Propositions à constituants nominaux multiples : une étude comparative de corpus de contes en français et en dagara (Burkina Faso)

Communication au colloque Sénélangues 2015
à l'Université Cheikh Anta Diop de Dakar,
le 24 et 25 avril 2015

Colin R. Mills
SIL

En théorie une proposition peut comporter un nombre illimité de constituants nominaux (CN). Cependant en pratique on ne trouve, le plus souvent, qu'un nombre très bas de CN dans une même proposition. En fait, plus une proposition a de CN, moins ses chances d'occurrence sont élevées.

Dans cette communication je voudrais explorer quelles pourraient être les raisons pragmatiques pour cette limitation. On verra comment cette limitation s'opère dans le cas de quatre corpus comparables de contes de la langue dagara du Burkina Faso et du français de France.

Je commencerai d'abord par une introduction théorique qui situe ce thème dans le contexte d'une discussion plus large de la densité informationnelle et qui donne un survol de la littérature pertinente. Ensuite il y aura une courte présentation de la méthodologie employée. Après cela, je vais présenter les résultats de l'analyse de ces corpus par rapport aux propositions à CNs multiples. Des exemples permettront de voir comment et pourquoi dans les corpus ces propositions à CN multiples sont évitées, ou bien pas évitées dans certains cas. Je voudrais ensuite terminer en élargissant l'optique pour considérer brièvement les implications non seulement théoriques, pour la densité informationnelle et la typologie linguistique, mais aussi pratiques, dans les domaines de l'éducation et de la traduction.

Donc, on commence par la discussion théorique et le survol de la littérature.

Si on veut comparer la densité informationnelle entre plusieurs propositions similaires, on peut examiner à combien d'entités chaque proposition se réfère, c'est-à-dire combien de choses ou personnes elle mentionne.

On va comparer ces trois propositions :

A. Marie joue.
B. Marie joue dans le jardin.
C. Marie joue avec Pierre dans le jardin.

On va imaginer que le contexte pragmatique reste le même pour les trois propositions.

Donc on voit que *Marie joue* contient moins d'informations que *Marie joue dans le jardin*. Et la proposition B contient moins d'informations que *Marie joue avec Pierre dans le jardin*. Donc, plus on a de CN, plus on a d'informations.

Selon la Functional Discourse Grammar de Hengeveld et Mackenzie (2008), chaque fois qu'une entité est nommée explicitement, le locuteur fait un sous-acte référentiel, et un effort cognitif est demandé à l'allocutaire pour identifier cette entité. Donc plus on mentionne d'entités dans une proposition, plus elle risque de demander d'efforts cognitifs. Ce sont des efforts pour le locuteur qui doit concevoir et énoncer la proposition et ce sont des efforts pour l'allocutaire aussi qui essaie de comprendre.

Dans la typologie linguistique, les propositions à quatre arguments sont très rares dans les langues du monde entier selon Hengeveld et Mackenzie : normalement on n'en trouve que trois, mais ils identifient quatre dans une construction causative en turc (2008: 189).[1] Munro et Gordon (1982), qui ont étudié des langues vernaculaires d'Amérique du Nord, trouvent au maximum quatre CN par proposition dans la langue chickasaw, si on n'inclut pas les déterminants possessifs. Elles trouvent que des phrases à CN multiples qu'on a tendance à trouver dans des grammaires ne représentent pas bien la langue et manque de naturalité. Miller et Weinert (1998) suggèrent que l'anglais ou le russe parlés, surtout dans la conversation spontanée, ne sont pas si différents de cette langue vernaculaire d'Amérique du Nord. Ils trouvent que l'anglais et le russe parlés se limitent normalement à un maximum de trois CN par proposition.

Dans le cadre de mes recherches pour une thèse doctorale (Mills, 2014), j'ai examiné ce phénomène de propositions à CN multiples, ainsi que d'autres phénomènes de la densité informationnelle. L'un des facteurs qui pourraient être primordiaux dans le choix du locuteur à utiliser ou non des propositions à CN multiples est le canal employé (écrit ou oral), comme, selon Miller et Weinert (1998), cette limitation s'opère dans l'anglais ou le russe parlés.

Un autre facteur très important est l'influence de la parole écrite dans la société à laquelle ce locuteur appartient. Il y a beaucoup de critères concrets qu'on peut utiliser pour évaluer l'influence de la parole écrite dans une société : par exemple le taux d'alphabétisation, le pourcentage de la population qui ont fini l'école primaire, le nombre de publications disponibles etc.

Donc, en ce qui concerne la méthodologie je me suis basé sur ces deux axes, à savoir le canal employé (écrit ou oral) et l'influence de la parole écrite dans la société, et j'ai construit quatre corpus de contes, les corpus dagara oral, dagara écrit, français oral et français écrit.

[1] Hengeveld and Mackenzie (2008:189): les gloses en anglais de cette construction causative en turc sont "I made Hasan-DAT pitcher-ACC cupboard-DAT put-CAUS-PST-1.sg"

Des efforts ont été faits pour essayer de faire des corpus comparables, aussi similaires que possible: le conte a été choisi parce que c'est un genre de texte qu'on trouve dans les deux cultures, bien que le conte ait une signification différente dans chacune des deux cultures. Autant que possible les corpus avaient vingt contes chacun, la moitié racontée par des hommes adultes, la moitié par des femmes adultes. Les contes en français venaient de publications ou enregistrements existants. Pour recueillir les contes en dagara j'ai organisé deux concours de contes, un concours oral et un concours écrit, dans un village dagara. Le but de la méthodologie de concours était d'essayer d'avoir des contes en dagara de contextes analogues que pour le français. Les contes dagara étaient eux aussi destinés pour la publication ou interprétés devant un public assez nombreux. Cependant, il n'était pas possible de rendre comparables tous les détails. Par exemple le niveau d'éducation formelle des locuteurs français était plus élevé que celui des locuteurs dagara, mais cette différence reflète la différence de l'influence de la parole écrite dans les deux sociétés.

Après la transcription et informatisation en utilisant le Text Encoding Initiative, les méthodes de la linguistique comparative ont été utilisées pour l'analyse des corpus.

Les contes étaient donc divisés en propositions. On a considéré que les verbes auxiliaires et les verbes qui exprimaient le temps, l'aspect ou la modalité ne formaient pas une proposition distincte mais appartenaient à la même proposition que le verbe qui exprimait une action, un état, un processus etc.

La prochaine étape était de marquer les propositions qui se référaient à plus de deux entités. La référence à une entité pourrait se faire par le moyen d'un constituant nominal entier ou un pronom, y compris des pronoms relatifs. On comptait aussi des entités qui étaient exprimées seulement par une proposition sans CN explicite, comme *avec* dans *pour manger avec.*

Donc selon cette méthode l'exemple

il se mit à s'en demander la raison
1 2 3 4 5

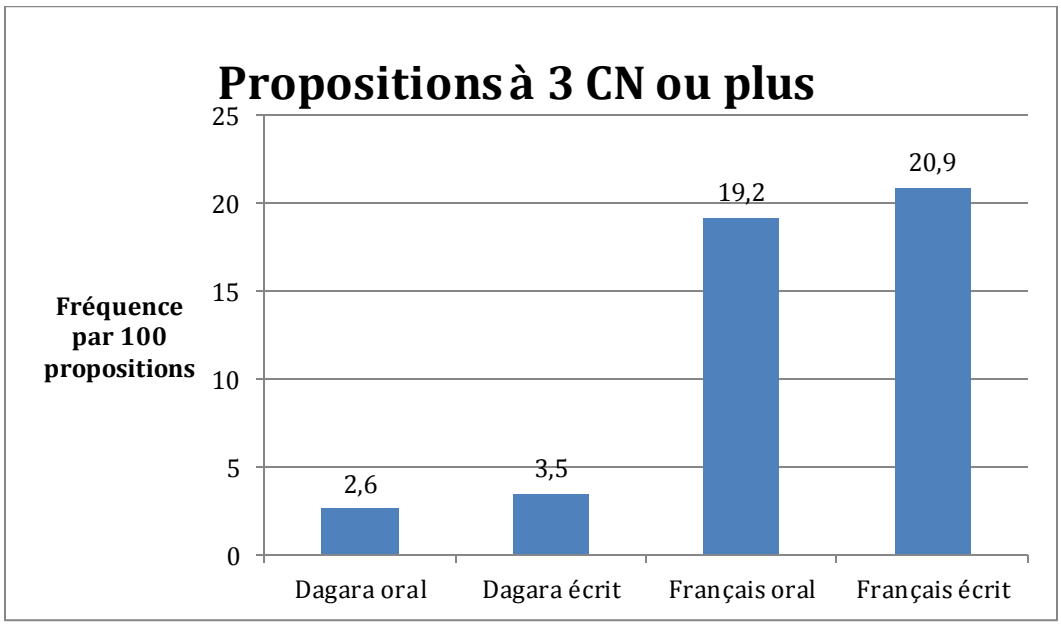
serait considéré comme une seule proposition et il y aurait cinq CN explicitement exprimées.

Plusieurs CN qui formaient un seul CN supérieur, par exemple "les mains de la jeune fille" étaient comptés comme un seul CN. La complexité au sein du CN était considérée dans une autre partie de l'étude.

Les appositions aussi, comme "sa Majesté la reine" étaient traitées ailleurs.

La fréquence de propositions ayant trois CN ou plus a donc était calculée.

Les **résultats** pour les quatre corpus étaient comme on voit dans l'histogramme :



Donc par 100 propositions il y avait 2,6 propositions à trois CN ou plus pour le dagara oral, 3,5 pour le dagara écrit, 19,2 pour le français oral et 20,9 pour le français écrit.

La fréquence est légèrement plus haute pour le corpus écrit que pour le corpus oral de chaque langue. Ceci s'explique par la difficulté de concevoir ou de comprendre un grand nombre de CN dans une même proposition. On a plus de temps à l'écrit pour réfléchir et gérer ces difficultés cognitives.

Mais on voit clairement que le plus grand écart est entre le dagara et le français, qui viennent de deux sociétés où la parole écrite a un statut bien différent. Donc on peut expliquer la grande fréquence de ces propositions à CN multiples dans les corpus français par l'entraînement cognitif que les conteurs ont reçu pendant les longues années d'éducation formelle et leurs fréquentes lectures etc. Du côté dagara, par contre, on voit qu'il y a une préférence d'éviter ces structures cognitivement plus difficiles.

Regardons un peu plus en détail certains exemples qui viennent des corpus pour voir les stratégies que les locuteurs des deux langues adoptent, soit pour cumuler les CN dans une même proposition ou pour éviter un cumul de CN.

On vient de voir qu'en dagara la fréquence de ces propositions à CN multiples est très basse. Elles sont moins de 4% des propositions dans les deux corpus dagara. Donc examinons les stratégies que les locuteurs utilisent pour éviter de cumuler les CN dans une même proposition.

L'exemple ici montre comment on fait en dagara pour éviter trop de CN dans une même proposition :

(et le roi a dit)

ka	**ba**	**yi/**
COMP	*3PL*	*sortir*

qu'ils sortent

de	**a**	**wii/**
prendre	*DET*	*cheval.PL*

prennent les chevaux

dɪw	**bɪɛrɛ**	**nʊ́/**
chasser	*suivre.IMPF*	*avec*

poursuivent avec

Les gens qu'on poursuit ne sont pas mentionnés, même pas par un pronom, parce qu'on peut les identifier par le contexte. Les serviteurs du roi sont introduits par le pronom *ba* dans la première proposition, mais ils restent le sujet de chaque proposition et n'ont plus besoin d'être explicités. Donc la troisième proposition parle implicitement de trois entités: les serviteurs qui poursuivent, les échappés qu'on poursuit et les chevaux qu'on utilise pour poursuivre, mais on ne les mentionne pas explicitement dans cette troisième proposition, parce que toutes ces entités sont déjà identifiables dans le contexte. Il ne reste que ce mot *nʊ́*, avec, qui est le vestige de l'instrument utilisé.

Et en fait, les deux premières propositions ajoutent seulement des informations inférables, comme le fait qu'ils devaient sortir du palais pour trouver les chevaux, et le fait qu'ils devaient *prendre* les chevaux avant de les utiliser pour la poursuite. Donc on voit que ces deux premières propositions n'ont pas un contenu sémantique très important. Elles sont principalement là pour alléger la troisième proposition, pour qu'elle ne contienne pas beaucoup de CN explicites.

Cette façon dont le locuteur dagara s'arrange pour éviter de se référer à trop d'entités dans une même proposition fait penser à ce que Chafe (1987) a dit sur le langage parlé. Chafe dit que dans une unité d'intonation, qui équivaut la plupart du temps à une proposition, il n'y a normalement qu'un seul nouveau concept.

On pense aussi à ce que des psycholinguistes comme Gernsbacher (1990)[2] ont découvert sur l'importance cognitive de la proposition. Le fait de finir une proposition permet à l'allocutaire de classer dans la mémoire les informations que cette proposition contient et de passer à la prochaine proposition avec un cerveau moins encombré. Donc s'il y a moins de CN par proposition, cela permet généralement plus de facilité dans la conception et la compréhension.

Mais pourquoi est-ce que les corpus en français avaient beaucoup plus de propositions à CN multiples?

En partie cela peut s'expliquer par la possibilité en français d'avoir des verbes pronominaux, et même des verbes pronominaux comme auxiliaires. On pense par exemple à la proposition *il se mit à s'en demander la raison* avec cinq CN.

il se mit à s'en demander la raison
1 2 3 4 5

Mais cela n'arrive pas tellement souvent dans les corpus en français, donc ce n'est pas toute l'explication. Une autre explication partielle est que les corpus français avaient dans une même proposition plus de circonstants que les corpus en dagara. Ces circonstants en français indiquent le temps, le lieu ou la destination. En dagara il n'y en avait souvent pas plus d'un de ces circonstants par proposition et même quelquefois on avait toute une proposition pour indiquer ce qu'on exprimait par un circonstant en français.

Prenons l'exemple de cette proposition à six CN en français:

le prince et la princesse avaient lancé un premier appel pressant un matin à l'aube sur R[adio] B[erry] S[ud].

Ici on a le circonstant "à l'aube" au milieu d'une autre proposition avec deux autres circonstants, "un matin" et "sur RBS" (Radio Berry Sud)

Tandis qu'en dagara on avait tout une proposition à un seul CN pour exprimer le concept "à l'aube":

k' **a** **zie** **wa** **cãanɩ**

COMP *DET* *endroit* *venir/PERF* *être.clair*

et quand le jour s'était levé

Pour résumer, alors, les propositions à CN multiples deviennent plus fréquentes en fonction de l'influence de la parole écrite, comme on voit dans l'histogramme. Donc il y a en général une corrélation entre la densité informationnelle et l'influence de la parole écrite. Cette influence peut venir

[2] "comprehenders represent each clause of a multi-clause sentence in its own substructure" (Gernsbacher, 1990:26).

d'une combinaison du canal utilisé, écrit ou oral, et de l'importance de la parole écrite dans la société, mais dans cette étude l'influence du canal écrit ou oral était moins grande que l'influence sociétale.

On trouve la même tendance en examinant la densité informationnelle dans d'autres structures linguistiques à plusieurs niveaux syntaxiques ou discursifs: au niveau du CN, au niveau de la proposition, et au niveau des épisodes discursifs. En général la densité informationnelle s'augmente du corpus dagara oral, au dagara écrit, encore plus en français oral et au maximum dans le corpus français écrit. Donc la conclusion globale est que dans ces corpus, plus la parole écrite a de l'influence, plus la densité informationnelle est élevée. Et cela veut dire que la difficulté cognitive sera élevée elle aussi.

Si c'est vraiment la difficulté cognitive qui occasionne les différences en densité informationnelle, il devrait être possible de généraliser la tendance qu'on trouve ici à d'autres genres et à d'autres langues. C'est une hypothèse à étudier à l'avenir!

Mais pour maintenant, je voudrais voir quelles sont les implications de ces résultats? On va examiner certaines implications pour l'éducation et la traduction et pour la linguistique.

Dans le domaine de l'éducation, on doit reconnaître que des propositions à CN multiples seront plus difficiles à comprendre, surtout pour des élèves d'une langue maternelle où ces structures sont moins fréquentes. Donc il faut commencer par des textes à moindre densité informationnelle, mais entraîner les élèves à avoir plus de facilité à comprendre les textes plus compliqués.

Dans le domaine de la traduction, on doit reconnaître que même si on peut traduire une proposition à CN multiples quasi telle quelle dans une autre langue, on ne trouve pas ces structures avec la même fréquence dans les deux langues. La même chose est vraie à un moindre degré si on veut adapter un texte écrit à l'oral, ou vice-versa. Mais si on traduit du français au dagara, par exemple, sans essayer de réduire le nombre de CN par proposition, la traduction en dagara sera plus difficile à comprendre que l'original n'avait été pour son public. Cette traduction en dagara ne sera probablement pas appropriée ou naturelle au contexte de la langue cible. Dans mes recherches à l'avenir, je voudrais bien appliquer les résultats de ces recherches sur la densité informationnelle à la traduction, en développant des idées comme celles de Fabricius-Hansen (1996, 1998; voir aussi Fabricius-Hansen et al. 2005) qui parle de *Information splitting* (scinder les informations).

Dans le domaine de la linguistique, la typologie linguistique a beaucoup à gagner si elle se laisse enrichir par des considérations sociolinguistiques et pragmatiques. Il faut qu'on soit conscient de la fréquence d'occurrence d'une structure et les contextes où on peut le rencontrer : autrement on risque de se concentrer sur des fragments de langage qui ne sont pas tellement représentatifs d'une langue. Pour mieux prendre en compte le contexte, on pourrait axer l'investigation typologique non seulement sur les structures mais

aussi sur l'arrière-plan socioculturel des langues. Comme ça, peut-être qu'on trouvera des similarités non seulement entre langues SVO par exemple, mais aussi entre les langues où la parole écrite a relativement peu d'influence. Et peut-être que si on comparait entre beaucoup de langues les structures qu'on trouve plus fréquemment à l'oral qu'à l'écrit ou le contraire, on aurait des aperçus nouveaux et intéressants.

En outre, je voudrais que la densité informationnelle soit plus généralement reconnue comme un concept très utile pour la linguistique. La densité informationnelle peut nous aider à mieux comprendre comment fonctionnent les processus de parler ou de comprendre le langage. Et elle peut nous aider à expliquer beaucoup de phénomènes linguistiques en explicitant le lien entre l'énoncé et son contexte pragmatique et socioculturel.

Bibliographie

Chafe, W. L. (1987a). Cognitive constraints on information flow. In R. S. Tomlin (Ed.), *Coherence and grounding in discourse* (pp. 21-51). Amsterdam: John Benjamins.

Fabricius-Hansen, C. (1996). Information density: A problem for translation and translation theory. In M. Doherty (Ed.), *Information structure: A key concept for translation theory. Special issue of Linguistics, 34,* 521-566. Berlin: Mouton.

Fabricius-Hansen, C. (1998). Informational density and translation, with special reference to German–Norwegian–English. In S. Johansson & S. Oksefjell (Eds.), *Corpora and cross-linguistic research: Theory, method, and case studies* (pp. 197-234). Amsterdam: Rodopi.

Fabricius-Hansen, C., Ramm, W., Solfjeld, K., & Behrens, B. (2005). Coordination, discourse relations, and information packaging — cross-linguistic differences. In M. Arnague, M. Bras, A. Le Draoulec, & L. Vieu (Eds.), *Proceedings of the Symposium on the Exploration and Modelling of meaning (SEM-05), Biarritz, France, November 14-15, 2005* (pp. 85-93). Also: SPRIK Report No. 31, September 2005, Department of Literature, Areas Studies, and European Languages, University of Oslo.

Gernsbacher, M. A. (1990). *Language comprehension as structure building.* Hillsdale, NJ: Erlbaum.

Hengeveld, K., & Mackenzie, J. L. (2008). *Functional discourse grammar: A typologically-based theory of language structure.* Oxford: Oxford University Press.

Miller, J., & Weinert, R. (1998). *Spontaneous spoken language: Syntax and discourse.* Oxford: Clarendon Press.

Mills, C. R. (2014). *Information density in French and Dagara folktales: A corpus-based analysis of linguistic marking and cognitive processing.* Thèse doctorale inédite, Queen's University Belfast.

Munro, P., & Gordon, L. (1982). Syntactic relations in Western Muskogean: A typological perspective. *Language*, *58(1*), 81-115.