<éveiller_obj, curiosité>
<éveiller_obj, conscience>
<éveiller_obj, attention>
etc.
<taux_de, intérêt>
<défendre_obj, intérêt>
<groupement_de, intérêt>
<servir_obj, intérêt>
<porter_obj, intérêt>
etc.

La similarité des distributions est évaluée grâce au score de *Lin*. Les mots ainsi rapprochés sont nommés *voisins distributionnels*

Construction de la ressource

Relations de voisinage distributionnel établies par Upéry :

`éveiller_obj / exciter_obj`

Contextes partagés

curiosité
convoitise
appétit
imagination
désir
passion
esprit
intérêt

Construction de la ressource

Relations de voisinage distributionnel établies par Upéry :

éveiller_obj / exciter_obj
éveiller_obj / raviver_obj

Contextes partagés

soupçon
nostalgie
crainte
souvenir
désir
inquiétude
sentiment
espoir
passion
intérêt

Construction de la ressource

Relations de voisinage distributionnel établies par Upéry :

éveiller_obj / exciter_obj
éveiller_obj / raviver_obj
éveiller_obj / endormir_obj

Contextes partagés

méfiance
vigilance
conscience
enfant
homme

Construction de la ressource

Relations de voisinage distributionnel établies par Upéry :

éveiller_obj / exciter_obj
éveiller_obj / raviver_obj
éveiller_obj / endormir_obj
intérêt / importance

Contextes partagés

juger_sans
attacher_obj
revêtir_obj
se mesurer_suj
prendre conscience_de
attester_obj
etc.

Description de la ressource

- La base obtenue pour l'ensemble de l'encyclopédie Wikipédia compte environ 4 millions de couples
- Les couples extraits couvrent un large éventail de relations de proximité sémantique, qu'il est difficile de typologiser (Fabre & Bourigault, 2006)
- Il est cependant nécessaire d'introduire des seuils lors des calculs de similarité, afin de filtrer cette ressource extrêmement pléthorique

Plan

Objectif de l'expérience

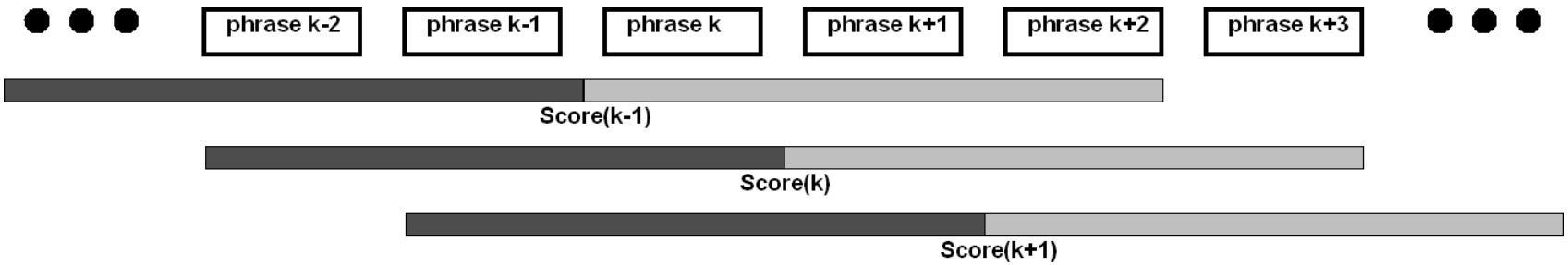
- Le but de cette expérience est de montrer la pertinence du voisinage distributionnel pour détecter les liens de cohésion lexicale, en s'appuyant sur les résultats d'un système de segmentation thématique.
- Pour cela, nous soumettons un corpus à un système de segmentation thématique basé uniquement sur la prise en compte de liens lexicaux, en spécifiant successivement différents liens :
  (a) des liens de répétition lexicale
  (b) des liens de synonymie
  (c) des liens de voisinage distributionnel

Caractérisation du corpus utilisé

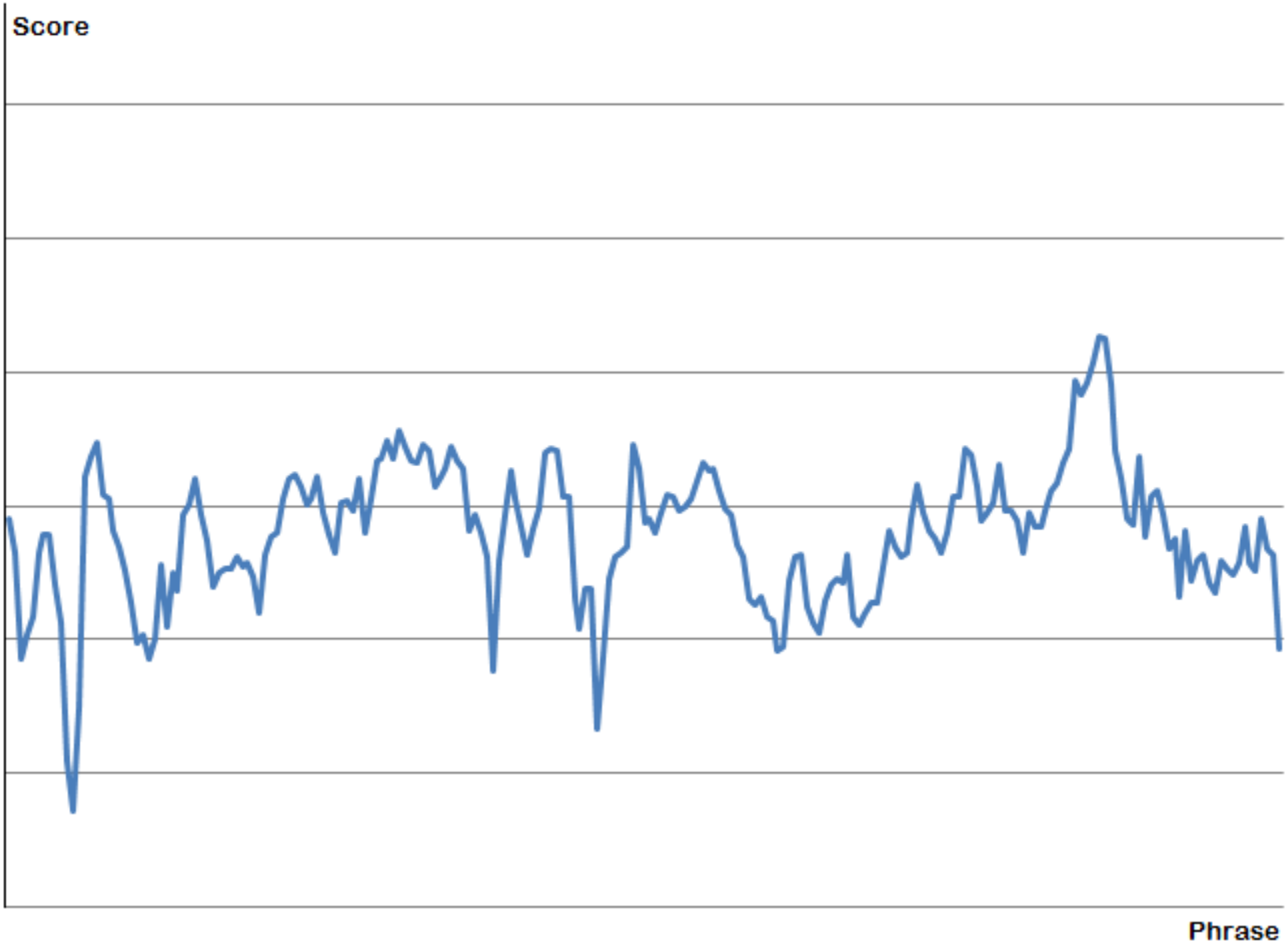
- 30 articles issus de l'encyclopédie en ligne *Wikipédia*
- Articles ayant pour sujet des lieux : pays (ex. Danemark) ou villes (ex. Salzbourg)
  Selon nos observations, dans cette catégorie d'articles, les différentes sections correspondent généralement à différents "thèmes" (histoire, géographie, culture, etc.).
- 1584 paragraphes (donc $1584 - 30 = 1554$ ruptures possibles)
- 302 titres de section (donc 302 ruptures de référence)

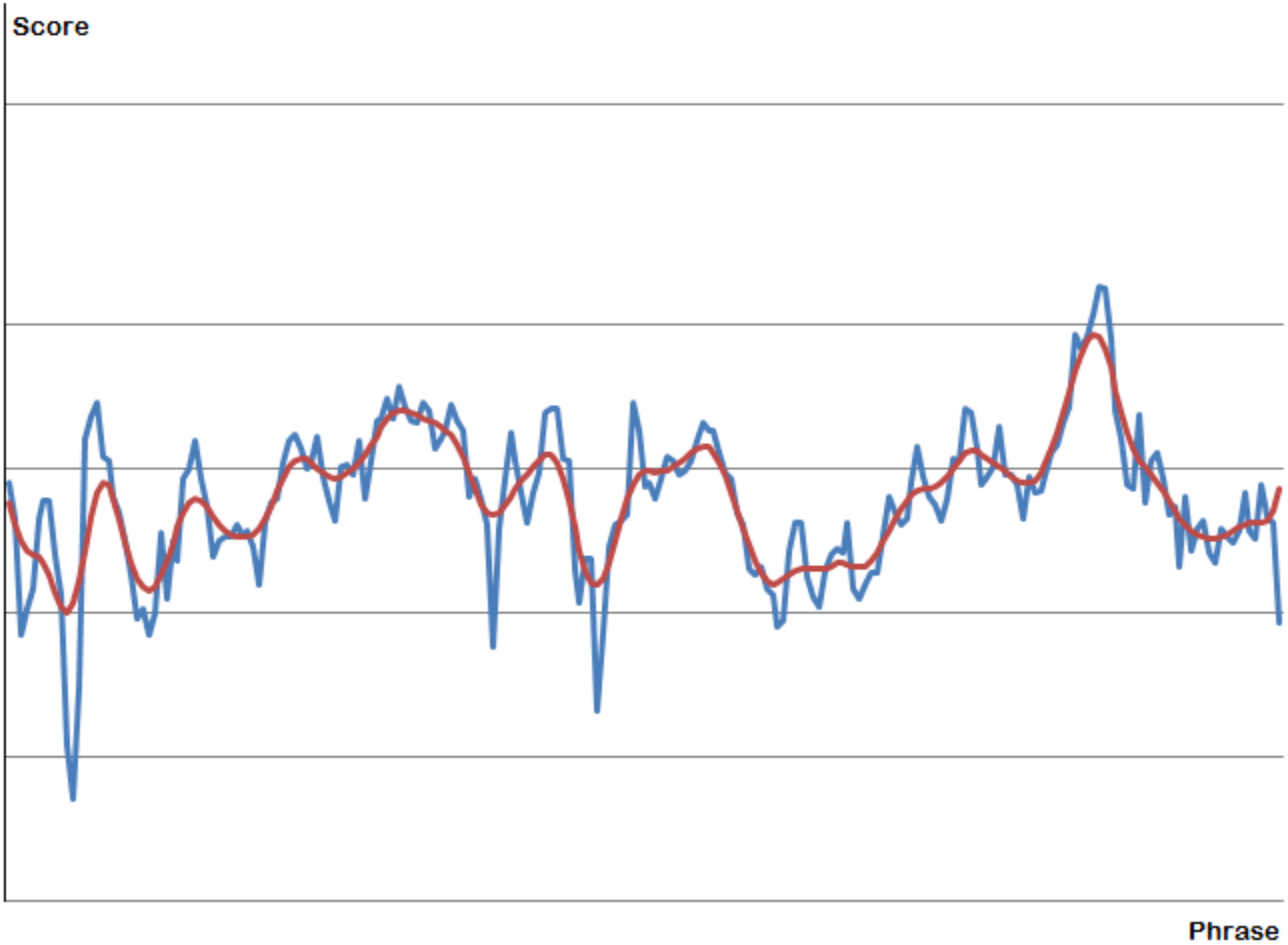
Algorithme de segmentation

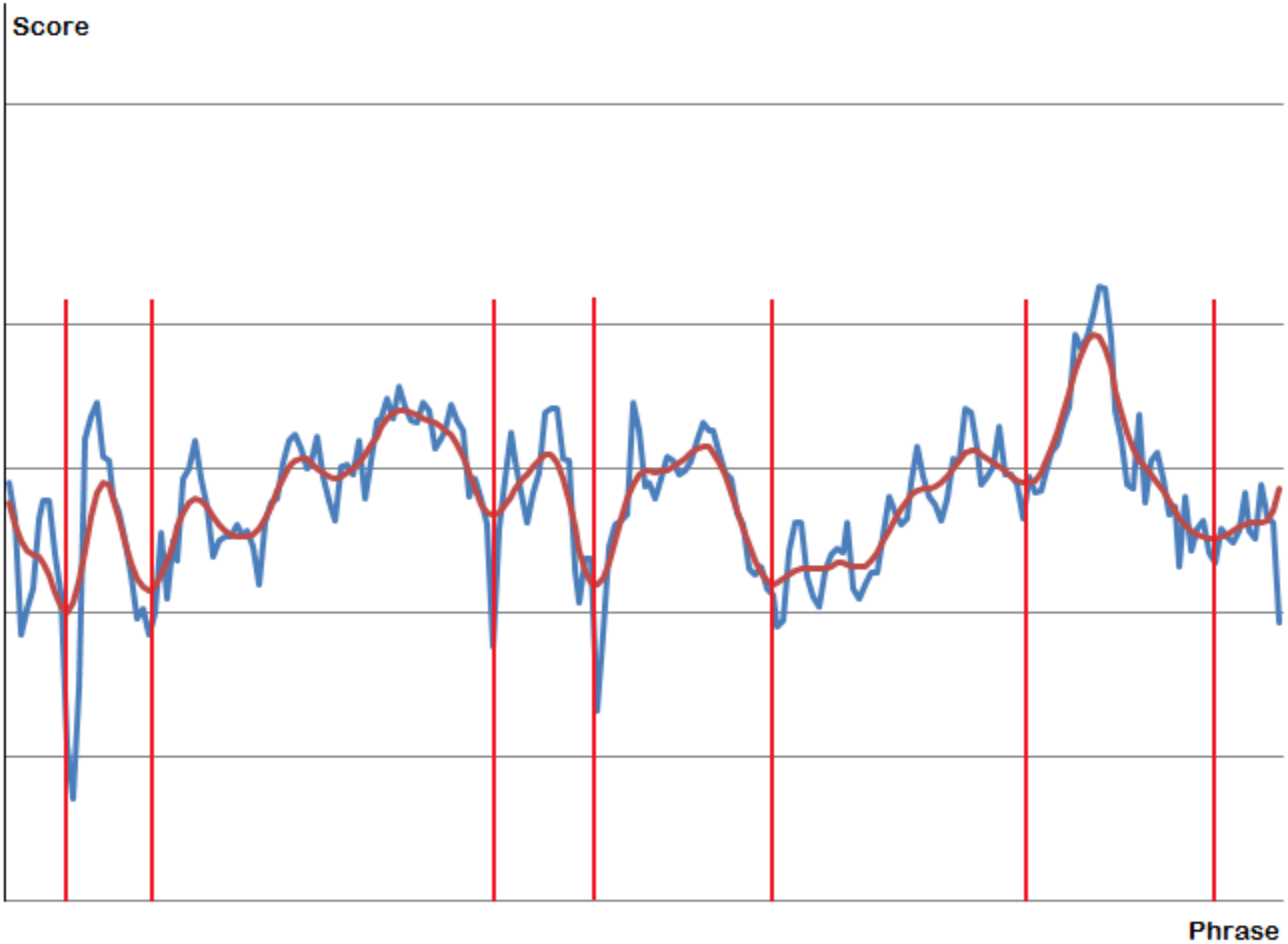
- Approche linéaire par fenêtre glissante, à la manière de Hearst (1997)
- Unité de base : la phrase. Fenêtre d'observation de 6 unités.



- $S = \log\left(\frac{N_{liens}}{N_{liens\ possibles}}\right)$


Score
Phrase


Score
Phrase


Score
Phrase

Projection des liens cohésifs sur le corpus

- On marque dans le corpus des liens non pondérés entre couples de mots appartenant à des phrases différentes
- Répétitions lexicales : répétitions de lemmes de noms / verbes / adjectifs
- Synonymes : couples recensés par le dictionnaire *Dicosyn* (Université de Caen)
- Voisinage distributionnel : couples de voisins dont le score de *Lin* dépasse 0.25 et pour lesquels chaque membre du couple est parmi les 15 meilleurs voisins de l'autre membre

Projection des liens cohésifs sur le corpus

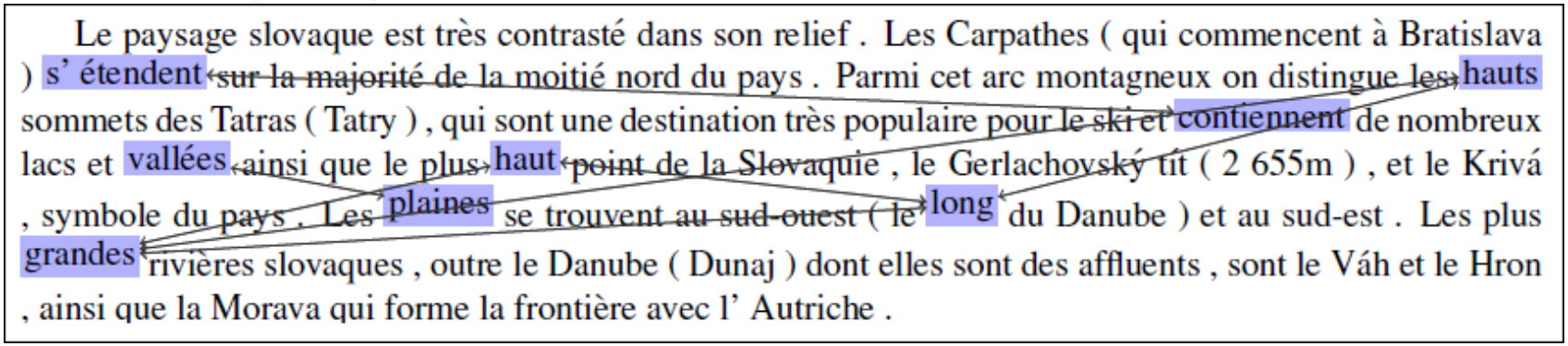
Liens de répétition lexicale :

Le paysage slovaque est très contrasté dans son relief . Les Carpathes (qui commencent à Bratislava) s' étendent sur la majorité de la moitié nord du pays . Parmi cet arc montagneux on distingue les hauts sommets des Tatras (Tatry) , qui sont une destination très populaire pour le ski et contiennent de nombreux lacs et vallées ainsi que le plus haut point de la Slovaquie , le Gerlachovský tít (2 655m) , et le Krivá , symbole du pays . Les plaines se trouvent au sud-ouest (le long du Danube) et au sud-est . Les plus grandes rivières slovaques , outre le Danube (Dunaj) dont elles sont des affluents , sont le Váh et le Hron , ainsi que la Morava qui forme la frontière avec l' Autriche .

- Couples repérés : *pays/pays haut/haut Danube/Danube*
- Seulement trois liens !

Projection des liens cohésifs sur le corpus

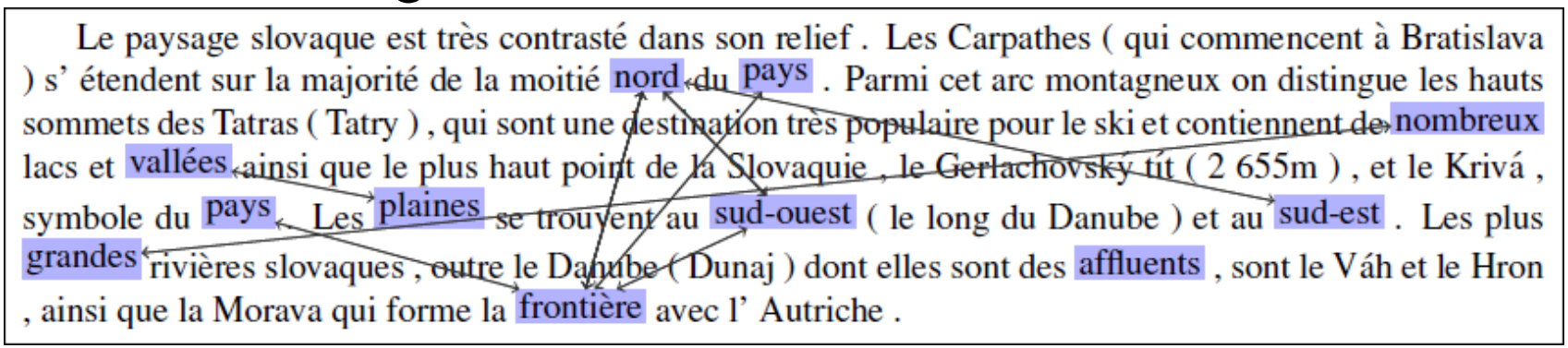
Liens de synonymie :



- Couples repérés : *s'étendre/contenir plaine/vallée haut/long long/grand grand/haut*

Projection des liens cohésifs sur le corpus

Liens de voisinage distributionnel :



- Couples repérés : *pays/frontière frontière/nord frontière/sud-ouest nord/sud-ouest nord/sud-est vallée/plaine grand/nombreux*
- Les voisins mettent au jour des liens qu'aucune ressource classique ne permet de capter.

Évaluation

Décisions prises sur l'évaluation

Quelle référence ?

- Options rejetées : annotations manuelles ou concaténations de textes
- Option choisie : utilisation de titres de section comme ruptures de référence

Quel score d'évaluation ?

- Les scores habituels de précision et de rappel ne sont pas adaptés pour l'évaluation d'un système de segmentation thématique
- Scores calculés : *WindowDiff* (Pevzner & Hearst, 2002)

Évaluation
Résultats

- Pour mettre en perspective les résultats : *baseline* "hasard", qui place les ruptures au hasard, leur nombre pour chaque texte étant approximativement connu
- Résultats :

Liens pris en compte	WindowDiff
Hasard	0.452
Répétition	0.359
Synonymie	0.358
Voisinage	**0.336**

Plan

Conclusion

- L'objectif de cette étude était de montrer la pertinence du voisinage distributionnel pour la détection de la cohésion lexicale. Nous avons à cette fin impliqué les voisins recensés par notre ressource dans un système de segmentation thématique. Les résultats obtenus montrent un apport significatif de la ressource mobilisée.
- Cette expérience mériterait d'être approfondie :
  - comparaison des voisins avec une ressource plus similaire (collocations) ?
  - combinaison des ressources ?
- La segmentation thématique n'est pas une fin en soi !

Bibliographie I

Beeferman, D., Berger, A., & Lafferty, J. 1997. Text segmentation using exponential models. *Pages 35–46 of : Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*.

Choi, Freddy Y. Y. 2000. Advances in domain independent linear text segmentation. *Pages 26–33 of : In Proceedings of NAACL*.

Choi, Freddy Y. Y., Wiemer-hastings, Peter, & Moore, Johanna. 2001. Latent semantic analysis for text segmentation. *Pages 109–117 of : In Proceedings of EMNLP*.

Bibliographie II

Fabre, C., & Bourigault, D. 2006. Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. *In : Actes de la 13^{e}conférence sur le Traitement Automatique de la Langue Naturelle*.

Ferret, Olivier. 2002 (24–27 juin). Segmenter et structurer thématiquement des textes par l'utilisation conjointe de collocations et de la récurrence lexicale. *In : TALN*.

Halliday, M. A. K., & Hasan, Ruqaiya. 1976. *Cohesion in English*.

Hearst, M. A. 1997. TextTiling : segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, **23**(1), 33–64.

Bibliographie III

Hernandez, Nicolas. 2004. *Détection et Description Automatique de Structures de Texte*. Ph.D. thesis, Université de Paris-Sud XI.

Hoey, M. 1991. *Patterns of lexis in text*. Oxford University Press (Oxford).

Kozima, Hideki. 1993. Text Segmentation Based on Similarity between Words. *Pages 286–288 of : Meeting of the Association for Computational Linguistics*.

Bibliographie IV

Lin, M., Nunamaker Jr., J. F., Chau, M., & Chen, H. 2004. Segmentation of lecture videos based on text : a method combining multiple linguistic features. *In : Proceedings of the 37th Annual Hawaii International Conference on System Sciences*.

Morris, J., & Hirst, G. 2004. Non-classical Lexical Semantic Relations. *Pages 46–51 of : Workshop on Computational Lexical Semantics*. Boston : D.M.a.R. Girju (Ed.).

Pevzner, Lev, & Hearst, Marti A. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, **28**, 1–19.