*TALN 2011, Montpellier, 27 juin – 1er juillet 2011*

# La complexité linguistique
# Méthode d'analyse

Adrien Barbaresi
ICAR, ENS LYON

**Résumé.** La complexité linguistique regroupe différents phénomènes dont il s'agit de modéliser le rapport. Le travail en cours que je décris ici propose une réflexion sur les approches linguistiques et techniques de cette notion et la mise en application d'un balayage des textes qui s'efforce de contribuer à leur enrichissement. Ce traitement en surface effectué suivant une liste de critères qui représentent parfois des approximations de logiques plus élaborées tente de fournir une image « raisonnable » de la complexité.

**Abstract.** Linguistic complexity includes various linguistic phenomena which interaction is to be modeled. The ongoing work described here tackles linguistic and technical approaches of this idea as well as an implementation of a parsing method which is part of text enrichment techniques. This chunk parsing is performed according to a list of criteria that may consist in logical approximations of more sophisticated processes in order to provide a « reasonable » image of complexity..

**Mots-clés :** Complexité, lisibilité, allemand, analyse de surface.

**Keywords:** Complexity, lisibility, German, chunk parsing.

## 1 Enjeux

L'analyse de la complexité se situe dans le cadre de l'assistance à la compréhension. Il s'agit ici de déterminer la lisibilité d'un texte pour des humains ou pour des machines, c'est-à-dire d'une part le niveau de maîtrise et de pratique de la langue requis et d'autre part le modèle formel et les instruments à utiliser.

Ce thème très riche invite à penser les langues avant de les analyser à différentes échelles et selon différents modes opératoires. Du point de vue disciplinaire, on peut le situer à la croisée de la linguistique, des études sur la lisibilité, des sciences cognitives et de la théorie de l'information. Le traitement envisagé aborde des notions d'informatique et l'intégration d'un marquage des textes étudiés, approches qui sont en prise directe avec la réflexion actuelle chez les chercheurs et les entrepreneurs sur les données, leur statut, leur forme et leur traitement.

Cette démarche s'inscrit en ce sens entre la réflexion en sciences humaines et l'exploitation technique. Elle est également en rapport avec la transmission d'une langue et son « outillage » (Auroux, 1994). En effet, au-delà d'une tentative consistant à modéliser les processus à l'œuvre lors du déchiffrage d'un texte, il s'agit d'équiper une langue, de l'enrichir d'une description utile.

De fait, pour (Gibson, 1998), étudier la complexité linguistique, c'est expliquer les étapes de l'apprentissage de sa langue maternelle par un enfant, donner des éléments pour aborder les problèmes syntaxiques chez les aphasiques, et fournir des applications dans lesquelles la compréhensibilité de la langue est importante, comme les correcteurs grammaticaux ou la génération automatique de textes.

L'intérêt premier porte sur la complexité de phrases, de paragraphes ou de textes écrits en allemand. L'attention portera spécifiquement sur un standard de cette langue, considéré comme une *koinè*, une langue commune qui dépasse des disparités régionales.

Il ne s'agit donc pas d'un travail de comparaison entre différentes langues. En revanche, la pertinence de la notion de sous-langage pourra être examinée. De même, dans un deuxième temps, une adaptation de la démarche et des outils à l'anglais et au français apportera peut-être quelques éléments qui viendront enrichir la compréhension du sujet en infirmant ou confirmant des hypothèses.

# 2 Méthode

## 2.1 Diviser pour mieux appréhender

Du point de vue linguistique, la complexité est avant tout un phénomène difficile à définir (Kusters & Muysken, 2001).

Dans la lignée du concept d'« architecture » de la complexité (Simon, 1962), il s'agit de diviser pour mieux comprendre. Tout système complexe est quasiment décomposable (« *nearly-decomposable* ») en sous-éléments, qui peuvent eux-mêmes être complexes. Même si cette approche ne tient pas compte des interactions entre les différents sous-systèmes, elle propose néanmoins de chercher des strates élémentaires en bas de la hiérarchie d'un système, postulant que celles-ci peuvent être comprises et modélisées efficacement.

Or, dans le cas qui nous intéresse, ces unités pourraient être les mots, qui permettent d'aborder la complexité morphologique ainsi que par extension les bouts de phrases (*chunks*), afin de saisir la complexité morpho-syntaxique et éventuellement de s'intéresser à la décomposition syntaxique d'un groupe ou d'une phrase. Les phénomènes interphrastiques représentent un niveau supérieur et bien plus délicat à analyser automatiquement, celui du discours et de la linguistique textuelle.

La division du texte en unités linguistiques pertinentes comme par exemple les groupes nominaux présente également l'intérêt de s'approcher des études psycho-linguistiques qui prennent en compte la notion de coût de rattachement des composants les uns aux autres. En effet, la tradition cognitiviste voit la complexité linguistique comme un coût supplémentaire de traitement pour un éventuel récepteur. La dimension syntaxique joue un grand rôle à travers notamment les principes d'intégration et de distance entre constituants. Le coût de rattachement des unités linguistiques est modélisé en termes d'unités d'énergie et d'unités de mémoire à très court terme (Gibson, 1998), dans la lignée des premiers travaux dans cette discipline (Miller, 1956).

Le cadre de pensée de la grammaire générative a incontestablement contribué à faire évoluer l'idée de complexité. Le fait qu'il s'agisse avant tout d'un modèle syntaxique du langage a aussi joué un rôle dans l'abord de cette notion en conférant une certaine importance à la syntaxe qui joue aujourd'hui encore un rôle. Cela dit, cette notion est loin de se limiter à cette dimension. La complexité d'un point de vue linguistique regroupe différents phénomènes dont il s'agit de modéliser le rapport, de même que différentes approches à croiser.

## 2.2 Croiser les approches et enrichir les pratiques

(Corbin, 1980) opposait « deux façons pour un linguiste de constituer les données sur lesquelles il travaille : l'introspection, le corpus », et deux méthodes, la « linguistique de bureau » et la « linguistique de terrain ». Cette distinction a également investi le traitement automatique des langues, où elle semble toujours d'actualité selon (Cori, 2008), qui oppose « TAL théorique » et « TAL robuste ». Elle concerne aussi bien la démarche (processus longs et diversifiés d'une part, traitement plus près de la surface d'autre part) que les objectifs (en lien avec la constitution d'un système linguistique d'un côté et avec des contraintes pratiques comme l'hétérogénéité des textes de l'autre).

On pourrait ajouter à la distinction établie la « linguistique à l'instrument » définie par (Habert, 2005), une discipline qui exploite des corpus tout en menant une réflexion sur les instruments qu'elle utilise. Réflexion que (Latour, 1985) appelait déjà de ses vœux :

> « Nous oublions toujours l'importance des inscriptions, de leurs strates successives et leur "mise en instrument" alors que nous parlons pourtant d'êtres qui ne sont visibles qu'ainsi. »

Dans un contexte informatique qui a vu augmenter la vitesse de calcul à un rythme soutenu, on a vu s'opérer un déplacement en TAL des modèles de la compétence, fondés sur des bases théoriques fortes qui précèdent le traitement automatique, vers ceux de la performance, qui ont une vision plus statistique du langage et donnent plus d'importance à l'apprentissage par des techniques d'intelligence artificielle, si bien que ces derniers ont aujourd'hui souvent la préférence dans la communauté scientifique.

Ce changement a bel et bien contribué à améliorer l'efficacité des méthodes sur des critères quantitatifs. Toutefois, cette réussite ne doit pas faire oublier qu'il s'agit de « croiser les approches » et d'aller dans le sens d'une

évaluation qualitative qui s'attache à

> « munir les données attestées d'annotations fines, multiples, permettant de progresser vers les régularités sous-jacentes » (Habert & Zweigenbaum, 2002)

Cette démarche se veut attentive à la pertinence des textes étudiés, en rapport avec le but de recherche poursuivi, à l'existence de nombreux critères d'annotation qui doivent pouvoir coexister, se confronter, interagir à travers leur marquage et enfin à un apport proprement linguistique concernant la compréhension d'un phénomène langagier résultant de cette analyse.

# 3 Travail sur corpus

Ce travail est à situer dans un contexte où il n'y a plus de séparation hermétique entre les approches fondées sur un corpus (*corpus-based*) et celles partant d'un modèle théorique (*theory-driven*), mais bien une approche hybride fondée sur la connaissance pour évaluer, tester, générer des hypothèses (Wallis & Nelson, 2001).

Le corpus est constitué de textes écrits en allemand standard, et devra permettre une analyse différenciée de la complexité. Pour ce faire, on peut imaginer de traiter des textes divers, dont un échantillon aura été relevé manuellement voire soumis à un panel test pour permettre d'étudier la corrélation des résultats obtenus à grande échelle avec les résultats obtenus sur l'échantillon.

L'étude sera comparative, et ce à deux niveaux différents : comparaison des résultats obtenus avec ceux d'un corpus étalonné d'une part, comparaison de corpus connus pour être deux versions d'un même texte d'autre part (littérature simplifiée pour les enfants ou les apprenants par exemple, articles de journaux sur le même thème, articles scientifiques et leur vulgarisation).

Par ailleurs, on peut imaginer d'effectuer d'autres comparaisons apportant un autre éclairage sur la notion de complexité, comme les régularités dans l'utilisation d'une langue apprise (ou seconde). On peut déterminer si d'après les mesures les textes écrits par des locuteurs natifs se caractérisent par la présence d'un plus grand nombre d'indicateurs de la complexité.

Concernant les textes pris en compte, l'apport des textes du domaine public est essentiel dans une optique de transmissibilité des corpus et des résultats. En effet, la question des droits d'auteurs est encadrée beaucoup plus strictement en Allemagne qu'en France.

Aussi l'étude d'articles de journaux, par exemple une comparaison sur les articles traitant des mêmes sujets dans un quotidien à très grand tirage (la *Bild-Zeitung*) et dans un hebdomadaire (*Die Zeit*) est soumise à caution. S'il est possible d'obtenir les articles en téléchargeant et nettoyant des pages web, s'il est possible de les analyser librement, il n'est pas permis de rendre disponible le texte enrichi et étiqueté sans recourir à des techniques dites de « masque » (Rehm *et al.*, 2007), par exemple en remplaçant les mots étiquetés par des autres choisis au hasard dans la catégorie correpondante, ou en mélangeant les phrases du corpus au hasard. Ces techniques ôtent toute notion de cohérence et de cohésion au texte obtenu, limitant fortement l'intérêt d'un tel corpus.

Les articles de journaux ne sont donc utilisés jusqu'ici que pour un échantillonnage et une analyse « à vue ». L'étude sur corpus a jusqu'à présent porté essentiellement sur des romans classiques, des textes philosophiques, des romans pour enfants et des romans de gare. Le genre narratif est sur-représenté, reflétant la nature des œuvres rassemblées par le Projet Gutenberg.

Néanmoins, les articles scientifiques constituent une piste intéressante à double titre : d'une part du point de vue technique, parce qu'il est souvent aisé d'en extraire le texte, voire de remonter automatiquement d'un article à un autre et parce qu'ils peuvent dans la plupart des cas être republiés, d'autre part du point de vue linguistique, puisqu'ils ne sont pas toujours écrits par des germanophones et qu'une étude comparée pourrait souligner l'existence ou non de régularités chez les uns et les autres.

Enfin, le corpus est susceptible d'être élargi à des textes en anglais et en français, afin de valider ou d'infirmer les résultats obtenus et d'en tirer d'autres conclusions par une étude comparative.

# 4 Traitement

## 4.1 Hypothèse de recherche

En prenant connaissance d'un texte, on se demande souvent s'il est difficile à comprendre. Un rapide survol permet d'en savoir un peu plus sur le temps, les connaissances et les ressources nécessaires à une lecture satisfaisante.

Il est possible de concevoir un programme capable d'effectuer une opération de « survol » qui détermine le degré de complexité de données textuelles destinées à être traitées automatiquement. Cet indice peut prendre en compte plusieurs dimensions de ce phénomène et plusieurs échelles, de la phrase au texte, de la morphologie à l'analyse du discours en passant par la syntaxe.

Il est possible d'isoler les parties qui ne poseront aucun problème et celles qui ont besoin d'une analyse en profondeur. Plutôt que d'appliquer une méthode qui privilégierait la rapidité à la justesse (ou vice-versa) on peut alors pondérer ces deux paramètres en appliquant un traitement différencié, voire même adapté aux difficultés rencontrées.

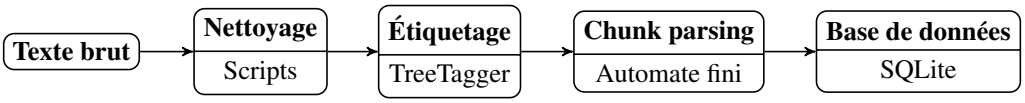
Il s'agit ici de traiter cette question avec les moyens de la linguistique informatique et de l'intégrer au sein de la chaîne d'opérations qui va du texte brut au texte enrichi syntaxiquement et sémantiquement. Mesurer la complexité peut amener à mobiliser des ressources et des traitements complexes (comme l'analyse grammaticale automatique par exemple).

Il importe dès lors d'expérimenter l'apport effectif dans un indicateur de complexité des informations construites à l'aide de ces ressources et traitements. Cela peut conduire à adopter des approximations de certains de ces traitements, à partir du moment où il s'avère que de tels traitements moins complexes fournissent une image « raisonnable » de la complexité.

## 4.2 Architecture de traitement

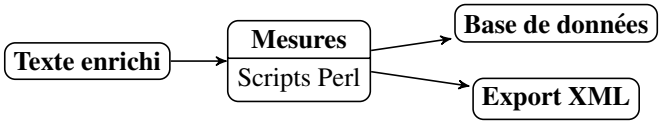
La principale contrainte du point de vue du fonctionnement est donc de maintenir un niveau de complexité algorithmique relativement faible et de rester au plus près d'une approche linéaire, balayant les mots au fil du texte.

La chaîne de traitement va du texte brut au texte enrichi selon le schéma suivant :




La chaîne de traitement est guidée par un script bash. Le nettoyage emploie notamment le logiciel sed, il comprend un découpage en tokens qui comme le reste des scripts est implémenté en Perl, en raison de la polyvalence et de l'adéquation de ce langage à la manipulation de texte. L'étiqueteur utilisé est le TreeTagger de l'IMS Stuttgart (Schmid, 1994), en raison de sa précision dans le cas de l'allemand.

Les mesures de complexité sont ensuite effectuées et exportées au choix sous deux formes différentes :




La décomposition du texte et la mesure se font par des cascades à états finis (Abney, 1996). On tente par là même d'éviter dans la mesure du possible une « dilution des heuristiques » (Trouilleux, 2009), qui pourrait résulter de l'utilisation de techniques d'intelligence artificielle nécessitant un apprentissage ciblé comme les machines à vecteurs de support (*Support Vector Machines*, SVM) ou les réseaux de neurones artificiels qui conduisent souvent à ce que l'on appelle des « boîtes noires ».

L'élaboration du programme de traitement se situe à un stade de recherche selon la classification de (Véronis, 2000) : les travaux réalisés sont de nature prospective mais ne donnent pas encore lieu à des implémentations utilisables en situation d'annotation réelle.

### 4.3 Critères

Cette liste n'a pas vocation à être exhaustive, elle concerne des phénomènes dont le repérage est envisageable, voire parfois déjà mis en pratique.

1. MOTS

   **longueur** par exemple en fixant un seuil à 17 caractères, ce qui représente en général un peu moins de 5 % des mots rencontrés dans la langue ;

   **fréquence relative** au sein du document et par rapport à une liste de mots fréquents établie à partir d'un grand corpus (par exemple la liste des 10000 mots les plus fréquents du Corpus Leipzig [1]) ;

   **lemme** savoir si le lemme est reconnu par l'étiqueteur morpho-syntaxique (le TreeTagger propose cette option) revient à approximer un ensemble de mots connus ;

2. GROUPES

   **taille** elle peut donner une idée des difficultés de rattachement des composants, par exemple dans un groupe nominal long ;

   **composition** des groupes nominaux atypiques, repérables par une suite d'étiquettes, peuvent signaler une structure plus complexe (de même qu'une erreur de l'étiqueteur, ce qui représente également une forme de complexité) ;

   **nombre** une phrase comportant de nombreux groupes posera sans doute des problèmes de rattachement, par exemple si cinq groupes nominaux sont trouvés ;

3. PHRASES

   **longueur** parmi les critères les plus souvent retenus dans les études de lisibilité, la longueur en caractères figure en bonne place (à l'usage, les seuils aux alentours de 130 et de 190 caractères semblent indiquer que la phrase se complexifie) ;

   **virgules** en allemand, les propositions sont obligatoirement suivies de virgules, c'est là un moyen efficace pour repérer d'éventuelles subordonnées, à condition d'éviter les énumérations. (Les seuils retenus jusque là sont de 0 comme indice de simplification et 3 comme complexification) ;

   **subordonnées** déterminer leur type de même que leur nombre par phrase par une observation des pronoms relatifs corrélée au critère précedent peut s'avérer pertinente ;

   **verbes** la forme et la rection des verbes peuvent corroborer l'examen des groupes nominaux ;

   **attaque d'énoncé** ce champ est flexible en allemand, il dénote parfois des phénomènes de linéarisation, par exemple une volonté de mettre en avant une partie de la phrase pour la rendre plus compréhensible.

Parmi les sources éventuelles de complexité pour lesquelles les critères manquent citons les ellipses et la densité conceptuelle en général, les ambiguïtés syntaxiques et sémantiques et à plus large échelle les phénomènes de cohérence et de cohésion textuelle.

## 5 Notion de complexité

On peut dire qu'au sein d'une tradition philosophique puis philologique les débats autour de la notion de complexité du langage sont anciens, notamment à travers la question qui consiste à savoir si certaines pensées ou idées sont en elles-mêmes complexes, si elles prennent une tournure complexe lors de leur expression ou s'il existe un lien entre les deux. Cette question se pose de manière particulière depuis le courant rationaliste, qui postule le primat de l'idée pure sur la matérialité de l'expression. Avec le concept de structure profonde, la complexité est

1. http ://wortschatz.uni-leipzig.de/

même en quelque sorte inhérente à la relecture d'une certaine « linguistique cartésienne » (Chomsky, 1966), c'est donc tout naturellement que l'attrait pour cette notion a connu une nette recrudescence lors du développement de la grammaire générative.

Plus tard, le mouvement de la « nouvelle lisibilité » dans les années 80, en butte à une conception machinale du cerveau vu comme un mécanisme de traitement de l'information, se réapproprie l'idée et l'applique au champ de la psycho-linguistique tourné vers des facteurs organisationnels comme la densité des idées et des concepts (Kemper, 1983) ou la structure et la présentation d'un document (Britton *et al.*, 1982). Ce changement de paradigme dans les sciences cognitives a partie liée avec l'émergence de la linguistique textuelle.

Plus récemment, le langage vu comme système complexe adaptatif (Beckner *et al.*, 2009) est une thèse qui séduit de plus de chercheurs qui souhaitent fournir une théorie de l'interdépendance des sous-systèmes que sont par exemple les traits phonétiques, morphologiques ou la syntaxe.

Toutes ces approches de la complexité sont autant de raisons de s'atteler à la mise en valeur de la proximité des « traces recombinées » (Latour, 1985).

# Références


ABNEY S. (1996). Partial parsing via finite-state cascades. *Natural Language Engineering*, **2**(4), 337–344.

AUROUX S. (1994). *La révolution technologique de la grammatisation*. Liège : Mardaga.

BECKNER C., ELLIS N., BLYTHE R., HOLLAND J., BYBEE J., KE J., CHRISTIANSEN M., LARSEN-FREEMAN D., CROFT W. & SCHOENEMANN T. (2009). Language Is a Complex Adaptive System : Position Paper. *Language As a Complex Adaptive System*, **59**(1), 1–26.

BRITTON B., GLYNN S., MEYER B. & PENLAND M. (1982). Effects of text structure on use of cognitive capacity during reading. *Journal of Educational Psychology*, **74**(1), 51–61.

CHOMSKY N. (1966). *Cartesian linguistics : A chapter in the history of rationalist thought*. Harper & Row.

CORBIN P. (1980). De la production des données en linguistique introspective. In *Théories linguistiques et traditions grammaticales*, p. 121–179. Villeneuve-d'Ascq : Presses Universitaires de Lille.

CORI M. (2008). Des méthodes de traitement automatique aux linguistiques fondées sur les corpus. *Langages*, **171**.

GIBSON E. (1998). Linguistic complexity : Locality of syntactic dependencies. *Cognition*, **68**(1), 1–76.

HABERT B. (2005). Portrait de linguiste(s) à l'instrument. *Texto !*, **10**(4).

HABERT B. & ZWEIGENBAUM P. (2002). Régler les règles. *TAL*, **43**(3), 83–105.

KEMPER S. (1983). Measuring the inference load of a text. *Journal of educational psychology*, **75**(3), 391–401.

KUSTERS W. & MUYSKEN P. (2001). The complexities of arguing about complexity. *Linguistic Typology*, **5**(2/3), 182–185.

LATOUR B. (1985). Les « vues » de l'esprit. *Culture technique*, **14**, 4–29.

MILLER G. A. (1956). The magical number seven, plus or minus two : Some limits on our capacity for processing information. *The Psychological Review*, **63**, 81–97.

REHM G., WITT A., ZINSMEISTER H. & DELLERT J. (2007). Corpus masking : Legally bypassing licensing restrictions for the free distribution of text collections. *Digital Humanities*, p. 166–170.

SCHMID H. (1994). Probabilistic Part-Of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, volume 12 : Manchester.

SIMON H. A. (1962). The Architecture of Complexity. *Proceedings of the American Philosophical Society*, **106**(6), 467–482.

TROUILLEUX F. (2009). Un analyseur de surface non déterministe pour le français. In *16ème Conférence sur le Traitement Automatique des Langues Naturelles*, Senlis.

VÉRONIS J. (2000). Annotation automatique de corpus : panorama et état de la technique. In J.-M. PIERREL, Ed., *Ingénierie des langues, Informatique et systèmes d'information*, p. 111–129. Paris : Hermès Science.

WALLIS S. & NELSON G. (2001). Knowledge discovery in grammatically analysed corpora. *Data Mining and Knowledge Discovery*, **5**(4), 305–335.