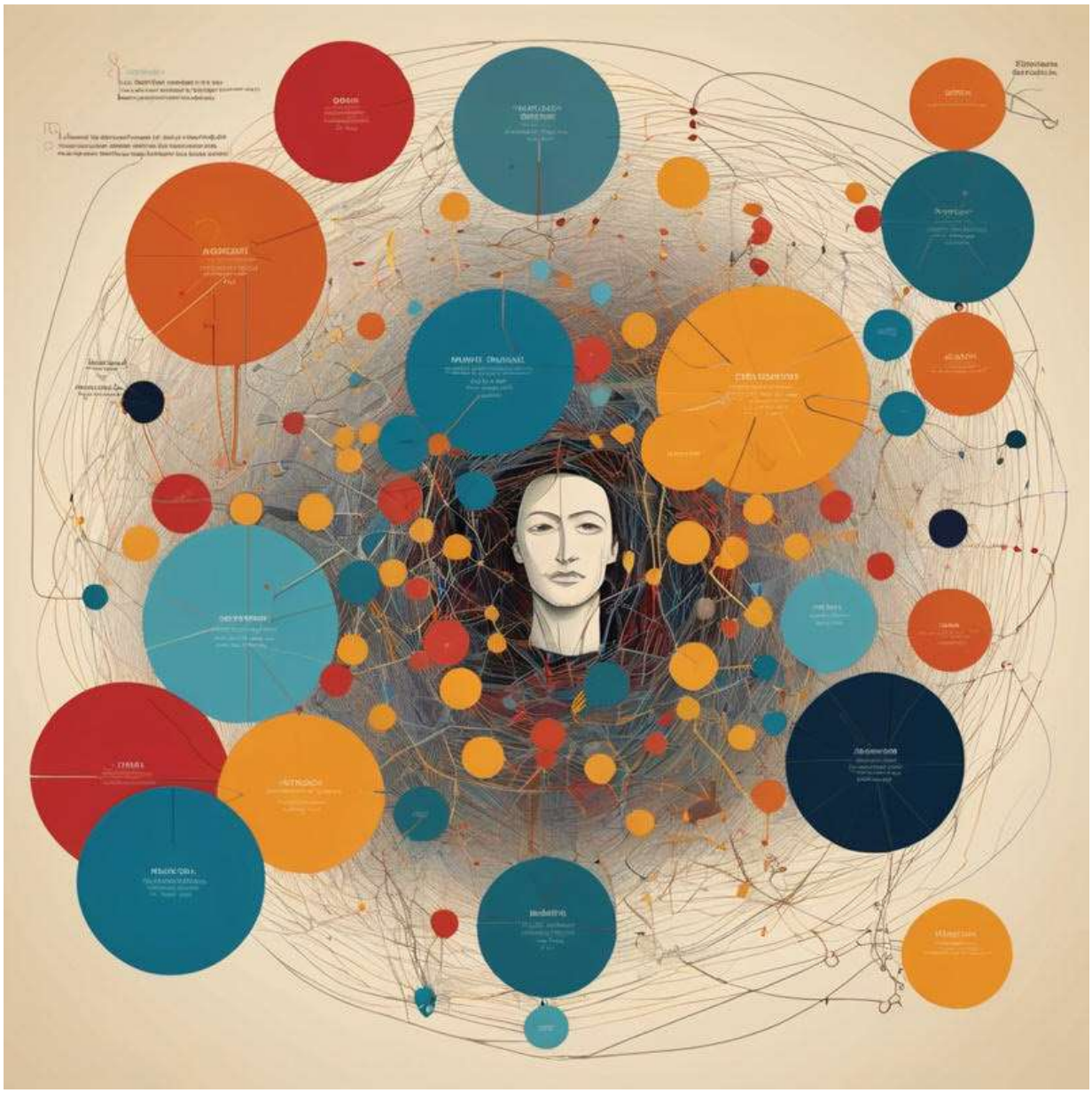
La complexité en sciences du langage

Livret des résumés

12 & 13 décembre 2024
Maison de la recherche
4 rue des irlandais, 75005 Paris



COMITE D'ORGANISATION

Delphine Battistelli (MoDyCo, CNRS-University of Paris Nanterre, France)
Georgeta Cislaru (CLESTHIA, University of Sorbonne Nouvelle, France)
Sascha Diwersy (Praxiling, CNRS-University of Paul-Valéry-Montpellier, France)
Anne Lacheret-Dujour (MoDyCo, CNRS-University of Paris Nanterre, France)
Dominique Legallois (Lattice, CNRS-University of Sorbonne Nouvelle, France)

Comité scientifique

Pierluigi Basso, ICAR, CNRS-Université Lumière Lyon 2
Delphine Battistelli, MoDyCo, CNRS-Université Paris Nanterre
Philippe Blache, LPP, CNRS-Université Aix Marseille
Alice Blumenthal-Dramé, Freiburg Institute for Advanced Studies, Université de Freiburg
Lisa Brunetti, LLF, Université Paris Cité
Georgeta Cislaru, CLESTHIA, Université Sorbonne Nouvelle
Chantal Claudel, MoDyCo, CNRS-Université Paris Nanterre
Sascha Diwersy, Praxiling, CNRS-University of Paul-Valéry-Montpellier
Quentin Feltgen, Université de Gand
François Thomas, Cental, UCLouvain
Núria Gala, LPL, CNRS-Aix Marseille Université
Didier Grandjean, Swiss Center for Affective Sciences, Université de Genève
Karin Heidlmayr, MoDyCo, CNRS-Université Paris Nanterre
Sylvain Kahane, MoDyCo, CNRS-Université Paris Nanterre
Anne Lacheret-Dujour, MoDyCo, CNRS-Université Paris Nanterre
Nicola Lampitelli, MoDyCo, CNRS-Université Paris Nanterre
Frédéric Landragin, Lattice, CNRS
Julie Lefevre, MoDyCo, CNRS-Université Paris Nanterre
Dominique Legallois, Lattice, CNRS-Université Sorbonne Nouvelle
Ewa Lenart, SFL, CNRS-Université Paris 8
Olga Nadvornikova, Université Charles, Prague
Olive, CeRCA, CNRS- Thierry Université de Poitiers
Sophie Prévost, Lattice, CNRS
Florence Villoing, MoDyCo, CNRS-Université Paris Nanterre
Marie-Albane Watine, BCL, CNRS-Université Côte d’Azur
Johannes Ziegler, CRPN, CNRS-Université Aix Marseille

Table des matières

Appel à communications

Si parler, écrire, écouter, lire sont des activités faciles, simples, naturelles pour celles et ceux qui les pratiquent au quotidien, que dire des processus cognitifs et langagiers qui les sous-tendent, des langues dans lesquelles ces activités sont pratiquées, des théories et des modèles développés pour expliquer et représenter les mécanismes en jeu ? Que dire également pour les sujets humains (locuteur, auditeur, scripteur, lecteur) qui n'ont pas encore fini l'apprentissage de l'une et ou l'autre de ces activités (enfant en phase d'acquisition du langage, adulte apprenant une langue seconde), apprentissage qui peut s'avérer particulièrement difficile, voire impossible (sourd écrivant en langue vocale) ?

Très vite la question de la complexité se pose ; la notion est d'ailleurs régulièrement convoquée dans les travaux en sciences du langage mais souvent de façon vague et intuitive. En pratique, cette question de la complexité revêt des modalités différentes en fonction de celle/ou celui qui la pose (psycholinguiste, linguiste, descriptiviste, modélisateur) et de celle/ou celui à qui elle se pose (sujet parlant, sujet percevant, sujet natif, sujet non natif, sujet apprenant, sujet atypique, etc.). Bref, complexité comment, pour qui et pourquoi ? Complexité nécessaire ou contingente ? Pour répondre à ces questions, encore faut-il savoir de quelle complexité on parle : conceptuelle (ex. : représentation du temps et de la référence dans les langues), formelle (ex. : structure phonologique, graphique, morphologique, syntaxique d'une langue) ou physiologique (geste articulatoire peu naturel à produire, contraintes matérielles) ? Une complexité en appelle-t-elle une autre (ex. la conception complexe du temps dans une langue convoque-t-elle une syntaxe complexe, la complexité formelle implique-t-elle la complexité cognitive et inversement ?)

Ces journées proposent de faire un état des lieux sur la complexité en Sciences du langage. Elles seront l'occasion de s'intéresser à l'histoire et l'usage de la notion de complexité en Sciences du langage, à travers divers éclairages théoriques et épistémologiques. Elles ont pour vocation de faire dialoguer linguistes de l'oral et linguistes de l'écrit, talistes et psycholinguistes, etc. autour de la complexité qui traverse, à des degrés variables, les différents composants de la langue et du discours (segmental, suprasegmental, morphologique, syntaxique, sémantique, pragmatique). Elles ont pour objectif de faire émerger à l'issue de cette rencontre un concept opératoire pour la communauté, aussi stratifié soit-il, les critères qui le fondent étant à l'évidence pluriels :

- Pour le linguiste, est complexe ce qui n'est pas simple à représenter et modéliser, parce que (i) peu prédictible (ex. constructions inattendues, productions qui échappent aux règles générales), (ii) de nature continue, et donc difficilement isolable ou catégorisable (ex. niveau suprasegmental vs. niveau segmental ; référence opaque ou indéfinie). Est complexe aussi un observable que l'on peut décrire mais qui résiste à l'explication (ex. les erreurs dans les écrits sourds) ;
- Pour le sujet humain, serait complexe ce qui n'est pas naturel et donc difficile à produire où à entendre (une langue étrangère), ce qui est sous-spécifié linguistiquement, parce qu'ambigu, ou implicite, et occasionnant un coût de traitement élevé.

Mais si pour le linguiste ou le sujet humain, la complexité constitue une difficulté, un obstacle dont on se passerait bien peut-être, pour les langues et leurs usages, la complexité est nécessaire, voire consubstantielle : pas de langue ni de discours sans complexité. Du point de vue synchronique, la complexité participe à la régulation du système linguistique et garantit son équilibre interne, qui reposerait sur une répartition entre éléments simples et complexes (ex. : morphologie pauvre vs. système tonal complexe en chinois). Reste à comprendre comment cet équilibre se construit dans les langues. Sous l'angle diachronique, la complexité semble jouer le même rôle, qu'il s'agisse de simplifier certains processus et de maintenir l'économie du système sur le plan formel (ex. : suppressions des oppositions phonologiques avec un rendement fonctionnel faible, processus de grammaticalisation) ou, au contraire, de réintégrer de la complexité (ex. : passage du pidgin au créole).

Se pose alors une question pour les sciences du langage : comment rendre compte de la complexité linguistique ? Quelle approche adopter ? Typologique et contrastive ou interne, expérimentale, ou inductive sur gros corpus ? Quelle mesure de

complexité et quel étalon de mesure proposer ? Quelle échelle fixer ? Quels descripteurs ? Ainsi en syntaxe, peut-on poser l'existence d'une phrase neutre SVO pour travailler sur des phrases dites complexes ? Les concepts de transformation et de mouvement proposés par la grammaire générative sont-ils opératoires pour travailler sur la complexité syntaxique ? Si oui, comment ? Sinon, par quels descripteurs les remplacer : des descripteurs "aisément" calculables (cf. travaux sur la lisibilité ou la simplification qui y recourent systématiquement) comme la longueur des phrases ou les types de dépendances entre les éléments (ex. : nombre, longueur, direction) ? La question du médium apporte un éclairage encore différent sur la complexité, en particulier pour la composante syntaxique. La structure syntaxique du message est-elle plus complexe à l'oral ou à l'écrit ? Et de quel point de vue ? De la production ou de la réception ? Du point de vue de l'activité de langage ou de sa représentation linguistique et de sa modélisation ? En sémantique et en pragmatique, comment traiter la relation sens-forme ? Comment représenter l'ambiguïté et l'implicite ? Un texte peut-il être simple, étant donné qu'il se construit autour d'unités et de constructions, elles-mêmes complexes. Si oui, comment et par quel mécanisme d'ajustement ou de changement qualitatif ? En linguistique textuelle, la notion de complexité a pu être appréhendée de nombreuses façons ; par exemple, dans le processus d'écriture des textes, par la mesure et la quantification des pauses des scripteurs. Ou encore par les méthodes employées en linguistique appliquée pour la simplification de textes jugés trop difficiles à comprendre et devant être adaptés pour un public particulier. Se pose enfin une question centrale en modélisation : comment gérer la complexité de l'objet que l'on souhaite représenter ? Comment décomposer un objet complexe en éléments simples sans perte d'information ? Comment comprendre quelles sont les propriétés nécessaires et suffisantes pour représenter le fonctionnement du système ? Comment aborder la question de l'articulation entre eux de multiples descripteurs à l'aide de formules mathématiques allant plus avant que les formules proposées dans le champ de la lisibilité ? Parmi les descripteurs, on pourra par exemple s'intéresser à des unités ou des relations de dépendance en syntaxe, à des contours ou à des tons dans les langues à prosodie accentuelle, à des opérations sous-jacentes à la description de la sémantique d'unités lexicales et grammaticales, à des opérations de référenciation à des espaces différents de validation de contenus prédicatifs (ex. des espaces hypothétiques), ou encore à des types de relations entre unités textuelles (ex. : enchâssement, inclusion, succession). On pourra également questionner, du point de vue de l'acquisition des compétences, la mise en corrélation des structures linguistiques et des étapes du développement cognitif de l'individu.

Conférences invitées

La complexité au cœur de l'émotion

Didier Grandjean
Faculté de psychologie et des sciences de l'éducation
Centre interfacultaire en sciences affectives
Université de Genève

Les processus émotionnels peuvent être conceptualisés de manières différentes selon les modèles théoriques proposés, lors de cet exposé je présenterai ces différents modèles et les appliquerai aux situations dans lesquelles une parole est produite lors d'un épisode émotionnel. Ces modulations émotionnelles d'une production de parole seront conceptualisées à travers la notion de prosodie émotionnelle et sa perception. Différents niveaux de complexité seront discutés à l'aune de ce concept de prosodie émotionnelle : premièrement les mécanismes psychologiques à l'œuvre dans la construction des percepts basés sur les paramètres acoustiques et les mécanismes de codage prédictif ; deuxièmement les corrélats cérébraux associés à ces mécanismes psychologiques qui nous mèneront à discuter de la complexité des processus cérébraux de manière plus générale et les différents niveaux de codage neuronaux à l'œuvre dans la perception, la signification et l'interprétation des percepts.

Le premier niveau de complexité discuté référant aux mécanismes de perception de prosodie émotionnelle mettra en évidence les relations non linéaires et donc complexes entre les dimensions acoustiques physiques, comme par exemple les dynamiques temporelles de la F0 ou de l'énergie, et la construction des percepts liés comme la perception de la hauteur ou pitch et le volume perçu.

Le deuxième niveau de complexité réfèrera aux mesures cérébrales et aux inférences liées lors de la perception de l'émotion vocale. Un premier volet portera sur les résultats empiriques en imagerie cérébrale à résonance magnétique fonctionnelle et les méthodes d'analyses utilisées dans ce contexte. La discussion se poursuivra sur d'autres mesures cérébrales portant sur les champs électriques et magnétiques générés par les populations neuronales. Ce deuxième volet permettra de mettre en exergue les différents processus électro-magnétiques et les analyses reliées telles que les potentiels évoqués, les analyses temps fréquences et les analyses portant sur les relations entre populations neuronales comme les mesures de synchronisations neuronales qu'elles soient basées sur les relations, par exemple, en termes d'énergie ou de phase ou la combinaison de ces deux grandeurs.

Finalement je présenterai un modèle théorique intégratif de ces différents résultats en imagerie à résonance magnétique fonctionnelle et en électro-magnéto-encéphalographie sur les mécanismes de perception de l'émotion vocale et ses effets sur les interactions sociales.

Longueurs des dépendances ou flux de dépendances : deux mesures de complexité syntaxique et deux façons de voir les contraintes sur la mémoire à court terme

Sylvain Kahane
MoDyco, Université Paris Nanterre & CNRS / IUF

Ce travail résulte d'une collaboration avec Chunxiao Yan (Kahane et al. 2017, Kahane & Yan 2019, Yan 2021). Parmi les mesures de la complexité syntaxique, une mesure simple a fait l'objet de nombreuses études : la longueur moyenne des dépendances syntaxiques (Hudson 1995, Gibson 2001, Liu 2008).

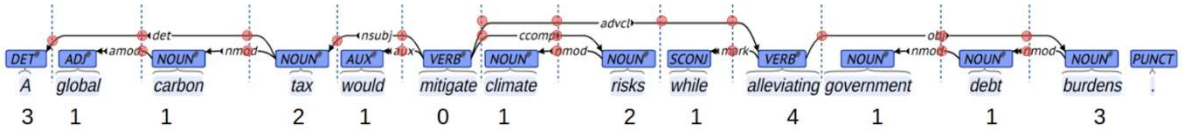


Figure 1. Longueur de dépendances

Chaque dépendance est représentée par une flèche allant du gouverneur au dépendant.
La longueur de chaque dépendance est indiquée sous le dépendant.

Il a pu être montré que les langues naturelles tendent à minimiser les longueurs des dépendances (DLM : Dependency Length Minimization). Cette propriété permet elle-même d'expliquer d'autres propriétés comme la tendance des langues à respecter la projectivité, c'est-à-dire à placer les dépendants au plus proche de leur gouverneur et donc à réaliser des constituants continus (les constituants étant les projections des têtes lexicales) (Ferrer i Cancho 2006). La DLM explique aussi la tendance à placer, à droite d'une tête, les constituants les plus courts avant les plus longs.

Depuis qu'on s'intéresse à la complexité, on a mis les contraintes sur la complexité des énoncés en rapport avec les limitations de nos capacités cognitives et notamment de la mémoire à court terme ou mémoire de travail. La DLM s'expliquerait alors par la nécessité de minimiser le temps de tenue en mémoire des informations traitées par la mémoire de travail.

Nous allons maintenant voir une autre explication possible de la DLM. Nous appelons flux de dépendance en un point de l'énoncé (essentiellement entre deux mots) l'ensemble des dépendances syntaxiques qui relient un mot avant ce point à un mot après (Kahane 2001). La taille du flux est le nombre de dépendances dans ce flux. Il se trouve que la taille moyenne du flux dans un énoncé est nécessairement égale à la taille moyenne des longueurs des dépendances. Il suffit pour s'en convaincre de compter le nombre de points d'intersection entre une dépendance et une position inter-mot. Ce nombre est à la fois égal à la somme des longueurs de dépendance (puisque la longueur d'une dépendance est le nombre de positions inter-mots qu'elle franchit) et à la somme des tailles des flux.

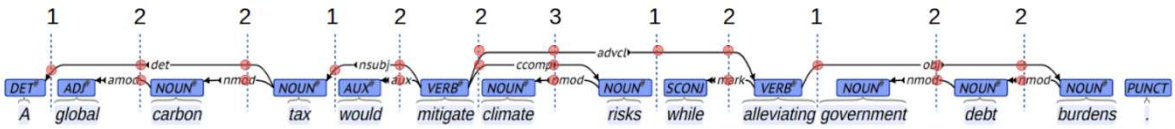


Figure 2. Flux de dépendances
Chaque flux inter-mot est symbolisé par un trait vertical.
La taille du flux est indiquée au-dessus.

Autrement dit, la DLM est en fait une minimisation de la taille moyenne du flux. Mais minimiser la taille, cela signifie, en termes de mémoire de travail, minimiser le nombre d'informations tenues simultanément en mémoire. Ceci rejoint les travaux de Miller (1956) qui avait montré, par diverses expériences, que la mémoire de travail est limitée à 7±2 éléments, chiffre aujourd'hui réactualisé autour de 4 selon Cowan (2001).

Le fait que les longueurs de dépendance et les tailles de flux ait la même moyenne ne signifie pas du tout qu'il se distribue de la même façon. La plupart des dépendances sont de longueur 1, mais certaines dépendances peuvent être très longues. A l'inverse les flux de taille 2 sont plus nombreux que les flux de taille 1 et le flux dépasse difficilement la dizaine (cela dépend aussi de la façon dont on définit les dépendances). Cela nous laisse penser que la contrainte est davantage sur le flux que sur la longueur des dépendances.

On peut maintenant se poser une question : la taille du flux est-elle vraiment la bonne mesure pour évaluer la complexité du flux ? Quand on regarde le flux, on note qu'il peut y avoir des configurations assez diverses. Deux configurations méritent d'être contrastées : les bouquets de dépendances, c'est-à-dire les configurations où plusieurs dépendances partagent une même extrémité ; les dépendances disjointes, c'est-à-dire les configurations où les dépendances ne partagent aucune extrémité. Notre hypothèse que les dépendances disjointes sont plus coûteuses pour la mémoire de travail, car les bouquets permettent de factoriser l'information sur l'extrémité commune. Nous introduisons une nouvelle mesure que nous appelons le poids du flux, c'est-dire le nombre maximal de dépendances disjointes dans le flux. Alors que la taille du flux ne semble pas bornée de manière évidente (même si elle l'est nettement plus que la longueur des dépendances), le poids du flux est fortement borné et nous n'avons relevé aucun exemple dépassant un poids de 5 dans les treebanks Universal Dependencies (12 millions de mots dans plus de 100 langues quand nous avons fait les calculs).

Bibliographie :

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1), 87-114.

Ferrer i Cancho, Ramon (2006). Why do syntactic links not cross?*, EPL (Europhysics Letters),* 76(6), 1228.

Gibson, Edward (1998) Linguistic complexity: Locality of syntactic dependencies. *Cognition,* 68(1), 1-76.

Hudson, Richard (1995) Measuring syntactic difficulty. Manuscript, University College, London.

Kahane, Sylvain (2001) Grammaires de dépendance formelles et théorie Sens-Texte, Tutoriel, *Actes de TALN*.

Kahane, Sylvain, Chunxiao Yan, Marie-Amélie Botalla (2017) What are the limitations on the flux of syntactic dependencies? Evidence from UD treebanks. *Conference on Dependency Linguistics (Depling)*, ACL.

Kahane, Sylvain, Chunxiao Yan (2019) Advantages of the flux-based interpretation of dependency length minimization. *First international conference on Quantitative Syntax (Quasy)*, ACL.

Liu, Haitao (2008) Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science,* 9(2), 159-191.

Miller, George A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review,* 63(2).

Yan, Chunxiao (2021) *Complexité syntaxique et flux de dépendance : études quantitatives dans les treebanks Universal Dependencies*, Thèse de l'Université Paris Nanterre.

Évaluer automatiquement la complexité textuelle : quels défis reste-t-il après 101 ans de recherches en lisibilité ?

Thomas François
Cental, IL&C, UClouvain

La lisibilité est un domaine situé aujourd'hui à l'intersection des sciences de l'éducation et du traitement automatique des langues dont l'objectif principal est d'évaluer automatiquement la complexité des textes pour une population donnée. Apparu dans les années 1920 aux États-Unis avec les travaux de Lively et Pressey (1923), le domaine a depuis connu de nombreux développements, donnant naissance à plusieurs centaines de formules de lisibilité (cf. Dubay, 2004 et François, 2011 pour une description des travaux du 20e siècle). Depuis le début du 21e siècle, l'évaluation automatique de la lisibilité des textes a connu des avancées significatives qui s'appuient sur le développement du traitement automatique du langage, de l'apprentissage automatisé et, plus récemment, de l'apprentissage profond. Ces modèles sont désormais capables de prendre en considération un nombre beaucoup plus riche de caractéristiques linguistiques liées à la complexité des textes et de produire des prédictions plus fiables (cf. Collins-Thompson, 2014 ; François, 2015 ; Vajjala, 2022 pour des synthèses des avancées récentes).

Dans notre communication, nous allons tout d'abord présenter un aperçu historique du domaine en mettant en lumière les dimensions linguistiques qui ont été successivement prises en compte pour mesurer la complexité de lecture des textes. Nous illustrerons les tendances actuelles en nous appuyant sur nos propres travaux, d'une approche basée exclusivement sur des variables linguistiques (François et Fairon, 2012 ; Wilkens et al., 2022) à l'emploi des larges modèles de langue (Yancey et al., 2021, Battistelli et al., 2022) en passant par des approches hybrides reposant à la fois sur l'apprentissage profond et des variables linguistiques (Deutsch et al., 2020 ; Lee et al., 2021 ; Wilkens et al., 2024).

Dans une seconde partie, nous adopterons une perspective critique. En effet, après plus d'un siècle de recherche en lisibilité, il apparaît opportun de s'interroger sur les acquis et les limites du domaine. Il est indéniable que les travaux classiques, dont le plus emblématique est la formule de Flesch (1948), eurent un impact sur la société américaine, contribuant, par exemple à abaisser le niveau de difficulté moyen des journaux américains d'un niveau universitaire à un niveau de fin de secondaire dans les années 1950 (Klare, 1963). De même, les travaux de Kincaid et al. (1973) ont conduit à la création de la formule de Flesch-Kincaid, largement utilisée dans l'armée américaine. Toutefois, il est préoccupant de constater que ces indices historiques, notamment les formules de Flesch et de Flesch-Kincaid, dominent encore largement les pratiques actuelles. Prenons pour exemple l'analyse de la lisibilité de la communication financière. Selon Le Maux (2015), les études menées entre 1950 à 2011 montrent une absence totale d'évolution dans les formules utilisés : Flesch et Fog (Gunning, 1952) dominent la scène en 1950 comme en 2011. Pire encore, dans le domaine de la simplification automatique de textes, qui relève pourtant du TAL comme la lisibilité, les évaluations automatisées de la difficulté des textes produits par les systèmes de simplification est toujours majoritairement réalisée avec la formule de Flesch-Kincaid, malgré des critiques (Tanprasert et Kauchak, 2021).

Plusieurs problèmes freinent aujourd'hui les avancées dans le domaine. Tout d'abord, l'approche basée sur le TAL exige des corpus d'entraînement considérablement plus vastes que ceux requis pour les formules traditionnelles. Alors que l'annotation de la difficulté des textes dans la recherche classique sur la lisibilité reposait sur des données de lecteurs, la lisibilité exploitant le TAL a abandonné la collecte de données auprès des lecteurs au profit de textes pédagogiques, dont la difficulté est alors évaluée par des experts (François, 2011). Cela a conduit à un glissement : au lieu de modéliser directement la compréhension en lecture – et cela bien que les dispositifs utilisées pour mesurer celle-ci, comme les tests à trous ou les questions à choix multiples, présentent des limites –, les modèles récents modélisent plutôt les jugements d'experts. De plus, au regard de la quantité d'efforts nécessaires pour constituer un corpus d'entraînement, les corpus actuels sont généralement très homogènes, peu nombreux et principalement disponibles pour l'anglais (WeeBit par Vajjala et Meurers, 2012 ; Newsela par Xu et al., 2015 ; OneStopEnglish par Vajjala et Lucic, 2018 ou CLEAR par Crossley et al., 2022). Dès lors, le risque que les prouesses atteintes par les formules computationnelles ne se généralisent pas à l'ensemble des textes théoriquement ciblés par ces modèles est bien réel. Ainsi, Nelson et al. (2012) ont souligné que ces modèles computationnels pourraient ne pas se comporter mieux que les formules traditionnelles.

Un second défi est lié à la nature intrinsèquement subjective de la lecture. La lisibilité vise pourtant à émettre un jugement à visée généraliste sur la difficulté des textes. Il y a donc là contradiction intrinsèque, un va-et-vient entre le particulier et le général qui se justifie par des considérations pratiques, à savoir que, pendant longtemps, les formules de lisibilité ont été

calculées manuellement. Avec les progrès technologiques, il devient toutefois envisageable de personnaliser davantage les prédictions, comme cela a été démontré dans le cadre de la tâche d'identification des mots complexes (Tack et al., 2016 ; Lee et Yeung, 2018 ; Gooding et Tragut, 2022).

Pour conclure, il est clair que le domaine de la lisibilité a considérablement évolué depuis ses débuts et les formules classiques des années 1940. Néanmoins, plusieurs défis demeurent, tant au niveau des recherches fondamentales que de l'adoption des résultats par le grand public. Les avancées récentes, notamment dans le TAL et l'apprentissage profond, ouvrent des perspectives prometteuses, mais il est essentiel de continuer à interroger les bases méthodologiques et à diversifier les approches pour mieux répondre aux besoins des utilisateurs.

Corpus-derived complexity measures and online language processing: Does one size fit all languages?

Alice Blumenthal-Dramé
English department of the University of Freiburg

This talk will explore the notion of complexity from three complementary perspectives: corpus linguistics, cognitive linguistics, and typology.

The first, more theoretical half will introduce several competing definitions of complexity currently in use. Building on this foundation, I will outline and illustrate a number argumentative patterns in the literature that link corpus linguistics, typology, and cognitive linguistics (Ehret et al., 2021, 2023). I will argue that some of these patterns rely on unwarranted or oversimplified assumptions. For example, it is often assumed that corpus-derived complexity metrics for a language necessarily reflect cognitive processing costs for its users. Another assumption is that the same metric correlates with identical cognitive effects across languages, implying that speakers do not adapt their processing strategies to the typological make-up of their language.

Following this critique, I will argue for the importance of benchmarking—establishing systematic associations between corpus-derived complexity metrics and online language processing experiments across languages.

The second half of the talk will put this approach into practice by presenting an experimental study comparing word order complexity in Armenian, Georgian, and Russian (Forker, Sazhumyan & Blumenthal-Dramé in preparation). First, complexity will be predicted using corpus-derived frequencies of features such as heaviness, definiteness, and animacy in relation to word order in each language. These predictions will then be compared to actual processing complexity for different feature combinations. The results will demonstrate that:

The same predictors do not have consistent effects across languages.

The alignment between corpus-derived predictions and cognitive data varies by language.

The final section will discuss methodological and practical challenges in conducting cross-linguistic benchmarking research on complexity. I will conclude with reflections on the comparability of languages in terms of their complexity.

Bibliography:

Ehret, K., Berdicevskis, A., Bentz, C., & Blumenthal-Dramé, A. (2023). Measuring language complexity: Challenges and opportunities. Linguistics Vanguard, 0(0). https://doi.org/10.1515/lingvan-2022-0133

Ehret, K., Blumenthal-Dramé, A., Bentz, C., & Berdicevskis, A. (2021). Meaning and Measures: Interpreting and Evaluating Complexity Metrics. Frontiers in Communication, 6, 640510 https://doi.org/10.3389/fcomm.2021.640510

Forker, D., Sazhumyan, H., & Blumenthal-Dramé, A. (in preparation). Approaching word order in Georgian, Armenian and Russian.

Communications

Levels of Complexity in Literary Language: A Preliminary Study

Pascale Feldkamp
Center for Humanties Computing
Aarhus University pascale.moreira@cc.au.dk

Yuri Bizzoni
School of Culture and Communication
Aarhus University yuri.bizzoni@cc.au.dk

Literary texts represent examples of language operating at its most virtuous and demanding: they are capable of generating an 'experience' (Starr, 2013; Girju and Lambert, 2021) – often emotional or evocative (Bizzoni and Feldkamp, 2024; Miall and Kuiken, 1994) – through the sheer force of words. In this domain, the capacity of language to evoke emotions, construct worlds, and shape experience is pushed to its limits. Computational literary analyses have for some time used both stylistic and syntactic aspects of text that can be linked to complexity to try distinguishing the textual profiles of different types of fiction. For example, a handful of studies focused on canonic literature (Barré et al., 2023; Brottrager et al., 2022; Wu et al., 2024; Algee-Hewitt et al., 2016) have recently shown that despite their great diversity, canonic texts tend to exhibit a high level of complexity under several respects: they have a denser nominal style (Wu et al., 2024), lower readability levels, higher language-model-based perplexities, as well as less predictable sentiment arcs (Bizzoni et al., 2023b). However, few studies have gone beyond measures of style and syntax to conceptualize what makes a literary text complex. The focus on stylistic and syntactic features is rooted in a long formalist tradition of literary theory, which holds that literary texts distinguish themselves through stylistic discomfort. For this approach, the 'literariness' of texts lies in their use of language to create linguistic strangeness – a "foregrounding" that slows down the reading process (Mukaˇrovský, 1964).
An especially overlooked aspect in the analysis of complexity in the literary use of language may in fact be found beyond the stylistic/syntactic level, at the level of the feelings expressed and evoked in texts. Research in psycholingustics has recently placed emphasis on the effect of sentiment in texts, finding that readers quickly respond to the valence of words (Pfeiffer et al., 2020), and that non-neutral words (Lei et al., 2023) and negative valence in stories (Arfé et al., 2023) to an extent increases processing time. Complexity at this level is difficult to define. While a metric like simple sentiment standard deviation can be used to gauge the width of the 'emotional palette' that authors are using in a novel, some more sophisticated measures for the complexity of sentiment arcs in novels have been developed in recent years, like the approximate entropy or the Hurst exponent of sentiment arcs (Bizzoni et al., 2021, 2022). Very little work has explored the connection between these different levels of complexity: the relation between complexity at the stylistic level, and complexity at the emotional level – e.g., as effected by more complex sentiment trajectories across a novel. We formulate two alternative hypotheses:
H1. Complexity at the stylistic and syntactic level requires a "simplification" at other levels of the narrative text. In this case, meaning a reduced sentimental complexity.
H2. Complexity at the stylistic and syntactic level is linked to an enhanced complexity at other linguistic levels of the narrative texts as well, in this case, for example, at the sentimental level.1

We argue that the two hypotheses carry different consequences. H1 stems from the idea, observed in other domains, that for optimized communication, readers' cognitive loads have to be distributed through different linguistic layers. From this 'cognitive compensation' hypothesis follows that, for example, an increased lexical complexity links to a simplification of syntactic structures, and so forth (Degaetano-Ortlieb and Teich, 2022).
H2 comes from the idea that complexity in literary expression tends to interest different layers at the same time: works that "dare" using more complex syntax will also reach into more challenging sentiment profiles. This scenario also carries an interesting possibility, that for the literary or artistic use of language, the complexity at one level could be functional for the complexity at another level, for example syntactic complexity (longer sentences, less predictable structures etc.) could link up with sentimental complexity. This may also further suggest that literary texts function as 'supernormal stimuli', intentionally amplifying complexity across levels to heighten engagement and elicit amplified responses (Dubourg and

[1] Naturally, the hypothesis zero (H0) would be that the levels have no connection with each other.

Baumard, 2022; Costa and Corazza, 2006). Naturally, such scenario might entail a much higher cognitive load for average readers. We carry out a preliminary study in the relation between stylistic and syntactic features often employed in computational literary analysis, and sentiment features of texts.

Data & features

Our data is the Chicago Corpus of 9,089 novels in the time period 1880-2000. The novels are predominantly by anglophone authors, and the corpus was compiled based on the number of libraries worldwide that hold the novel, with a preference for higher holdings. As such, since library holdings reflect both popular demand and more prestigious curated literature, the corpus spans various genres, from Agatha Christie to James Joyce.2

Features used in this study were used in other studies seeking to distinguish textual profiles of different types of literature. The details on each measure can be found in the Appendix (Table 2). Here, we focused on features that supposedly reflect some stylistic or syntactic complexity. For features at the sentiment level that have been used in previous studies (Bizzoni et al., 2023b, 2022), we choose to focus on measures of local and global complexity of the sentiment arc. As such, we first annotated all novels at the sentence level for sentiment valence (where 1 represents the positive and -1 the negative polarity) using the Syuzhet package (Jockers, 2015) – an SA tool developed explicitly for literary language, which has shown the best performance for English in the literary domain compared to transformer-based models (Bizzoni et al., 2023a). We then calculated the standard deviation, approximate entropy, and Hurst exponent of sentiment arcs for all novels – taking these features to represent the variance, as well as the local and global predictability – in other words, complexity – of novels sentiment profiles in this preliminary experiment. We proceed to juxtapose stylistic/syntactic and sentiment features of complexity across all novels, gauging the correlation between them. We then assess the link between stylistic/syntactic and the sentiment level by trying to predict individual sentiment variables using all the stylistic/syntactic features.

Feature	F-stat	R2	adj. R2
Sentiment SD	1803.0	0.787	0.786
ApEn	364.2	0.427	0.426
Hurst	123.1	0.201	0.2

Table 1: Linear regression: predicting sentiment features based on stylistic/syntactic features. Here for all, $p < 0.01$. The R2 (and adjusted R2) score represents the proportion of variance in the predicted sentiment variable explained by the stylistic and syntactic features.

Results and Conclusions

We show a high correlation between sentimentlevel features and some of the stylistic/syntactic features in Fig. 1. Especially readability formulas, word- and sentence length, dependency length, lexical richness ('MSTTR'), indicators of heavy nominal style ('freq of' and 'nominal verb ratio') and LLM perplexity – i.e., features that are strongly associated to harder-to-process and more informationrich text – show a particularly strong correlation with sentiment SD. Approximate entropy seems to show some of the same patterns, while less correlated with, e.g., LLM perplexity. The more global uncertainty measured with the Hurst exponent also appears related to these complexity metrics.Our results support H2. Rather than a balance between different aspects of language, we find that complexity at one level tends to correspond to complexity at another. As mentioned in the introduction, increased stylistic and syntactic complexity might enhance the expressive ability of the text, "allowing" it to communicate a more extreme sentimental profile, without the obvious need for a balance between different levels of linguistic expression.

Moreover, testing linear regression, we find that complexity at the sentimental level – most strongly in the case of sentiment SD – appears so tightly linked with textual complexity metrics, that the latter work well as a model to predict the former (Table 1). See Fig. 2 in Appendix for a visualization of the predicted and actual values in sentiment SD. In the future, we intend to explore the relationship between these levels of complexity in literary language further, better formalizing the relation

[2] See Bizzoni et al. (2024b) for details on the corpus, used in various recent studies, like Wu et al. (2024). The corpus has been recently made available at https://github.com/ centre-for-humanities-computing/chicago_corpus.

and role of each of the selected components. We would also examine the relationship between perceived complexity or difficulty of a text and these features in an experiment setting.

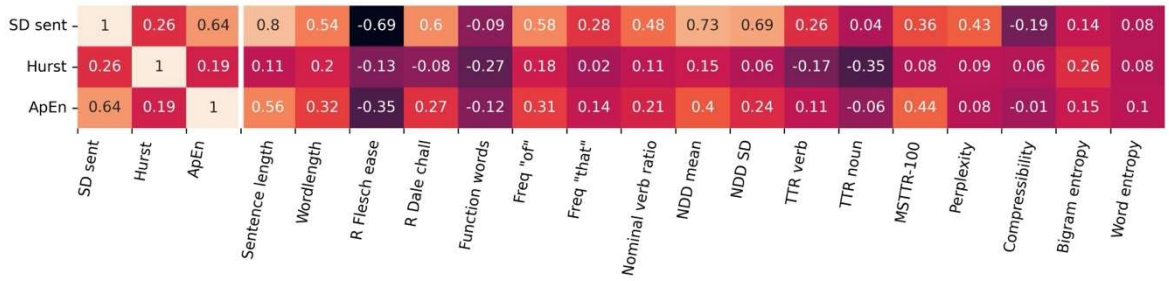


Figure 1: The correlation (Spearman's ρ) between stylistic/syntactic features and sentiment features. See table 2 in Appendix for details on the computation of these features and for the label explanations.

Bibliography:

Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser. 2016. Canon/Archive. Large-scale Dynamics in the Literary Field. Stanford Literary Lab.

Barbara Arfé, Pablo Delatorre, and Lucia Mason. 2023. Effects of negative emotional valence on readers' text processing and memory for text: an eye-tracking study. Reading and Writing, 36(7):1743–1768.

Jean Barré, Jean-Baptiste Camps, and Thierry Poibeau. 2023. Operationalizing Canonicity: A Quantitative Study of French 19th and 20th Century Literature. Journal of Cultural Analytics, 8(3).

Jonah Berger, Yoon Duk Kim, and Robert Meyer. 2021. What Makes Content Engaging? How Emotional Dynamics Shape Success. Journal of Consumer Research, 48(2):235–250.

Yuri Bizzoni and Pascale Feldkamp. 2024. Below the sea (with the sharks): Probing textual features of implicit sentiment in a literary case-study. In Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language, pages 54–61, Malta. Association for Computational Linguistics.

Yuri Bizzoni, Pascale Feldkamp, Ida Marie Lassen, Mia Jacobsen, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2024a. Good Books are Complex Matters: Gauging Complexity Profiles Across Diverse Categories of Perceived Literary Quality. ArXiv:2404.04022 [cs].

Yuri Bizzoni, Pascale Feldkamp Moreira, Emily Öhman, and Kristoffer L. Nielbo. 2023a. Comparing Transformer and Dictionary-based Sentiment Models for Literary Texts: Hemingway as a Case-study. In NLP4DH (forthcoming), Tokyo, Japan.

Yuri Bizzoni, Pascale Moreira, Nicole Dwenger, Ida Lassen, Mads Thomsen, and Kristoffer Nielbo. 2023b. Good reads and easy novels: Readability and literary quality in a corpus of US-published fiction. In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), pages 42–51, Tórshavn, Faroe Islands. University of Tartu Library.

Yuri Bizzoni, Pascale Moreira, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2023c. Sentimental matters - predicting literary quality by sentiment analysis and stylometric features. In Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 11–18, Toronto, Canada. Association for Computational Linguistics.

Yuri Bizzoni, Pascale Feldkamp Moreira, Ida Marie S. Lassen, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2024b. A matter of perspective: Building a multi-perspective annotated dataset for the study of literary quality. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LRECCOLING 2024), pages 789–800, Torino, Italia. ELRA and ICCL.

Yuri Bizzoni, Pascale Feldkamp Moreira, Mads Rosendahl Thomsen, and Kristoffer L. Nielbo. 2023d. The fractality of sentiment arcs for literary quality assessment: the case of nobel laureates. Journal of Data Mining & Digital Humanities, NLP4DH.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022. Fractality of sentiment arcs for literary quality assessment: The case of nobel laureates. In Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities, pages 31–41, Taipei, Taiwan. Association for Computational Linguistics.

Yuri Bizzoni, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2021. Sentiment dynamics of success: Fractal scaling of story arcs predicts reader preferences. In Proceedings of the Workshop on Natural Language Processing for Digital Humanities, pages 1–6, NIT Silchar, India. NLP Association of India (NLPAI).
Lloyd R. Bostian. 1983. How active, passive and nominal styles affect readability of science writing. Journalism quarterly, 60(4):635–670.
Judith Brottrager, Annina Stahl, Arda Arslan, Ulrik Brandes, and Thomas Weitin. 2022. Modeling and predicting literary reception. Journal of Computational Literary Studies, 1(1):1–27.
Davida H. Charney and Jack R. Rayman. 1989. The Role of Writing Quality in Effective Student Résumés. Journal of Business and Technical Communication, 3(1):36–53. Publisher: SAGE Publications Inc.
Marco Costa and Leonardo Corazza. 2006. Aesthetic Phenomena as Supernormal Stimuli: The Case of
Eye, Lip, and Lower-Face Size and Roundness in Artistic Portraits. Perception, 35(2):229–246. Publisher: SAGE Publications Ltd STM.
Scott A. Crossley, Rod Roscoe, and Danielle S. Mc-Namara. 2014. What Is Successful Writing? An Investigation Into the Multiple Ways Writers CanWrite Successful Essays. Written Communication, 31(2):184–214. Publisher: SAGE Publications Inc.
Stefania Degaetano-Ortlieb and Elke Teich. 2022. Toward an optimal code for communication: The case of scientific english. Corpus Linguistics and Linguistic Theory, 18(1):175–207.
Edgar Dubourg and Nicolas Baumard. 2022. Why and How Did Narrative Fictions Evolve? Fictions as Entertainment Technologies. Frontiers in Psychology, 13. Publisher: Frontiers.
Katharina Ehret and Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. In Complexity, Isolation, and Variation, pages 71–94. De Gruyter.
Gerardo Febres and Klaus Jaffe. 2017. Quantifying literature quality using complexity criteria. Journal of Quantitative Linguistics, 24(1):16–53. ArXiv:1401.7077 [cs].
Richard S. Forsyth. 2000. Pops and flops: Some properties of famous english poems. Empirical Studies of the Arts, 18(1):49–67.
Craig L. Garthwaite. 2014. Demand spillovers, combative advertising, and celebrity endorsements. American Economic Journal: Applied Economics, 6(2):76– 104.
Roxana Girju and Charlotte Lambert. 2021. Inter-Sense: An Investigation of Sensory Blending in Fiction. ArXiv:2110.09710 [cs].
Qiyue Hu, Bin Liu, Mads Rosendahl Thomsen, Jianbo Gao, and Kristoffer L Nielbo. 2020. Dynamic evolution of sentiments in Never Let Me Go: Insights from multifractal theory and its implications for literary analysis. Digital Scholarship in the Humanities, 36(2):322–332.
Arthur M. Jacobs and Annette Kinder. 2022. Computational analyses of the topics, sentiments, literariness, creativity and beauty of texts in a large Corpus of English Literature. ArXiv:2201.04356 [cs].
Matthew L Jockers. 2015. Syuzhet: Extract sentiment and plot arcs from text. Matthew L Jockers blog.
Justine Kao and Dan Jurafsky. 2012. A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. In Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature, pages 8–17, Montréal, Canada. Association for Computational Linguistics.
Corina Koolen, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout. 2020. Literary quality in the eye of the Dutch reader: The national reader survey. Poetics, 79:1–13.
Anqi Lei, Roel M. Willems, and Lynn S. Eekhof. 2023. Emotions, fast and slow: processing of emotion words is affected by individual differences in need for affect and narrative absorption. Cognition and Emotion, 37(5):997–1005.
Lei Lei and Matthew L. Jockers. 2020. Normalized Dependency Distance: Proposing a New Measure. Journal of Quantitative Linguistics. Publisher: Routledge.
Suraj Maharjan, John Arevalo, Manuel Montes, Fabio A. González, and Thamar Solorio. 2017. A multi-task approach to predict likability of books. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1217–1227, Valencia, Spain. Association for Computational Linguistics.
Claude Martin. 1996. Production, content, and uses of bestselling books in quebec. Canadian Journal of Communication, 21(4).
Carey McIntosh. 1975. Quantities of qualities: Nominal style and the novel. Studies in Eighteenth-Century Culture, 4(1):139–153.

David S. Miall and Don Kuiken. 1994. Foregrounding, defamiliarization, and affect: Response to literary stories. Poetics, 22(5):389–407.
Mahdi Mohseni, Volker Gast, and Christoph Redies. 2021. Fractality and variability in canonical and noncanonical english fiction and in non-fictional texts. 12.
Mahdi Mohseni, Christoph Redies, and Volker Gast. 2022. Approximate entropy in canonical and noncanonical fiction. Entropy, 24(2):278.
Jan Mukaˇrovský. 1964. Standard language and Poetic
Language. In Paul L. Garvin, editor, A Prague School Reader on Esthetics Literary Structure, and Style, pages 17–30. 1932. Georgetown University Press.
Christian Pfeiffer, Nora Hollenstein, Ce Zhang, and Nicolas Langer. 2020. Neural dynamics of sentiment processing during naturalistic sentence reading. NeuroImage, 218:116934.
Emily Sheetz. 2018. Evaluating Text Generated by Probalsitic Language Models.
Sanja Stajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity? In Proceedings of Workshop on natural language processing for improving textual accessibility, pages 14–22, Istanbul, Turkey. Association for Computational Linguistics.
G. Gabrielle Starr. 2013. Feeling Beauty: The Neuroscience of Aesthetic Experience. The MIT Press.
Joan Torruella and Ramon Capsada. 2013. Lexical statistics and tipological structures: A measure of lexical richness. Procedia - Social and Behavioral Sciences, 95:447–454.
Andreas van Cranenburgh and Rens Bod. 2017. A dataoriented model of literary language. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1228–1238, Valencia, Spain. Association for Computational Linguistics.
Yaru Wu, Yuri Bizzoni, Pascale Moreira, and Kristoffer Nielbo. 2024. Perplexing canon: A study on GPTbased perplexity of canonical and non-canonical literary works. In Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), pages 172–184, St. Julians, Malta. Association for Computational Linguistics.
Claire M. Zedelius, Caitlin Mills, and Jonathan W. Schooler. 2019. Beyond subjective judgments: Predicting evaluations of creative writing from computational linguistic features. Behavior Research Methods, 51(2):879–894.

Appendix

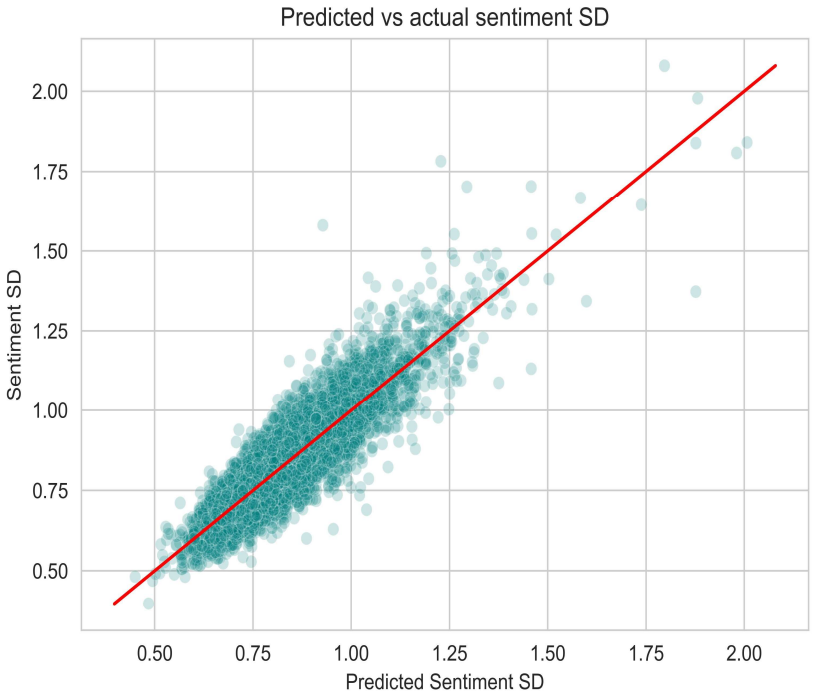


Figure 2: Predicted and actual sentiment SD in the linear regression based on the stylistic/syntactic complexity features.

Feature	Description	Type	Reference
Type-Token Ration (MSTTR-100), TTR Noun, TTR Verb	Measures lexical diversity by comparing the variety of Stylistic Ratio words (types) to the total number of words (tokens), indicating a text's vocabulary complexity and inner diversity. A high TTR represents a richer prose: a higher diversity of elements and a lower lexical redundancy (Torruella and Capsada, 2013). TTR of nouns or of verbs quantifies diversity within these Parts-of-Speech categories.[a]	Stylistic	Forsyth (2000)*, Kao and Juraf- sky (2012)*, Algee-Hewitt et al. (2016), Maharjan et al. (2017), Koolen et al. (2020), Brottrager et al. (2022), Jacobs and Kinder (2022), Bizzoni et al. (2023c)
Readability (R Flesch Ease, R Dale Chall)	Estimate reading difficulty based variously on sentence length, syllable count, and word length/difficulty. Assessed using five different classic formulae that remain widely used (Stajner et al., 2012).[b]	Stylistic	Martin (1996), Garthwaite (2014), Maharjan et al. (2017), Febres and Jaffe (2017), Zedelius et al. (2019)*, Berger et al. (2021)*, Brottrager et al. (2022), Bizzoni et al. (2023b)
Compressibility	Measures the extent to which the text can be compressed, serving as an indirect indicator of redundancy and lexical variety (Ehret and Szmrecsanyi, 2016).[c]	Stylistic	van Cranenburgh and Bod (2017), Koolen et al. (2020), Bizzoni et al. (2023c)
Word and bigram entropy	Measures the unpredictability in word choices and combinations, with higher entropy indicating greater variety and stylistic complexity.	Stylistic	Algee-Hewitt et al. (2016)
Normalized Dependency Distance, mean & SD (NDD Mean, NDD STD)	Quantifies the mean and SD in dependency length, following the procedure proposed in Lei and Jockers (2020).	Stylistic/ Syntactic	Lei and Jockers (2020)
Nominal verb ratio	Quantifies the proportion of nouns and adverbs (over verbs) in the text, reflecting the nominal tendency in style, which is often associated with complex linguistic structures, denser communicative code, expert-to-expert communication (McIntosh, 1975; Bostian, 1983).	Stylistic/ Syntactic	Charney and Rayman (1989)*, Crossley et al. (2014)*, Wu et al. (2024)
"Of"/"that" frequencies	Frequency of these function words have been seen to indicate, in the case of "of", a more nominal prose, and in the case of "that", a more declarative and verb-centered prose.	Stylistic/ Syntactic	Wu et al. (2024)
Function words	Frequency of function words (normalized for text length), suggesting a more information-rich prose when lower.	Stylistic/ Syntactic	Bizzoni et al. (2024a)
Perplexity	Represents the predictability of the prose through a self-trained large language models (GPT), as outlined in Wu et al. (2024).[d] Higher values indicate greater complexity or unpredictability.	Hybrid	Sheetz (2018), Wu et al. (2024)
Sentiment SD (SD Sent)	Represents the average variability in sentiment, indicating the range of sentiment within the narrative.[e]	Narrative/ Sentiment	Berger et al. (2021)*, Bizzoni et al. (2023c)
Hurst exponent	Quantifies the long-term auto-correlation of the sentiment arc,e with higher values suggesting a more complex, self-similar structure across different scales. *f*		Mohseni et al. (2021), Bizzoni et al. (2021), Bizzoni et al. (2023d)

Approximate en- tropy (ApEn)	Assesses the predictability of sequences of the sentiment arc, [e] with lower values indicating greater regularity or simplicity. *f*		Hu et al. (2020), Mohseni et al. (2022), Bizzoni et al. (2023c)

Table 2: Used features related to stylistic and sentiment complexity. "References" refer to studies that have used the complexity feature showing some relation between it and reader appreciation. * Denotes studies in domains other than established prose fiction (e.g., online stories, movies).

b Flesch Reading Ease and New Dale–Chall Readability Formula.
c We calculated the compression ratio (original bit-size/compressed bit-size) for the first 1500 sentences of each text using bzip2, a standard file-compressor.
d All perplexity calculations were via gpt2 models, done on the byte pair encoding tokenization used in the series of gpt2 models. To get the mean perplexity per novel, we used a sliding window due to maximum input length. For details on the computation, see Wu et al. (2024).
e All sentiment analysis was performed using the Syuzhet implementation on a sentence-basis (compound score). f For details on the measure, please refer to Bizzoni et al. (2023d).

Language as a complex system from a structural and diachronic perspective

Quentin Feltgen
Ghent University (Belgium)

Fifteen years ago, a position paper by the multi-disciplinary Five Graces group (2009) asserted that "Language is a complex adaptive system". Albeit their definition is far from settled (Ottino 2003, Ladyman et al. 2013, Estreda 2024), complex systems typically involve a large number of heterogenous interactions between their constituents, leading to emergent properties that are unpredictable from these interactions, and rather rely on their collective and repeated character.

The complex systems perspective on language has mostly focused on the sociolinguistic aspect of it, where language is not very different from a disease or an opinion that can spread, and the complexity is that of the social medium. As such, it has often been reduced to a competition between individual words (Chambers 1995, Ghanbarnejad et al. 2014, Amato et al. 2018, Pijpops 2022), or individual languages treated in a monolithic way (Abrams & Strogatz 2003, Scialla et al. 2023), constraining the linguistic phenomena that can be addressed from a quantitative complex systems perspective (Feltgen et al. 2017). Other works have treated language as a mapping between a set of meanings and a set of words (Pawlowitsch 2007, Baronchelli & Steels 2010), or between intervals of meanings (e.g. categories in the perceptual space) and words (Baronchelli et al. 2010), or intervals of phonemes (Victorri 2004). In these latter works, language appear as a resulting collective consensus over these mappings, achieved through repeated social interactions among agents.

On the other hand, the search for scaling laws, which is typical of a complex systems approach, have been applied to language (Altmann & Gerlach 2016), especially with regard to Zipf's law (Condon 1928), which has been seen as a reflect of a least effort principle in communication (Ferrer I Cancho & Solé 2003), or as a result of the nestedness property of the matrix of syntagmatic transitions (Thurner et al. 2015), a nestedness which derives from scaling properties (Payrató-Borras et al. 2019), that could themselves be the outcome of a preferential attachment growth of the network of syntagmatic relations (Barabási & Albert 1999). These works highlight the complexity of the language structure itself, even though, as Piantadosi (2014) highlighted, they only scratch its surface.

As a result, it appears that the complexity of the language system, that is, the complex ways through which language units are related on different levels within the linguistic organization (Sommerer & Smirnova 2019), has seldom been addressed. Yet, language is puzzling by its profuse character (Hoffmann 2004) and polysemous tendencies (Fonteyn 2021), two notions in seeming contradiction with each other and at odds with a simple form-to-meaning mapping. In this contribution, we shall illustrate these properties and how they contribute to language complexity using the French adverb carrément ('straight out') as a guideline example. We will address the systemic complexity at tree levels, through three kinds of interactions: the filler-host relationships for schematic units, the dynamical semiotic articulation between form and function in a diachronic perspective, and finally, the interactions between the forms themselves as they compete and align with each other.

First, we will consider the structural complexity associated with the adverb, by focusing on one specific context of use of the adverb, as an intensifier of an adjective (e.g. Les deux autres films que nous avons regardés sont moins notables, voire carrément oubliables. 'The other two movies that we watched are less remarkable, if not straight out forgettable.'), showing with data from the frTenTen23 corpus that the adjectives' collocational frequencies follow a construction-specific Zipfian distrubtion, as expected from past observations of this structure in the literature (Evert & Baroni 2005, Ellis & Ferreira-Junior 2009, Ellis 2012, Zeldes 2012, Ellis et al. 2014).

Second, we will consider the functional complexity of the adverb, by identifying several syntagmatic contexts of use: as an intensifier, with an adjective; isolated; as a manner adverb, immediately after a verb, (Álvarez-Prendes 2018). Tracking their frequency across time thanks to data from the Frantext corpus (ATILF 1998-2024), we show that these different functions clearly compete with each other over the form, highlighting its multi-functionality; but the emergence of a new function also triggers an increase of use of the other functions, hinting at the "unified" nature of this functional bundle.

Third, we will consider the paradigmatic complexity of the set of adverbs of which carrément is part, by showing that the frequency dynamics of carrément closely follows those of other adverbs of the same type (stance-marking adverbs in -ment; cf. Rhee 2016 for reflections on a similar category in English), revealing a highly structured and unexpected paradigmatic organization of this category. To do so, we resort to a clustering of the frequency share dynamics (i.e. how much the individual forms contribute to the total frequency of the paradigm of stance-marking adverbs), and show that carrément is part of a larger cluster, competing for the paradigm with the formerly established cluster.

In conclusion, we argue that focusing on the diachronic dynamics of specific linguistic forms, and of the syntagmatic contexts they associate with, is a fruitful entry point to unravel in full the highly complex, tightly intricate, and ever changing structure of language.

Bibliography:

Abrams, D. M., & Strogatz, S. H. (2003). Modelling the dynamics of language death. Nature, 424(6951), 900-900.

Altmann, E. G., & Gerlach, M. (2016). Statistical laws in linguistics. In Creativity and universality in language, Springer, 7-26.

Álvarez-Prendes, E. (2018). Polyfonctionnalité adverbiale, grammaticalisation et subjectivation: le cas de sérieusement, seriamente et en serio. Zeitschrift für romanische Philologie, 134(2), 471-486.

Amato, R., Lacasa, L., Díaz-Guilera, A., & Baronchelli, A. (2018). The dynamics of norm change in the cultural evolution of language. Proceedings of the National Academy of Sciences, 115(33), 8260-8265.

ATILF. (1998-2024). Base textuelle Frantext (online). ATILF-CNRS & Université de Lorraine. https://www.frantext.fr/

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. science, 286(5439), 509-512.

Baronchelli, A., & Steels, L. (2010). The minimal Naming Game: a complex systems approach. In Experiments in Language Evolution. John Benhamins Publ. Co.

Baronchelli, A., Gong, T., Puglisi, A., & Loreto, V. (2010). Modeling the emergence of universality in color naming patterns. Proceedings of the National Academy of Sciences, 107(6), 2403-2407.

Chambers, J. K. (1995). The Canada-US border as a vanishing isogloss: the evidence of chesterfield. Journal of English Linguistics, 23(1-2), 155-166.

Condon, E. U. (1928). Statistics of vocabulary. Science, 67(1733), 300-300.

Ellis, N. C., & Ferreira–Junior, F. (2009). Construction learning as a function of frequency, frequency distribution, and function. The Modern language journal, 93(3), 370-385.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. Annual review of applied linguistics, 32, 17-44.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014). Does language zipf right along?. Georgetown University Round Table on Languages and Linguistics, 33-50.

Evert, S., & Baroni, M. (2005). Testing the extrapolation quality of word frequency models. In P. Danielsson & M. Wagenmakers (Eds.), Proceedings of Corpus Linguistics 2005.

Estrada, E. (2024). What is a complex system, after all?. Foundations of Science, 29, 1143-1170.

Feltgen, Q., Fagard, B., & Nadal, J. P. (2017). Modeling language change: the pitfall of grammaticalization. In Language in Complexity: The Emerging Meaning, Springer, 49-72.

Ferrer I Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. Proceedings of the National Academy of Sciences, 100(3), 788-791.

“Five Graces Group”, Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., ... & Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. Language learning, 59, 1-26.

Fonteyn, L. (2021). Varying Abstractions: a conceptual vs. distributional view on prepositional polysemy. Glossa: a journal of general linguistics, 6(1).

Ghanbarnejad, F., Gerlach, M., Miotto, J. M., & Altmann, E. G. (2014). Extracting information from S-curves of language change. Journal of The Royal Society Interface, 11(101), 20141044.

Hoffmann, S. (2004). Are low-frequency complex prepositions grammaticalized? In Corpus approaches to grammaticalization in English, John Benjamins, 171-210.

Ladyman, J., Lambert, J., & Wiesner, K. (2013). What is a complex system?. European Journal for Philosophy of Science, 3, 33-67.

Ottino, J. M. (2003). Complex systems. American Institute of Chemical Engineers. AIChE Journal, 49(2), 292.

Pawlowitsch, C. (2007). Finite populations choose an optimal language. Journal of Theoretical Biology, 249(3), 606-616.

Payrató-Borras, C., Hernández, L., & Moreno, Y. (2019). Breaking the spell of nestedness: the entropic origin of nestedness in mutualistic systems. Physical Review X, 9(3), 031024.

Piantadosi, S. T. (2014). Zipf’s word frequency law in natural language: A critical review and future directions. Psychonomic bulletin & review, 21, 1112-1130.

Pijpops, D. (2022). Lectal contamination: Evidence from corpora and from agent-based simulation. International Journal of Corpus Linguistics, 27(3), 259-290.

Rhee, S. (2016). On the emergence of the stance-marking function of English adverbs: A case of intensifiers. Linguistic Research, 33(3).
Scialla, S., Liivand, J. K., Patriarca, M., & Heinsalu, E. (2023). A three-state language competition model including language learning and attrition. Frontiers in Complex Systems, 1, 1266733.
Sommerer, L., & Smirnova, E. (Eds.). (2020). Nodes and networks in diachronic construction grammar (Vol. 27). John Benjamins Publishing Company.
Thurner, S., Hanel, R., Liu, B., & Corominas-Murtra, B. (2015). Understanding Zipf's law of word frequencies through sample-space collapse in sentence formation. Journal of the Royal Society Interface, 12(108), 20150330.
Victorri, B. (2004). Continu et discret en sémantique lexicale. Cahiers de praxématique, (42), 75-94.
Zeldes, A. (2012). Productivity in argument selection: From morphology to syntax (Vol. 260). Walter de Gruyter.

Linguistic Communication as Information Compression and Extraction: a Unified Framework for Complexity in Human Communication

Evie A. Malaia
Department of Speech, Language, and Hearing
University of Alabama

Keywords: complexity, comprehension, processing cost, compression, transmission

Human languages are systems optimized for efficient communication, balancing the compression of information with cognitive processing costs. I propose a unified framework based on Marr's levels of analysis to quantify - measure and model - complexity in multimodal linguistic communication. I consider the relationship between linguistic structures (syntax, morphology) and cognitive resources equivalent to those between Marr's algorithmic level and computational levels. Languages can optimize the transmission of messages by increasing information content per unit of time (Andres et al., 2021; Bentz & i Cancho, 2016; Malaia & Wilbur, 2020). While compact, information-dense linguistic structures enhance transmission efficiency, they also increase cognitive load during processing, as observed in both spoken and signed languages.

This framework integrates insights from psycholinguistics, neurolinguistics, and computational modeling to explore complexity trade-offs. For example, complex syntax enables rapid information transfer but requires significant neural resources, reflecting a universal encoding-decoding tradeoff. Individual factors, such as familiarity with contexts or frequent vocabulary use, further influence perceived complexity. For example, complex syntactic structures allow for highly efficient encoding of information, minimizing redundancy and enabling rapid transmission of meaning; however, this efficiency comes at a cognitive cost (Blumenthal-Dramé et al., 2017). Linguistic structures that are compact and information-dense (such as reduced relative clauses, or use of non-subject noun to mark the topic at the beginning of a sentence) turn out to be metabolically expensive (i.e. more cognitively taxing) to process during decoding (Krebs et al., 2018; Malaia et al., 2009). Human brain must engage more neural resources to process complex syntax and morphology, or access less frequently used vocabulary (Blumenthal-Dramé & Malaia, 2019). This encoding-decoding tradeoff reflects the balance between socially and temporally distributed optimization for efficient communication during language evoluation, and individual cognitive load management required for comprehension of immediate communication.

By synthesizing approaches from formal linguistics and complex systems theory, this framework offers a robust tool for analyzing linguistic phenomena across modalities. Approaches such as entropy-based analyses (Borneman et al., 2018; Torre et al., 2019) offer promising avenues for revealing complexity tradeoffs between levels of linguistic anlaysis (e.g. syntax vs. morphology), or levels of communicaiton (production vs. perception) for overall optimization of communication. Exploring novel ways to relate variables across the scales of analysis (from phonology to pragmatics) may provide a more comprehensive understanding of language as a system. Integration of traditional linguistic analysis with complex systems methods used in neuroscience promises a robust framework for modeling language complexity that accounts for both individual cognitive processes and broader social dynamics of language use. The framework emphasizes the need for interdisciplinary collaboration and the development of cross-disciplinary methods to quantify and comprehensively model language complexity.

Bibliography:

Andres, J., Benešová, M., & Langer, J. (2021). Towards a fractal analysis of the sign language. Journal of Quantitative Linguistics, 28(1), 77–94.

Bentz, C., & i Cancho, R. F. (2016). Zipf's Law of Abbreviation as a Language Universal. Universitätsbibliothek Tübingen.

Blumenthal-Dramé, A., Glauche, V., Bormann, T., Weiller, C., Musso, M., & Kortmann, B. (2017). Frequency and chunking in derived words: A parametric fMRI study. Journal of Cognitive Neuroscience, 29(7), 1162–1177.

Blumenthal-Dramé, A., & Malaia, E. (2019). Shared neural and cognitive mechanisms in action and language: The multiscale information transfer framework. Wiley Interdisciplinary Reviews: Cognitive Science, 10(2), e1484.

Borneman, J. D., Malaia, E. A., & Wilbur, R. B. (2018). Motion characterization using optical flow and fractal complexity. Journal of Electronic Imaging, 27(05), 1. https://doi.org/10.1117/1.JEI.27.5.051229

Krebs, J., Malaia, E., Wilbur, R. B., & Roehm, D. (2018). Subject preference emerges as cross-modal strategy for linguistic processing. Brain Research. https://doi.org/10.1016/j.brainres.2018.03.029
Malaia, E. A., & Wilbur, R. B. (2020). Syllable as a unit of information transfer in linguistic communication: The entropy syllable parsing model. Wiley Interdisciplinary Reviews: Cognitive Science, 11(1), e1518.
Malaia, E., Wilbur, R. B., & Weber-Fox, C. (2009). ERP evidence for telicity effects on syntactic processing in garden-path sentences. Brain and Language, 108(3), 145–158. https://doi.org/10.1016/j.bandl.2008.09.003
Torre, I. G., Luque, B., Lacasa, L., Kello, C. T., & Hernández-Fernández, A. (2019). On the physical origin of linguistic laws and lognormality in speech. Royal Society Open Science, 6(8), 191023.

Context-dependency in measures of articulatory complexity

Charles Redmon,[1] Meghavarshini Krishnaswamy,[2] and Indranil Dutta[3]
[1]Department of Language and Linguistics, University of Essex
[2]Department of Linguistics, University of Arizona
[3]Department of Linguistics, Jadavpur University

Keywords: coarticulation, lexicon, Malayalam, asymmetries, coronals

In modeling articulatory complexity, particularly in the context of gestural coordination and coarticulation, much of the focus has been on the muscular and biomechanical characteristics of the articulators involved. This can be seen more formally, for instance, in the Degree of Articulatory Constraint model of Recasens et al. (1997), but is also apparent in a practical sense in the operations of articulatory synthesis models such as that of ArtiSynth (Lloyd et al., 2012) and TADA (Nam et al., 2004, 2012). For instance, it has long been known that consonants with greater articulatory and motor constraints, especially those involving the tongue-dorsum, tend to resist coarticulatory influence from neighbouring vowels relative to those with lesser or no tongue-dorsum involvement (Bladon and Al-Bamerni, 1976). This result has been most thoroughly studied via the locus equation (LE) model (cf. Sussman et al., 1998 and Lindblom & Sussman, 2012 for review), wherein consonants (largely plosives) with greater articulatory complexity and therefore greater resistance to coarticulation with neighbouring vowels exhibit flatter LE slopes; i.e., the second formant at consonant offset/onset varies less with variation in vocalic F2.

However, such notions of complexity inherent to a particular consonant or consonant-vowel combination are just one component of the problem, as they do not incorporate the wider system in which these sounds operate. For instance, the size and structure of the contrastive inventory is known to impact the range of acoustic variation languages tolerate in production (Manuel, 1990, 1999), and similarly as the ultimate function of the sound system is to encode higher-order units such as words and phrases there are also well-known correspondences between coarticulation and lexical frequency (Berry & Weismer, 2013). These latter findings also connect critically to usage-based frameworks (e.g., Bybee, 1999) as well as wider debates on the influence of lexical neighbourhood structure on speech production (Munson & Solomon, 2004; Wright, 2004; Baese-Berk & Goldrick, 2009). Therefore, any account of articulatory complexity must integrate this multiplicity of factors, meaning cross-linguistic comparison and generalisation on the basis of inventories alone is likely to be under-informative at best, and at worst can generate substantially misleading predictions.

We illustrated one such case where this was borne out in Malayalam (Dutta et al., 2019), where we showed that the coronal plosive system can be seen to reflect competing constraints of density in the inventory (the dental-alveolar-retroflex contrast dense in terms of place of articulation and unsurprisingly is cross-linguistically rare), context-dependency (such sounds only occur as voiceless geminates intervocalically), and heterogeneity in usage statistics (alveolars are generally sparse in the lexicon for historical reasons; retroflexes are much more common post-vocalically than pre-vocalically). In particular, contrary to general predictions that retroflexes should be most resistant to coarticulation, alveolars were found to coarticulate the least with following vowels (see Figure 1), while retroflexes coarticulated the most with preceding vowels but only via retroflexion of the vowel itself. More recent analysis of ultrasound data has corroborated this finding by showing alveolars to exhibit the least variation in tongue dorsum position in different vowel contexts.

In this presentation we will recontextualise these results to address the general question of the conference regarding how complexity is properly defined and integrated within linguistic theory, which in the case of articulation must involve a more comprehensive modeling of the dynamical process of speech encoding/decoding. In particular, articulatory complexity must be studied as a function of the distribution of coupled articulations across the language, consistent with the fundamental principles of Articulatory Phonology (Browman & Goldstein, 1986, 1989), but with an emphasis on scaling models to the wider encoding system in the lexicon. Malayalam presents an excellent test case wherein motor complexity intersects with inventory density, phonotactic asymmetries, and heterogeneity in lexical distributions. This work fits within a wider framework in development which seeks to integrate more directly phonetic modeling with the structure of higher-order systems that the articulatory apparatus serves to encode (Redmon & Jongman, 2024).

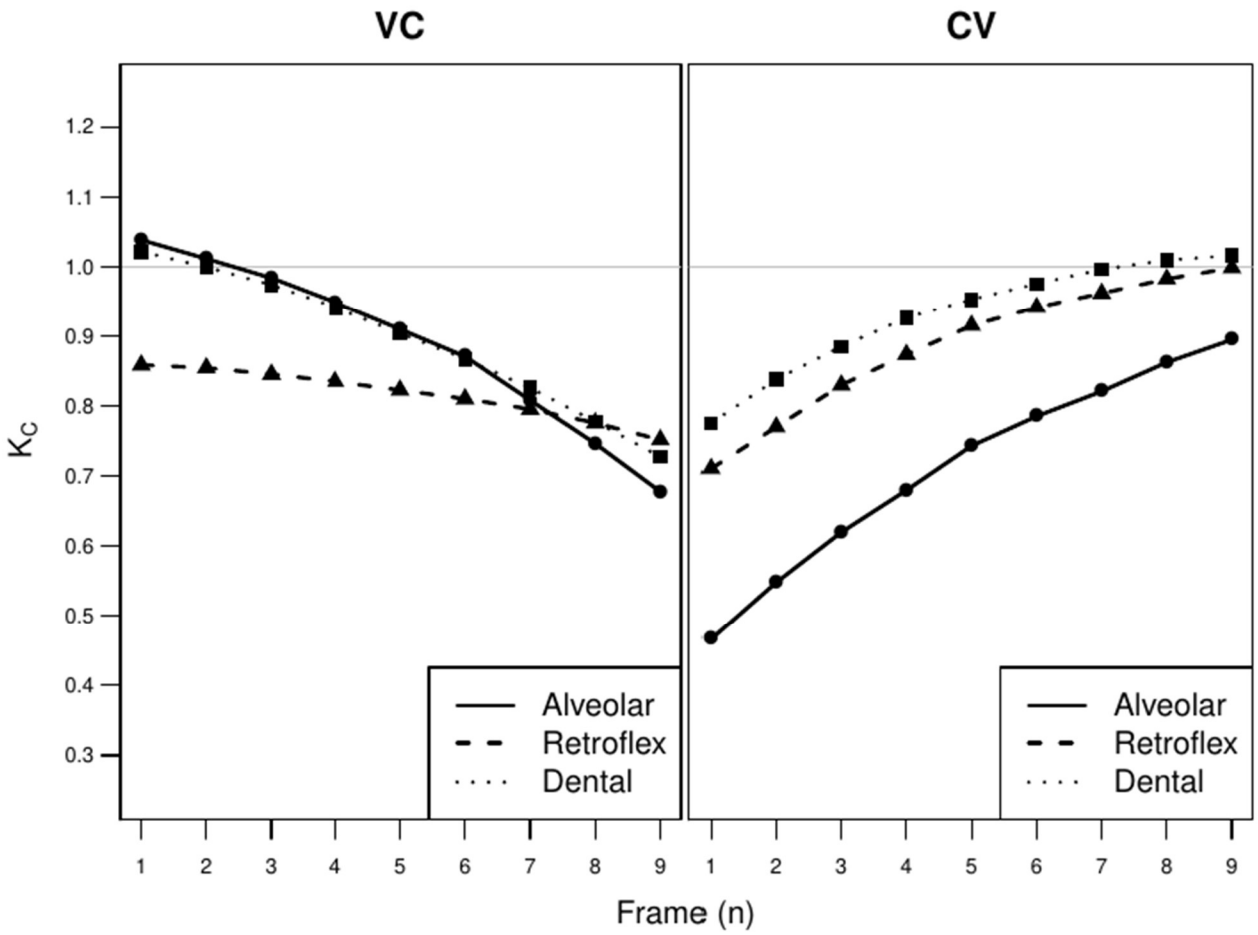


Figure 1. Scaling formant transitions from target-locus models by stop POA, for VC and CV transitions, aggregated across participants.

Bibliography:

Baese-Berk, M., & Goldrick, J. (2009). Mechanisms of interaction in speech production. Language and Cognitive Processes, 24, 527-554.

Berry, J., & Weismer, G. (2013). Speaking rate effects on locus equation slope. Journal of Phonetics, 41(6), 468-478.

Bladon, R. A., & Al-Bamerni, A. (1976). Coarticulation resistance in English /l/. Journal of Phonetics, 4, 135–150.

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. Phonology, 3, 219-252.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. Phonology, 6(2), 201-251.

Bybee, J. (1999). Usage-based phonology. Functionalism and Formalism in Linguistics, 1, 211-242.

Dutta, I., Redmon, C., Krishnaswamy, M., Chandran, S., & Raj, N. (2019). Articulatory complexity and lexical contrast density in models of coronal coarticulation in Malayalam. In Proceedings of the 19th International Congress of Phonetic Sciences.

Lindblom, B., & Sussman, H. M. (2012). Dissecting coarticulation: How locus equations happen. Journal of Phonetics, 40(1), 1-19.

Lloyd, J. E., Stavness, I., & Fels, S. (2012). ArtiSynth: A fast interactive biomechanical modeling toolkit combining multibody and finite element simulation. In Payan, Y. (Ed.), Soft Tissue Biomechanical Modeling for Computer Assisted Surgery, 355-394.

Munson, B., & Solomon, P. N. (2004). The effect of phonological neighborhood density on vowel articulation. Journal of Speech, Language and Hearing Research, 47, 1048-1058.

Nam, Hosung, Louis Goldstein, Elliot Saltzman & Dani Byrd. 2004. TADA: An enhanced, portable task dynamics model in MATLAB. The Journal of the Acoustical Society of America, 115, 2430.

Nam, H., Mitra, V., Tiede, M., Hasegawa-Johnson, M., Espy-Wilson, C., Saltzman, E., & Goldstein, L. (2012). A procedure for estimating gestural scores from speech acoustics. The Journal of the Acoustical Society of America, 132(6), 3980-3989.

Recasens, D., Pallarès, M. D., and Fontdevila, J. (1997). A model of lingual coarticulation based on articulatory constraints. The Journal of the Acoustical Society of America, 102(1), 544–561.

Redmon, C., & Jongman, A. (2024). From interfaces to system embedding: Phonetic contrasts in the lexicon. In Schlechtweg, M. (Ed.), Interfaces of Phonetics (pp. 23-69). Berlin: De Gruyter.
Sussman, H. M., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: The orderly output constraint. Behavioral and Brain Sciences, 21(2), 241-259.
Wright, R. (2004). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden & R. Temple (Eds.), Papers in Laboratory Phonology VI (pp. 26-50). Cambridge: Cambridge University Press.

Vers une mesure de la complexité des mots dérivés : que mesure-t-on et dans quel but ?

Gala, Núria, Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
Di Garbo, Francesca, Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
Colé, Pascale, Aix Marseille Univ, CNRS, CRPN, Marseille, France

Mots clés : Morphologie dérivationnelle, mesure de la complexité morphologique, procédés morphologiques, difficultés d'apprentissage/compréhension.

Le domaine de la morphologie a fait l'objet de nombreuses études à propos de la notion de complexité (Çöltekin & Rama, 2023). Cet intérêt pour la notion de 'complexité' s'est traduit par la proposition de métriques et indices permettant de quantifier la diversité de formes et la productivité de leurs combinaisons. En sciences du langage, les approches sont majoritairement typologiques, c'est-à-dire qu'elles s'intéressent à la comparaison de systèmes morphologiques entre langues (Baermann et al, 2017 ; Stump, 2017). Très souvent le focus porte sur les phénomènes de flexion (par exemple, l'index de complexité morphologique ou MCI, morphological complexity index, de Brezina & Palloti (2019), qui mesure la diversité de formes fléchies dans un texte). La notion de productivité, une mesure corrélée à la fréquence d'apparition des formes dans des corpus (Baayen, 1992), est souvent mobilisée, par exemple, lorsqu'on mesure la taille d'un paradigme, d'une famille morphologique, ou la variété/diversité lexicale dans un corpus. La notion de productivité des procédés morphologiques est, par ailleurs, liée à la prédictibilité sémantique : les mots créés avec les affixes les plus productifs seraient aussi plus prédictibles sémantiquement (ibid.).

À notre connaissance, la notion de complexité spécifique à la morphologie dérivationnelle demeure assez peu explorée en sciences du langage, en dépit du rôle fondamental de la morphologie dérivationnelle dans l'apprentissage de la lecture, l'acquisition de vocabulaire nouveau et la compréhension orale et écrite (Seiderberg, 1989 ; Goodwin et al., 2017, Colé et al. 2018). Toutefois, une mesure de la complexité dérivationnelle (MCD) pourrait s'avérer très utile dans l'enseignement primaire et secondaire. Elle permettrait de comparer différentes formes dérivées de façon systématique et par rapport à un certain nombre de traits structurels qui sont souvent utilisés en linguistique générale pour étudier la relation entre forme et sens, telle que la transparence ou la régularité. Une MCD devrait alors prendre en compte : (1) les caractéristiques liées aux formes linguistiques isolées et à leurs familles morphologiques d'appartenance (respectivement, longueur, transparence/opacité sémantique, polysémie ; taille, productivité), mais aussi (2) les propriétés inhérentes aux procédés morphologiques mis en œuvre pour la création de ces formes (régularité/quasi-régularité, productivité). Certaines mesures, comme la taille de la famille morphologique et le nombre de morphèmes, ont déjà montré leur corrélation positive avec la notion de difficulté lexicale (Gala et al. 2014). D'autres, comme la régularité ou quasi-régularité d'un procédé morphologique, n'ont pas été, à notre connaissance, réellement exploitées et elles pourraient amener à une meilleure compréhension des facteurs qui contribuent à certaines difficultés d'apprentissage et mémorisation chez les enfants en âge scolaire.

Une mesure de la complexité dérivationnelle pourrait ainsi être utilisée dans l'enseignement des langues pour établir une graduation entre les formes proposées pour l'apprentissage de différents procédés de formation des mots et/ou de différentes familles de mots. Par exemple, un enseignant pourrait se demander s'il y a une différence en 'complexité' entre les suffixes -aire et -ique en français, tous les deux permettant de construire des adjectifs à partir de noms, comme 'lunaire' ou 'poétique'. Est-ce que les dérivations avec ces suffixes sont régulières ? transparentes ? ; on pourrait aussi vouloir comparer les familles morphologiques de deux bases, non seulement en termes de productivité mais également selon leur régularité, opacité, cohésion, etc. Si on prend un exemple concret, quel mot dérivé entre 'antiquité' et 'antirouille' est le plus difficile à lire/comprendre pour un apprenant, sachant qu'ils sont tous les deux des noms de longueur proche, respectivement 9 et 11 caractères, et composés de deux morphèmes fréquents en français ? Est-ce que la chaîne 'anti' dans 'antiquité' amène à confusion par rapport au préfixe anti- présent dans 'antirouille' ? Est-ce que la préfixation en anti- est plus fréquente/productive que la suffixation en -ité ? Est-ce que le fait que 'antiquité' soit un mot plus fréquent que 'antirouille' le rend plus 'facile' à déchiffrer et à comprendre ? D'autres questions qu'une mesure de la complexité pourrait amener à envisager, d'un point de vue linguistique, seraient les suivantes : est-ce que la complexité d'une forme dérivée est déterminée par le procédé morphologique (ex. préfixation vs suffixation vs supplétisme vs composition savante) ? Est-ce

qu'elle est liée au type d'affixe (un affixe fréquent rendrait le mot dérivé plus 'facile') ? Quelle est l'impact/le poids de la phonologie, l'orthographe, la sémantique dans le calcul d'une MCD ?

En tenant compte de ce type d'interrogations et dans le but de concevoir une mesure qui soit utile à des fins pédagogiques, dans cette communication nous présenterons, dans un premier temps, un état de la question qui nous mène à conclure qu'à ce jour il n'existe pas, au-delà de la notion intuitive de longueur (nombre de morphèmes), une mesure spécifique permettant de comparer la complexité de deux mots polymorphématiques au sein d'une même langue comme le français. Dans un deuxième temps, nous proposerons un ensemble de variables participant à la caractérisation de la complexité d'une forme dérivée. Ces variables se justifient dans le cadre de la « théorie de la convergence » qui conçoit le codage en mémoire de la morphologie comme une représentation graduelle et inter-niveaux capturant les corrélations entre l'orthographe, la phonologie et la sémantique (Seidenberg & Gonnerman, 2000).

Une telle mesure pourrait se révéler indispensable à la création de matériels pédagogiques : elle permettrait d'introduire une graduation objective dans la proposition d'exemples à apprendre, elle permettrait d'observer la progression d'exemples de mots dérivés dans des corpus de différents niveaux (scolaires pour le français langue maternelle, ou en FLE). Une MCD serait également utile dans l'étude des langues comme le français (ou les langues romanes en général) où la morphologie dérivationnelle est le procédé le plus important pour la création de nouveaux mots (Duncan et al, 2009 ; Haspelmath & Sims, 2010).

Bibliographie :

Baayen, H. (1992). Quantitative aspects of morphological productivity. In: Booij, G., van Marle, J. (eds) Yearbook of Morphology. Springer, Dordrecht. https://doi.org/10.1007/978-94-011-2516-1_8

Baerman, M., Brown, D. & Corbett, G (2017). Morphological Complexity. Cambridge: Cambridge University Press.

Brezina, V., & Pallotti, G. (2019). Morphological complexity in written L2 texts. Second language research, 35(1), 99-119.

Colé, P., Cavalli, E., Duncan, L. G., Theurel, A., Gentaz, E., Sprenger-Charolles, L., & El-Ahmadi, A. (2018). What is the influence of morphological knowledge in the early stages of reading acquisition among low SES children? A graphical modeling approach. Frontiers in Psychology, 9, 547.

Çöltekin, Ç., & Rama, T. (2023). What do complexity measures measure? Correlating and validating corpus-based measures of morphological complexity. Linguistics Vanguard, 9(s1), 27-43.

Duncan, L., Casalis, S. & Colé, P. (2009). Early metalinguistic awareness of derivational morphology: Observations from a comparison of English and French. Applied Psycholinguistics. 30. 405 - 440. 10.1017/S0142716409090213.

Gala, N., François, T., Bernhard, D., & Fairon, C. (2014). Un modèle pour prédire la complexité lexicale et graduer les mots. Actes de la Conférence Traitement Automatique des Langues (TALN'2014), Marseille, 91-102.

Goodwin, A. P., Petscher, Y., Carlisle, J. F., & Mitchell, A. M. (2017). Exploring the dimensionality of morphological knowledge for adolescent readers. Journal of research in reading, 40(1), 91-117.

Haspelmath, M. & Sims, A. D. (2010). Understanding Morphology (2nd edition). London : Hachette.

Seidenberg, M. S. & Gonnerman, L. M. (2000) Explaining derivational morphology as the convergence of codes. Trends in Cognitive Sciences, Volume 4, ISSUE 9, 353-361.

Seidenberg, M.S. (1989). Reading Complex Words. In: Carlson, G.N., Tanenhaus, M.K. (eds) Linguistic Structure in Language Processing. Studies in Theoretical Psycholinguistics, vol 7. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-2729-2_3

Stump, G. (2017). The nature and dimensions of complexity in morphology. Annual Review of Linguistics, 3(1), 65-83.

Understanding morphosyntactic complexity through a functionalist, psycholinguistic perspective: The resilience of noun-phrase agreement structures in standard German

Daniel Walter
Associate Professor of German and Linguistics
Emory University, Oxford College

Keywords: Functionalism, psycholinguistics, German, morphosyntax

For many Indo-European languages , the existence of case, grammatical gender (hereafter gender), and number information make noun-agreement structures quite complex. Not only does gender and number information from the noun have to be transferred to all agreeing elements within the noun-phrase, such as determiners and adjectives, but noun-phrase external case information must also be applied in tandem, which is assigned as a syntactic role from a verb or passed down from a preposition. The result is that noun-internal and noun-external information must be simultaneously integrated to compute corresponding surface forms. This complexity has ramifications for both language processing and production, especially as the noun often appears after agreement structures have already been formed (Pearlmutter, Garnsey, & Bock, 1999).

German is no exception in this language family, and while it does not have the most grammatical cases, it has arguably one of the more difficult noun-phrase agreement systems due to the homophony/-graphy of surface forms (Kempe & MacWhinney, 1998). This is because German noun-phrases can be formed in 16 unique case/gender/number combinations (four cases applied to three genders in the singular and one plural form), but the result is not 16 distinct surface forms, rather only 6 in the case of definite articles (der, die, das, den, dem, des). The resulting system is therefore not only complex due to the number of feature combinations, but also processing requirements in real-time that may necessitate the inclusion of other contextual information for disambiguation.

Many languages do not require this complex nominal morphosyntactic system for speakers to interpret syntactic roles, so it is legitimate to ask, why does this complex system continue to exist in German, given that it has been diluted in other languages Indo-European languages like Dutch (Audring, 2006) and French (Ashdowne & Smith, 2005; Laubscher, 1921), or died out (almost) completely in others like English (Markus, 1988)?

If we take two perspectives into account, the function of grammar and the psycholinguistics of language learning and use, which is outlined in the unified competition model (MacWhinney 2008; 2013; 2022), we can understand this sustained complexity.

First, from a functional perspective (e.g., Bybee, 1998; Smith & Nordquist, 2018), grammar exists to create meanings that are relevant and useful for its speakers, and it needs to do so in a way that allows the grammatical forms to be sustained, or otherwise be subsumed by other grammatical features and in turn fade from the language. In German, the complex case-gender-number marking system plays an important functional role of enabling topicalization or thematicization that cannot be done in other languages, like English, which employ (almost solely) syntactic information to assign syntactic roles, like subjects and objects. German speakers are enabled by the case-marking system to vary word order, and the fact that German speakers must pay attention to nominal morphology is a feedback loop from its function in assigning syntactic role. If this word-order variation were random or served no purpose, it would likely die out, but in German, sentence-initial positioning of non-subjects creates topicalization that forefronts new or requested information (Cardinaletti, 1986). For example, if someone asks, in German, where an event is taking place, German speakers freely provide the location information in the sentence-initial position and restructure the sentence to have the subject follow the verb. Corpora studies like the one by Bader and Häussler (2010) also show variation in SVO versus OVS word-order preferences by object-type and verb. Thus, German speakers are continually forced to process case-morphology to make correct interpretations of utterances. And because the case-morphology is only interpretable with knowledge of the gender assignment of nouns, genders (which make sense here to think of as noun-classes that form from common colocations, especially in L1 language learning) maintain their importance in the language.

We can see in certain instances, where case and gender information do not serve much of a functional purpose, like in prepositional constructions, that case marking loses some relevance. For example, the German genitive case, which is used to form the possessive, has been dying out in German for, in my analysis, two reasons. First, after genitive prepositions, many Germans will use the dative form (wegen dem Wetter instead of wegen des Wetters) (Braunmüller, 2018). A misinterpretation of which case to use after genitive prepositions does not, in any meaningful way, alter the meaning of the

construction and therefore serves little to no purpose. In addition, the surface forms of the genitive and dative feminine are identical, which I would describe as a chink in the armor protecting the grammar from change. In other words, there is a place where two surface forms, which are not functionally very distinguishable from one another, overlap, leading to an inroad for misinterpretation by speakers that the dative form is the appropriate case and thus an application of dative endings to masculine and neuter nouns in the same position. The other issue is that possessive constructions can also be formed using the preposition von (of), and therefore is not the sole grammatical option to produce an intended meaning (Campe, 1999). These two reasons, the confusion between two surface forms and multiple available grammatical options for a single meaning, have led to a decline in the use of genitive case in colloquial German. And if one looks further at German spoken varieties around the world, we see significant differences in the application of the standard noun-phrase morphology in varying dialects (Baechler & Pröll, 2018).

On the other hand, the complexity of case-gender-number morphology in German should also be analyzed from a psycholinguistic perspective, incorporating language learning, processing, production, and attrition. We can view grammar as a give and take, where certain grammatical features allow for, or force particular ways of processing information (Frost, Gringer, & Rastle, 2005). In the case of German, distinction of nouns into various genders may seem complicated, especially from an L2 learner perspective (Walter & MacWhinney, 2015). However, for L1 learners who rely heavily on statistical learning, the positioning of the surface forms and their colocation with certain nouns and not others allow for distinct classes of nouns to form, just as other word categories can be formed through simple collocational processing (Li, Farkas, & MacWhinney, 2004). The apparent added complexity of having multiple noun classes simplifies noun processing because cues to gender in the pre-nominal morphology (i.e., on articles and adjectives) actually speeds up processing, because it delimits the number of possible nouns that can come after them (Taraban & Kempe, 1999). Combined with contextual knowledge, this grammatical knowledge aids in both speed and accuracy of lexical processing and prediction.

In sum, adopting a model for understanding complexity in languages that 1) accounts for the function of the complexity and 2) investigates the potential upsides for learning and processing the complexity allow us to better understand how complexity in the world's languages is maintained across time over generations of speakers. It also allows us to find a foothold to investigate why and how complexity is lost (or gained) over time and introduce other forces like language contact, literacy rates and practices, and second language learners on language change over time.

Bibliography:

Ashdowne, R. & Smith, J. C. (2005). Some semantic and pragmatic aspects of case-loss in Old French. In J. C. Salmons & S. Dubenion-Smith (Eds.) Historical Linguistics 2005 (pp. 191-206). John Benjamins.

Audring, J. (2006). Pronominal gender in spoken Dutch. Journal of Germanic Linguistics, 18(2), 85-116.

Bader, M., & Häussler, J. (2010). Word order in German: A corpus study. Lingua, 120(3), 717-762.

Baechler, R., & Pröll, S. (2018). Loss and preservation of case in Germanic non-standard varieties. Glossa: A Journal of General Linguistics 3(1): 1-35.

Baldi, P. (2018). Indo-European languages. In B. Comrie (Ed.) The world's major languages (pp. 23-50). Routledge.

Braunmüller, K. (2018). On the role of cases and possession in Germanic. In T. Ackermann, H. J. Simon, & C. Zimmer (Eds.) Germanic genitives (pp. 301-323). John Benjamins.

Bybee, J. L. (1998). A functionalist approach to grammar and its evolution. Evolution of communication, 2(2), 249-278.

Campe, P. (1999). Genitives and von-datives in German: A case of free variation. In M. H. Verspoor, K. D. Lee, & E. Sweetser (Eds.) Lexical and syntactical constructions and the construction of meaning (pp. 165-186). John Benjamins.

Cardinaletti, A. (1986). Topicalization in German: Movement to comp or base-generation in top? GAGL: Groninger Arbeiten zur Germanistischen Linguistik 28, 202-231.

Frost, R., Grainger, J., & Rastle, K. (2005). Current issues in morphological processing: An introduction. Language and cognitive processes, 20(1-2), 1-5.

Kempe, V., & MacWhinney, B. (1998). The acquisition of case marking by adult learners of Russian and German. Studies in second language acquisition, 20(4), 543-587.

Laubscher, G. G. (1921). The syntactical causes of case reduction in Old French (No. 7). Princeton University Press.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. Neural networks, 17(8-9), 1345-1362.

MacWhinney, B. (2008). A unified model. In P. Robinson & N. C. Ellis (Eds.) Handbook of cognitive linguistics and second language acquisition (pp. 351-381). Routledge.

MacWhinney, B. (2013). The logic of the unified model. In S. Gass & A. Mackey (Eds.) The Routledge handbook of second language acquisition (pp. 211-227). Routledge.
MacWhinney, B. (2022). The competition model: Past and future. In J. Gervain, G. Csibra, & K. Kovacs (Eds.) A life in cognition: studies in cognitive science in Honor of Csaba Pléh (pp. 3-16). Springer Cham.
Markus, M. (1988). Reasons for the loss of gender in English. In D. Kastovsky & G. Bauer (Eds.) Luick Revisited: Papers read at the Luick-Symposium at Schloß Liechtenstein 15.-18. 9. 1985 (pp. 241-58). Gunter Narr Verlag.
Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. Journal of Memory and language, 41(3), 427-456.
Smith, K. A., & Nordquist, D. (Eds.). (2018). Functionalist and usage-based approaches to the study of language: In honor of Joan L. Bybee (Vol. 192). John Benjamins Publishing Company.
Taraban, R., & Kempe, V. (1999). Gender processing in native and nonnative Russian speakers. Applied Psycholinguistics, 20(1), 119-148.
Walter, D., & MacWhinney, B. (2015). US German majors' knowledge of grammatical gender. Die Unterrichtspraxis/Teaching German, 48(1), 25-40.

Clausal complexity across registers in German and Persian

Nico Lehmann
Humboldt-Universität zu Berlin

Keywords: register, variation, complexity, clausal embedding, communicative situations

Introduction: Several studies have shown that the degree of clausal complexity, here "increased hierarchical organization" (Givo´n 2009: 2) in terms of embedding (see Verhoeven & Lehmann 2018), varies between situational contexts. Karlsson (2007), for instance, found that the phonic mode is more constrained in terms of clausal center embedding than the graphic mode. The contrast in syntactic complexity between situational contexts distinguished by the phonic and graphic mode has generally received wide attention (see e.g. Halliday 1979; Beaman 1984; Sakel & Stapert 2010; Maas 2010; Kornai 2014). Yet complexity also differs between varying phonic situational contexts, see e.g. Verhoeven & Lehmann (2018: 20) who show that public spoken communication displays a higher depth of embedding in contrast to non-public spoken interactions.
Assuming that syntactic choices are functionally motivated (Giv´on 1979: 81f.), higher or lower clausal complexity serve a communicative function that is related to certain situational-functional parameters, which means clausal complexity differs in particular socially recurring varieties, i.e. registers (Biber 2012; Biber & Conrad 2019; Lu¨deling et al. 2022). Since embedded structures are acquired later than coordinative ones (see Ochs 1979: 68; Weiss & Meurers 2019: 386), language users increase clausal complexity, thereby displaying intra-individual variation, in order to navigate the communicative and socio-cultural space. An increase in embedding complexity in graphic communication, for instance, might compensate for the lack of face-to-face monitoring of comprehension while its explicitness may be favoured between unacquainted interlocutors to manage the common ground (Giv´on 1979: 105).
In many studies, however, the situational contexts compared differed in more than only one situational-functional parameter; Biber & Gray (2010) and Biber et al. (2011; 2022b), for example, compared academic writing and everyday spoken conversations, which next to mode vary in informational vs. non-informational communicative purposes. The prerequisite for understanding register variation in complexity and determining what motivates language users to vary in syntactic complexity is looking at individual parameters and their contribution to the variation.
In a comparative corpus study, I examine intra-individual variation in clausal complexity in situational contexts that only differ in a single parameter, looking at Persian and German which both have similar community sizes, a high degree of literacy as well as rich written traditions, making them pertinent for variation in clausal complexity across a variety of situational contexts. Moreover, Persian is said to have an even sharper distinction between formal and informal registers, sometimes characterized as a diglossic situation (Ferguson 1959; Modaressi-Tehrani 1978).
Method: The Lang*Reg corpus (Adli et al. 2023) used for the study contains intraindividual variation for six communicative situations: the same participant telling a travelling story to a friend in two modes as well as talking freely about travelling with various kinds of interlocutors (friend, stranger, taxi driver or professor), see the parameter setup in Table 1. This design allows for comparing intra-individual variation for the parameters mode, social hierarchy and social distance. All other situational parameters were controlled for, e.g. social factors such as age, communicative topic and purpose. There were 20 participants in Persian (21,648 clauses) and 12 participants in German (10,472 clauses). Clausal complexity was measured using the average number of embedded clauses and average depth of embedding. In addition, the number of embedding of individual clause types, i.e. complement clause, adverbial clause and nominal dependent clause, was examined since they vary in (syntactic) functions (Biber et al. 2022a).

	Interlocutor	Interactivity	Mode	Length	Space	Distance	Hierarchy
1	friend	−interactive	graphic	–	−accessible	−distant	P = I
2	friend	−interactive	phonic	2min	−accessible	−distant	P = I
3	friend	+interactive	phonic	15min	−accessible	−distant	P = I
4	stranger	+interactive	phonic	15min	−accessible	+distant	P = I
5	taxi driver	+interactive	phonic	15min	+accessible	+distant	P > I
6	professor	+interactive	phonic	15min	+accessible	+distant	P < I

Table 1: Differentiating parameters of the Lang*Reg corpus design. Next to mode, two main social relation criteria are varied: social distance between interlocutors (related to level of acquaintance) and social hierarchy between interlocutors, i.e. whether the participant (P) or their interlocutor (I) exerts more power in the scope of the activity.

Results: Overall, Persian participants employed a higher level of clausal complexity than German participants (Persian: 29.6% embedded; German: 19.9% embedded). A generalized linear mixed-effects model (Poisson distribution) shows that social hierarchy and mode significantly impact the number of embedded clauses per non-embedded clause as well as the depth of embedding in the respective communicative situations in both languages whereas social distance has no effect in either language, as seen in Table 2. While both languages display the same effect direction for mode, i.e. greater clausal complexity in the graphic mode, the effects are reversed for social hierarchy, i.e. German shows higher clausal complexity with the higher ranking addressee (the professor) whereas Persian displays fewer clausal complexity with the higher ranking in contrast to the lower ranking addressee (the taxi driver). The fact that social distance did not have an effect for either language means that – all else being equal – a difference in common ground does not appear to cause participants to change their linguistic behaviour with respect to clausal complexity when talking about travelling experiences in these socio-cultural contexts.

		Social Hierarchy	Mode	Social Distance
German	number of embedding	p<.001	p<.05	p>.05
	depth of embedding	p<.001	p<.1	p>.05
Persian	number of embedding	p<.001	p<.01	p>.05
	depth of embedding	p<.001	p<.05	p>.05

Table 2: Generalized linear mixed-effects models on the number and depth of clausal embedding (Poisson distribution; random intercept: participants) for German and Persian.

An analysis by type of clause reveals that it is mostly adverbial clauses that are used significantly more often in the graphic mode in both languages compared to the phonic mode, thereby relying on more explicit characterizations about the events described in the written storytelling context. The effect of social hierarchy, however, is carried by the lower rate of complement clauses in Persian with the higher ranking addressee whereas in German it is the rates of adverbial clauses and nominal dependent clauses that are significantly higher with the higher ranking addressee. The German results are in line with the expectations that adverbial clauses serve to be particularly precise with higher ranking interlocutors, yet the Persian results are puzzling. The decrease in complement clauses might point towards a reluctance to verba sentiendi / verba diciendi, which trigger complement clauses, in Persian with higher ranking addressees not found in German.

Bibliography:

Adli, Aria & Verhoeven, Elisabeth & Lehmann, Nico & Mortezapour, Vahid & Vander Klok, Jozina (eds.). 2023. Lang*Reg: A multi-lingual corpus of intra-speaker variation across situations.
Version 0.1.0. [Data set]. Zenodo. Berlin, K¨oln: Humboldt-Universit¨at zu Berlin, Universit¨at zu K¨oln. https://doi.org/10.5281/zenodo.7646320

Beaman, Karen. 1984. Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In Tannen, Deborah (ed.), Coherence in spoken and written
discourse: Advances in discourse processes (Coherence in Spoken and Written Discourse), 45–80. Norwood, NJ:: ABLEX Pub. Corp.

Biber, Douglas. 2012. Register as a predictor of linguistic variation. Corpus Linguistics and Linguistic Theory 8(1). 9–37. https://doi.org/10.1515/cllt-2012-0002

Biber, Douglas & Conrad, Susan. 2019. Register, Genre, and Style. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108686136

Biber, Douglas & Gray, Bethany. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. Journal of English for Academic Purposes 9(1). 2–20. https://doi. org/10.1016/j.jeap.2010.01.001

Biber, Douglas & Gray, Bethany & Poonpon, Kornwipa. 2011. Should We Use Characteristics of Conversation to Measure Grammatical Complexity in L2 Writing Development? TESOL Quarterly 45(1). 5–35. https://doi.org/10.5054/tq.2011.244483

Biber, Douglas & Gray, Bethany & Staples, Shelley & Egbert, Jesse. 2022a. Introduction. In The Register-Functional Approach to Grammatical Complexity: Theoretical Foundation, Descriptive Research Findings, Application, 1–5. New York, London: Routledge.

Biber, Douglas & Gray, Bethany & Staples, Shelley & Egbert, Jesse. 2022b. Theoretical and Descriptive Linguistic Foundations of the Register-Functional Approach to Grammatical Complexity. In The Register-Functional Approach to Grammatical Complexity: Theoretical Foundation, Descriptive Research Findings, Application, 6–22. New York, London: Routledge.

Ferguson, Charles A. 1959. Diglossia. Word 15. 325–340.
Giv´on, Talmy. 2009. Introduction. In Giv´on, Talmy & Shibatani, Masayoshi (eds.), Syntactic complexity: Diachrony, acquisition, neuro-cognition, evolution, 1–19. Amsterdam, Philadelphia: John Benjamins. https://www.jbe-platform.com/content/books/9789027290144.
Giv´on, Talmy. 1979. From Discourse to Syntax: Grammar as a Processing Strategy. In Giv´on, Talmy (ed.), Discourse and Syntax (Syntax and Semantics), 81–112. New York: Academic Press. https://doi.org/10.1163/9789004368897 005
Halliday, Michael A. K. 1979. Differences between spoken and written language. In Page, Glenda
& Elkins, John & O'Connor, Barrie (eds.), Communication through reading: Proceedings of Fourth Australian Reading Conference. 37–52. Adelaide: Australian Reading Association.
Karlsson, Fred. 2007. Constraints on multiple center-embedding of clauses. Journal of Linguistics 3(43). 365–392. https://doi.org/10.1017/S0022226707004616
Kornai, Andr´as. 2014. Resolving the infinitude controversy. Journal of Logic, Language and Information 23(4). 481–492. https://doi.org/10.1007/s10849-014-9203-2
Lu¨deling, Anke & Alexiadou, Artemis & Adli, Aria & Donhauser, Karin & Dreyer, Malte & Egg,
Markus & Feulner, Anna Helene & Gagarina, Natalia & Hock, Wolfgang & Jannedy, Stefanie & Kammerzell, Frank & Knoeferle, Pia & Krause, Thomas & Krifka, Manfred & Kutscher, Silvia
& Lu¨tke, Beate & McFadden, Thomas & Meyer, Roland & Mooshammer, Christine & Mu¨ller, Stefan & Maquate, Katja & Norde, Muriel & Sauerland, Uli & Solt, Stephanie & Szucsich, Luka & Verhoeven, Elisabeth & Waltereit, Richard & Wolfsgruber, Anne & Zeige, Lars Erik. 2022. Register: Language Users' Knowledge of Situational-Functional Variation: Frame text of the First Phase Proposal for the CRC 1412. Register Aspects of Language in Situation 1(1).
1–59. https://doi.org/https://doi.org/10.18452/24901
Maas, Utz. 2010. Literat und orat. Grundbegriffe der Analyse geschriebener und gesprochener Sprache. Grazer Linguistische Studien (73). 21–150.
Modaressi-Tehrani, Yahya. 1978. A Sociolinguistic Analysis of Modern Persian. Lawrence, KS: dissertation.
Ochs, Elinor. 1979. Planned and Unplanned Discourse. In Giv´on, Talmy (ed.), Discourse and
Syntax (Syntax and Semantics), 51–80. New York: Academic Press. https://doi.org/10.1163/ 9789004368897 004
Sakel, Jeanette & Stapert, Eugenie. 2010. Pirah˜a – in need of recursive syntax? In Hulst, Harry van der & Sakel, Jeanette & Stapert, Eugenie (eds.), Recursion and Human Language, 1–16. Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110219258.1
Verhoeven, Elisabeth & Lehmann, Nico. 2018. Self-embedding and complexity in oral registers. Glossa: a journal of general linguistics 3(1). 1–30. https://doi.org/10.5334/gjgl.592
Weiss, Zarah & Meurers, Detmar. 2019. Analyzing Linguistic Complexity and Accuracy in Academic Language Development of German across Elementary and Secondary School 380–393.
https://doi.org/10.18653/v1/w19-4440

Text complexity as a regression task

Trung Hieu Ngo[1], Nicolas Béchet[2], Delphine Battistelli[3]
[1]LS2N, CNRS-Nantes University
[2]IRISA, CNRS-South Brittany University
[3]MoDyCo, CNRS-Paris Nanterre University

Keywords: complexity, features, regression, natural language processing

Our communication aims to present the experiment we conducted for observing which are the most discriminating linguistic features in order to produce an efficient regression model for text complexity in French.

Text complexity is currently not an area with extensive studies in natural language processing (NLP). Text complexity can either be studied as a classification task or as a regression task. In the classification approach, text complexity has been studied as a text readability classification task (Mesgar and Strube, 2018; Balyan et al., 2020; Yancey et al., 2021), or as an age prediction task (Bayot and Gonçalves, 2017; Chen et al., 2019). In the regression approach, text complexity has been studied as a task of age prediction (Nguyen et al., 2011; Bayot and Gonçalves, 2017; Chen et al., 2019; Wilkens et al., 2022). The regression approach shows promise over classification approach due to the more fine-grained nature of regression over classification and this is the one we choose.

Our work takes place in the ANR project TextToKids [1] which is a multidisciplinary project, combining experts from linguistics, psycholinguistics and NLP. This project tackles the problem of access informational content of genre diversified texts (journalistic, fictional, encyclopedic) for children from 7 y. old to 12 y. old. Thus, it is directly concerned with the question of how to evaluate complexity of texts. For solving this question, the project focuses on how to describe complexity into elementary linguistic features. The project led to the development of an automatic extraction chain [2] of 347 linguistic features divided into six levels of linguistic analysis – described in a first version in (Blandin et al. 2020) and with a revised version in (Battistelli et al., 2022). Among these features, 64.3% translate a complexity marker identified in the psycholinguistic literature, 46.2% in the literature on readability or simplification of texts and 32% an exploratory marker, i.e. a marker intuitively considered as relevant for the analysis of complexity but not yet identified either in the psycholinguistic, readability or simplification literature (e.g. temporal devices, discursive connectors, emotional lexical units). The project has also made a major achievement in creating an automatic tool for predicting recommended minimal age ranges for texts' readers (Rahman et al., 2020, 2023) starting from a corpus annotated in age ranges as proposed by official publishers. We decided to conduct an experiment consisting of applying these two tools to a parallel corpus made up of pairs of texts in their original version and in their hand-simplified version; and thus to prove that the combination of our two automatic tools is able to: (i) to detect in a pair of texts which one is the simplified one ; (ii) identify which features are the most impactful ones for representing the difference in complexity between the original texts and their simplified versions. This parallel corpus made up of pairs of texts in their original version and in their hand-simplified version is named Alector corpus (Gala et al., 2020). Alector corpus is drawn from 79 French literary and scientific texts commonly used in schools for children from 7 to 9 y. of age. The corpus is organized into age grade levels of CE1 level (7 y. old), CE2 level (8 y. old), and CM1 level (9 y. old), along with samples from International Reading Tests to enrich the corpus. The simplifications were manually done at the lexical, morpho-syntactic, and discourse level.

To carry out our experiment, the linguistic features have been first extracted from all pairs of texts of Alector corpus using the extraction tool described in (Battistelli et al., 2022)[2]. As we said, this tool allows the extraction of features from 6 levels (namely Phonetic, Morphology, Morphosyntax, Lexical, Syntax, and Semantic). Using interpretability tools such as SHAP (Lundberg and Lee, 2017), we can find which features are more important to apprehend text complexity. The extracted features have been then used as input representation for our trained regression models such as XGBoost (Chen and Guestrin, 2016) to predict the recommended age for a given pair of original and simplified texts. The interpretability module has been used to show the most impactful features that can highlight the difference in complexity between the two textual versions.

The results of our experiment allow us to position ourselves in relation to the objectives (i) and (ii) cited above:

- About the objective (i): the results show that each original version is automatically calculated as more complex than the simplified one. Using the differences between predicted age ranges as a measure, we see that the original texts have indeed on average +1.295 years higher recommended age than the simplified texts.
- About the objective (ii): the results show that, when using the interpretability module, we extract "niveau_lexical", "phonetique", "pronoms", "parties_du_discours", "dependances_syntaxiques", and "flexions_verbales" as the most

impactful features to describe the difference between two versions of texts, and then to describe texts' complexity (see fig. 1). These results are in line with the expected ones as they are reported in (Gala et al., 2020): the groups of features that the experts use to manually simplify texts are "lexique", "phonetique", "morphosyntaxe", and "syntaxe".

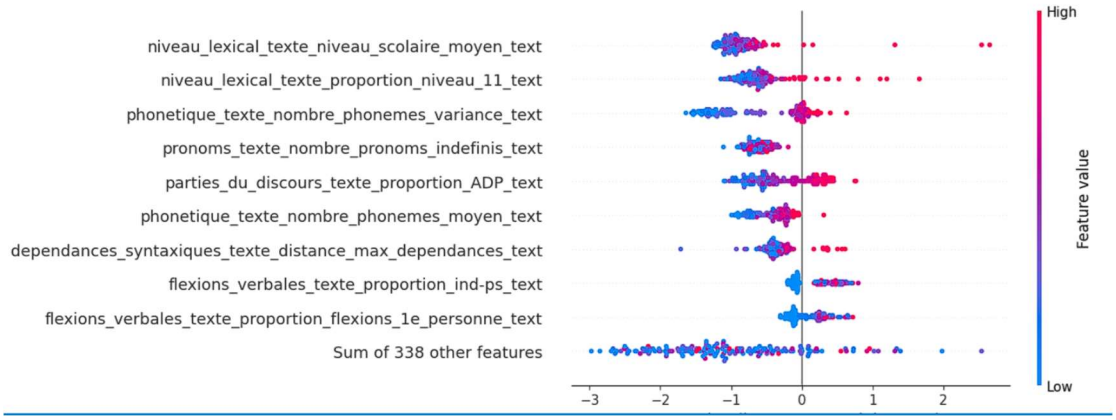


Figure 1. Most impactful features in recommending the age ranges for the Alector corpus

[1] https://texttokids.irisa.fr/
[2] https://texttokids.ortolang.fr/chain/

Bibliography:

Balyan, R., McCarthy, K. S., & McNamara, D. S. (2020). Applying natural language processing and hierarchical machine learning approaches to text difficulty classification. International Journal of Artificial Intelligence in Education, pp. 1–34

Battistelli, D., Etienne, A., Rahman., R., & Teissèdre, C. (2022). Une chaîne de traitements pour prédire et appréhender la complexité des textes pour enfants d'un point de vue linguistique et psycho-linguistique. In Proceedings of TALN 2022 (Traitement Automatique des Langues Naturelles), pp. 236-246

Bayot, R. K., & Gonçalves, T. (2018). Age and gender classification of tweets using convolutional neural networks. International Workshop on Machine Learning, Optimization, and Big Data Lecture Notes in Computer Science 2018, vol. 10710 LNCS, pp. 337-348

Blandin A., Lecorvé G., Battistelli D., & Etienne A. (2020) - "Recommandation d'âge pour des textes". In Proceedings of TALN 2020 (Traitement Automatique des Langues Naturelles), pp. 164–171, Nancy, France

Chen, T., Guestrin, C. (2016). XGBoost : un système de boosting d'arbres évolutif. http://arxiv.org/abs/1603.02754

Chen, J., Cheng, L., Yang, X., Liang, J., Quan, B., & Li, S. (2019). Joint learning with both classification and regression models for age prediction. Journal of Physics: Conference Series

Gala, N., Tack, A., Javourey-Drevet, L., François, T., & Ziegler, J.C. (2020). Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. Proceedings of LREC 2020 (12th Conference on Language Resources and Evaluation), pp. 1353-1361

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. In Proceedings of NIPS 2017 (31th International Conference on Neural Information Processing Systems). Curran Associates Inc., Red Hook, NY, USA, pp. 4768–4777

Mesgar, M., & Strube, M. (2018). A neural local coherence model for text quality assessment. In Proceedings of EMNLP 2018 (Conference on Empirical Methods in Natural Language Processing), pp. 4328-4339

Nguyen, D., Smith, N. A., & Rose, C. (2011). Author age prediction from text using linear regression. In Proceedings of ACL-HLT (5th workshop on language technology for cultural heritage, social sciences, and humanities), pp. 115–123

Rahman, R., Lecorvé, G., Etienne, A., Béchet, N., Chevelu, J. & Battistelli, D. (2020). Mama/Papa, Is this Text for Me?. In Proceedings of COLING 2020 (28th International Conference on Computational Linguistics), 8-13 december 2020, Barcelona, Spain

Rahman, R., Lecorvé, G., & Béchet, N. (2023). Age Recommendation from Texts and Sentences for Children. Retrieved from https://arxiv.org/abs/2308.10586: https://doi.org/10.48550/arXiv.2308.10586

Wilkens, R., Alfter, D., Wang, X., Pintard, A., Tack, A., Yancey, K. P., & François, T. (2022). Fabra: French aggregator-based readability assessment toolkit. In Proceedings of LREC 2022 (13th Language Resources and Evaluation Conference), pp. 1217–1233

Yancey, K., and Pintard, A. François, T. (2021). Investigating readability of French as a foreign language with deep learning and cognitive and pedagogical features. In Lingue e Linguaggio, pp. 229-258

Detection of Complexity in General and Medical-language Texts Using Eye-Tracking Data

Oksana Ivchenko, Natalia Grabar
Univ. Lille, CNRS, UMR 8163 - STL - Savoirs Textes Langage,

Keywords: Medical Texts, Types of Texts, Simplification, Reading, Eye-Tracking, Fixations

This study explores the complexity of medical texts, aiming to objectively detect difficult words and passages using eye-tracking data. We investigate how four types of texts (general topics from Wikipedia, medical topics from Wikipedia, clinical cases, and their manually simplified versions) are read and understood by participants. These simplifications were made manually at both lexical and syntactic levels, following text simplification guidelines. Using eye-tracking technology, we measure the cognitive load imposed by these texts, focusing on key eye-tracking metrics that serve as reliable indicators of reading difficulty (Ekstrand et al., 2021; Clifton et al., 2007; Singh et al., 2016). The goal is to identify which lexical and syntactic constructions in medical texts hinder comprehension and to provide insights for simplifying such texts. To date, we analyze data from 50 participants. Eye-tracking metrics are being analyzed both quantitatively and qualitatively to identify reading patterns and assess how reading behavior varies across different text types.

In a preliminary analysis, we compare the number and duration of fixations across four types of texts. Fixations, characterized by brief pauses during reading, are crucial for information processing and serve as indicators of cognitive engagement with the text. Longer fixations often signal processing difficulty or heightened interest, while more frequent fixations may suggest that the text is either challenging or highly engaging for the reader. The average duration of an eye fixation on a word during reading varies depending on several factors, including the complexity of the text, the reader's familiarity with the content, and the purpose of reading (Hyönä & Kaakinen, 2019). Generally, studies on eye movements during reading indicate that adults typically fixate on a word for about 200 to 250 milliseconds (ms) when reading in their native language under normal circumstances (Rayner & Reingold, 2015).

Quantitative analysis using various statistical tests indicates that text type significantly impacts reading ease (Figure 1), as evidenced by fixation measures : both the duration and number of fixations. In our experiment, clinical cases were the most difficult to read, followed by medical and general language encyclopedia articles, while the simplified version of clinical cases considerably eased the reading process. The statistical analysis also reveals a correlation between fixation duration and the number of fixations : complex words typically require longer and more frequent fixations, indicating a greater cognitive effort.

Interestingly, the quantitative analysis shows that clinical cases present the highest level of difficulty, yet their simplification significantly improves readability and hypothesized comprehension. In the qualitative analysis, we identified the top words that demand the most attention from readers in each text type. These words typically correspond to technical medical terms. In this way, we aim to conduct a comprehensive analysis of other eye-tracking metrics (saccades, regressions, reading and rereading times, etc.) to identify patterns of difficult passages within the text. In addition to traditional statistical analysis, we are training a neural network to predict eye-tracking features. The goal is to use eye-tracking data to identify the complexity of texts, particularly their type (medical, clinical, general, or simplified). By leveraging these predictions, the neural network can help classify text complexity based on reading patterns, which could ultimately assist in automatic text simplification.

Bibliography :

Clifton, C., Staub, A. & Rayner, K. (2007). Eye movements in reading words and sentences. Eye movements : A window on mind and brain.

Ekstrand, A., Nilsson, M. & Öqvist Seimyr, G. (2021). Screening for reading difficulties : Comparing eye tracking outcomes to neuropsychological assessments. Frontiers in Education, 6.

Hyönä, J. & Kaakinen, J. K. (2019). Eye Movements During Reading, In C. Klein & U. Ettinger, Eds., Eye Movement Research : An Introduction to its Scientific Foundations and Applications, pp. 239–274. Springer International Publishing : Cham.

Rayner, K. & Reingold, E. (2015). Evidence for direct cognitive control of fixation durations during reading. Current Opinion in Behavioral Sciences, 1, 107–112.

Singh, A., Mehta, P., Husain, S. & Rajakrishinan, R. (2016). Quantifying sentence complexity based on eye-tracking measures.

Les sourds signeurs à l'épreuve de la complexité syntaxique du français écrit : le cas de la subordonnée relative

Adrien Dadone
SFL - Université Paris 8

Contexte. Notre travail se veut une contribution à une didactique de l'écrit adressée aux personnes sourdes ayant pour langue première la langue des signes française (LSF). Malgré un essor relatif de la recherche dans ce domaine [1], l'élaboration d'une didactique tenant compte des spécificités de ce public reste à venir. C'est dans cette perspective que s'inscrit notre recherche. En particulier, à travers l'étude de l'appropriation d'une structure syntaxique complexe, celle de la subordonnée relative du français, nous avons cherché à repérer ce qu'il pouvait y avoir de spécifique dans l'apprentissage de l'écrit pour ce public. Pour ce travail, le cadre d'analyse de la complexité syntaxique se rapporte à celui de la subordination, soit lorsqu'une phrase est construite sur un lien de dépendance entre une proposition dite subordonnée et une proposition principale [2].

Objectifs. La LSF offre la possibilité d'exploiter l'espace devant le corps et d'articuler divers segments du corps simultanément. Cela ouvre à ces langues des potentiels d'expression de la complexité différents de ceux offerts par les langues vocales, notamment par la spatialisation, procédé décrit plus loin. Nous avons fait l'hypothèse d'une incidence de ces propriétés sur l'appropriation d'une L2 comme le français écrit qui, lui, est essentiellement linéaire : notre hypothèse de travail a donc été que les différences liées au canal (audio-phonatoire vs visuo-gestuel) peuvent fournir des explications quant aux spécificités des écrits produits par les sourds signeurs. Nous avons fait le choix de nous centrer sur les usages de la relative en français écrit chez des sourds signeurs en production et en compréhension. Un ensemble de questions ont guidé notre recherche : la production et la compréhension écrites laissent-elles apparaitre des marques liées à la différence de canal lorsque les personnes sourdes sont confrontées à des relatives ? Quelles formes prennent ces marques ? Révèlent-elles une appropriation de l'écrit spécifique au public sourd ?

Méthodologie. Nous avons d'abord mené une étude contrastive LSF-français écrit visant à identifier les équivalents fonctionnels en LSF de la relative du français écrit. Pour cela, nous avons collecté diverses traductions existantes de textes/discours contenant des phrases avec des relatives du français vers la LSF mais aussi de la LSF vers le français (c'est-à-dire lorsque la traduction vers le français fait apparaître une relative). Notre corpus est constitué d'articles issus du site d'information Média-Pi! qui propose des publications écrites bilingues français/LSF [3]. Nous avons également inclus deux romans traduits en LSF : une nouvelle, La Parure, et un journal imaginaire, Le Horla, tous deux de Maupassant. Toutes les traductions du français vers la LSF sont le fait de traducteurs sourds diplômés. Conformément à l'approche fonctionnelle, pour laquelle les formes linguistiques sont façonnées par les intentions de communication, notre étude contrastive a considéré ensemble les relations entre structures, fonctions et contexte. Le travail d'annotation des discours signés a été réalisé avec le logiciel ELAN [4]. Cette étude contrastive des deux systèmes linguistiques, LSF et français écrit, nous a permis de pouvoir identifier les effets potentiels de la langue source (LS) dans l'usage de la langue cible (LC).

Des tâches de production et de correction à l'écrit ont été ensuite demandées à des sourds signeurs ayant été scolarisés en France. Pour la tâche de production, nous nous sommes inspiré du protocole expérimental de projet Spencer [5] qui avait pour but d'étudier l'apprentissage de l'écrit dans différents contextes et pour différentes langues. Les locuteurs ont d'abord visionné une vidéo rapportant une série de problèmes à l'école. Ils ont été ensuite invités à produire en LSF et en français écrit un récit personnel en lien avec la vidéo. Il leur a été demandé dans un second temps de produire un contenu de type expositif également en LSF et en français écrit. Pour la tâche de correction, nous avons imaginé la situation d'une personne non francophone qui, se plaignant du bruit, écrit un mot à ses voisins. La consigne était de repérer les « erreurs » du texte soi-disant produit par cette personne. Nous avons également soumis le protocole à un groupe contrôle constitué d'apprenants entendants du français, aux L1 typologiquement plus ou moins proches du français ainsi qu'à des natifs francophones. Afin d'avoir des éléments de comparaison homogènes, notre recherche s'est concentrée sur des individus présentant un niveau entre B1 et B2 du CECRL. Au total, des passations ont été effectuées auprès de 20 sourds, et, parallèlement, de 18 entendants.

Résultats et discussion. Au moins deux spécificités ont pu être observées dans les écrits des sourds de notre panel ainsi que dans leur tâche de correction. La première suggère que la relative n'est pas toujours perçue comme une subordonnée. Plusieurs phrases ne comportent pas le verbe de la proposition hiérarchiquement supérieure, à l'instar de (1) ; comme si la relative suffisait à la formation d'une phrase complète.

(1) Ça s'est un peu mal passé, les camarades de la classe qui n'avaient jamais eu d'expériences avec des personnes sourdes auparavant. (B. 31 ans)

Au total, ce type de phrase apparait dans les écrits de 7 individus du groupe sourd (sur 20) et dans aucun de ceux du groupe contrôle. Cette marque apparait surtout dans les niveaux les plus faibles de la cohorte mais se rencontre également chez une personne de niveau B2. En parallèle, le texte de la tâche de correction comportait une phrase avec une proposition relative mais sans verbe au sein de la proposition principale. Encore une fois, le groupe sourd se distingue nettement du groupe contrôle. Alors qu'un seul participant des groupes contrôle n'a pas repris la phrase, 7 individus du groupe sourd n'ont pas cherché à reformuler la phrase.

Le deuxième résultat significatif porte sur un certain usage idiosyncrasique du mot-outil « que », usage déjà repéré dans des travaux antérieurs [6]. Notre corpus révèle qu'il peut être investi de multiples façons, ce qui donne lieu à des expressions de formes variées. « Que » semble être avant tout le connecteur syntaxique par défaut (2), qui peut s'ajouter à un autre connecteur déjà présent dans la séquence et qui s'étend au-delà du cas de la relative (3).

(2) Je voyais les questions que j'ai des trous de la mémoire. (J. 23 ans)

(3) …je les avais aperçu de loin et qu'elles me voyaient aussi. (N. 36 ans)

Du côté du groupe contrôle, on trouve une seule occurrence de ce type, chez un apprenant sinophone.

(4) …quand le deuxième fois que j'essayer de chercher la charactaire pour répondre la questionne. (Z. 26 ans)

Si la complexité syntaxique de la LC est une zone de difficulté potentielle désormais bien identifiée [7], l'écrit sourd, même à des niveaux avancés laisse apparaitre des marques propres à ce public, alors même que notre corpus montre que les autres catégories de la LC sont mieux maitrisées. On peut mettre les deux résultats discutés en regard du mode de fonctionnement de la LSF. L'analyse contrastive nous a en effet permis de montrer que pour exprimer les valeurs sémantiques de la relative, la LSF pouvait se passer de subordonnant et de pronom par le recours à la spatialisation. Ce procédé permet d'introduire, de maintenir ou de réintroduire une référence en activant/réactivant une portion d'espace sémantisée. La spatialisation des unités lexicales et l'absence d'un équivalent du mot-outil « que » dans la LSF pourraient alors expliquer, en partie, la construction syntaxique idiosyncrasique dans les interlangues observées, les sourds signeurs devant reconstruire la structure syntaxique de la LC depuis leur LS. De même, la perception de la subordonnée relative chez certains sourds suggère que les productions signées structurées spatialement ne trouvent pas de correspondance structurale avec une complexité syntaxique reposant sur la linéarité et structurée en principales et subordonnées introduites par un mot-outil. Ces résultats nous incitent à poursuivre le travail de l'analyse contrastive afin de constituer un support de travail pour l'enseignement de l'écrit destiné aux sourds signeurs.

Bibliographie

[1] Perini. & Dadone. (2024). Surdité et accès à la littératie : quel rôle pour les langues des signes ? Langages, 235(3), 71-85.

[2] Riegel & al. (2009). Grammaire méthodique du français (4e éd.). Paris, Presses universitaires de France.

[3] Média-Pi! (2024). Media-Pi! Retrieved September 3, 2024, from http://www.media-pi.fr

[4] ELAN (Version 6.8) [Computer software]. (2024). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

[5] Berman. & Strömqvist. (1999). Typological perspectives on developing literacy, Developing Literacy Across Genres, Modalities, And Languages, 1, 21-22.

[6] Perini (2013). Que peuvent nous apprendre les productions écrites des sourds ? Analyse de lectes écrits de personnes sourdes pour une contribution à la didactique du français écrit en formation d'adultes, Thèse de doctorat, Université Paris 8.

[7] Ortega. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. Applied linguistics, 24(4), 492-518.

Éléments de complexité syntaxique dans des écrits de scripteurs sourds en langue française

Mireille Esther Gettler Summa, Raphaël Prénovec, Chunxiao Yan, Caroline Bogliotti, Anne Lacheret Dujour
Modyco UMR 7114 CNRS & Université Paris Nanterre

Mots clés : corpus, scripteurs sourds, compétence syntaxique, complexité syntaxique, flux de dépendance

Nous présentons dans ce travail des résultats concernant la complexité syntaxique, issus d'approches quantitatives des domaines de l'apprentissage non supervisé et du traitement automatique du langage, sur le corpus ALPHA (Prenovec, 2022, 2024), corpus d'écrits de scripteurs sourds et dont la langue de communication quotidienne est majoritairement la Langue des Signes Française (LSF). Les langues des signes présentent des caractéristiques structurales différentes des langues vocales, en particulier concernant leurs propriétés syntaxiques (Jepsen et al., 2015) Ces spécificités syntaxiques ont des conséquences sur les propriétés formelles de leurs productions langagières en situation d'écriture d'une langue vocale, notamment concernant la linéarisation syntaxique. Le travail que nous avons mené sur ces écrits en français a pour objectif d'en rendre compte, à travers un indice de complexité syntaxique qui repose sur une métrique basée sur les flux de dépendance (Figure 1 ; Yan, 2021). Nous avons mis à l'épreuve cette mesure de complexité sur quatre profils construits par classification multidimensionnelle des scripteurs à partir d'annotations d'erreurs syntaxiques dans les textes et de relevés d'autres éléments syntaxiques : les constructions parataxiques et les subordonnées.
Le Corpus ALPHA comprend 218 textes (descriptifs, narratifs, explicatifs ou argumentatifs) produits par soixante-six scripteurs collégiens sourds (âge médian 16 ans).
Les erreurs épinglées dans chaque copie ont été doublement annotées sur les bases d'un guide d'annotation développé spécifiquement pour l'étude dans le cadre de la thèse de doctorat de Prenovec (2024).
Sur l'ensemble des erreurs annotées (orthographe lexique, morphologie, syntaxe et sémantique), 3197 erreurs de productions syntaxiques ont été relevées, soit 37,6% de l'ensemble des erreurs annotées. Deux analyses principales ont été conduites : i) la fréquence des subordonnées dans les textes des sourds, ii) la nature et la fréquence du mot subordonnant qui les introduit (sauf dans les cas particuliers des propositions infinitives). Les résultats nous permettent de constater que certains scripteurs utilisent couramment les subordonnées comme dans l'exemple (1) ci-dessous :

(1) A ce moment **quand** elle m'a annoncé **qu**'on allait partir j'étais à la fois contente mais à la fois inquiète car je devais recommencer une nouvelle vie dans une ville **que** je ne connaissais pas

D'autres scripteurs, comme le scripteur 3-B-3GB1-25/02/2020 qui n'a produit qu'une subordonnée, utilisent très rarement les subordonnées.

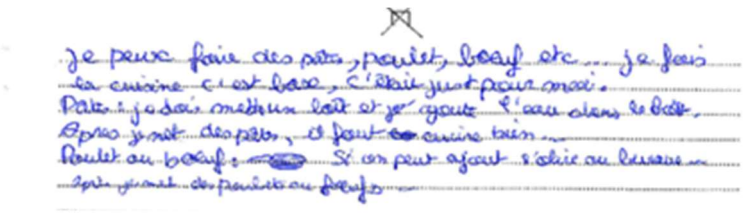
Code	SUBORDONNEES	PARATAXES	NB PROPOS. SANS PREDICAT	LONGUEUR DE L'ECRIT	LINEARITE
1-A-2G-20/12/2019	4	Un peu	2	Bonne	Bonne
3-B-3GB1-25/02/2020 et 28/02/2020	1	Beaucoup	2	Bonne	Assez Bonne
4-B-3GB1-25/02/2020 et 28/02/2020	9	Pas du tout	0	Bonne	Bonne

Tableau 1. Extrait de décomptes effectués à partir de l'analyse du corpus ALPHA. Dans ce tableau, les six colonnes représentées parmi les 22 du tableau global, représentent pour l'ensemble de la production d'un scripteur de gauche à droite : l'identifiant du scripteur, le nombre de subordonnées dans ses textes, la fréquence de la pratique de la parataxe catégorisée sur trois modalités, le nombre de propositions sans prédicat, le nombre de mots de l'écrit catégorisé sur trois modalités, le respect de la linéarité de la langue française catégorisé sur trois modalités

La parataxe est également une caractéristique intéressante dans la mesure où la présence d'au moins 2 prédicats dans une phrase la conduit à être une phrase complexe. En effet, l'absence de marqueur explicite de dépendance entre les propositions peut impacter le traitement de la phrase.
Un algorithme de classification multidimensionnelle, appliqué sur le tableau global dont est extrait le tableau 1, a conduit à une proposition de typologie des scripteurs. Nous avons dégagé quatre profils de compétence syntaxique que nous avons libellés : bonne syntaxe, assez bonne syntaxe, assez mauvaise syntaxe, mauvaise syntaxe.
Nous prouvons ainsi, en interprétant les classes par leurs variables statistiquement significatives, que les sourds ont des compétences syntaxiques variables, mais aussi bien caractérisables. C'est donc sur ces quatre profils respectifs que nous

avons mis à l'épreuve un algorithme de calcul de complexité syntaxique introduit dans le cadre du développement des Tree banks Universal Dependencies (Nivre et al., 2016). L'indice associé est en relation d'une part avec l'ordre des mots et d'autre part avec la mémoire de travail. Il s'agit là de la contrainte psycholinguistique de mémorisation, limitée à 7 +/- 2 unités (ici, des mots) en amont d'un mot donné dans un parcours linéaire du texte. Selon notre hypothèse, les scores de complexité des textes seraient corrélés de façon ordinale aux niveaux de compétence syntaxique : meilleurs sont les scores, plus robuste est la syntaxe. C'est une considération peut-être trop simple par rapport à la difficulté à cerner le concept de compétence syntaxique. Par conséquent, nous attendons des scores significativement liés à la qualité syntaxique de chaque profil, et ce pour chacun des profils construits de scripteurs sourds. Nous prenons ici l'option d'une métrique dans le cadre d'une syntaxe de dépendance.
Les calculs sont donc effectués sur chacun des quatre sous-ensembles des textes profilés. Il est à remarquer que le parser n'a pas pris en charge certaines parties des textes, cependant de faibles occurrences, qui comportent des segments inconnus comme des dessins en substitution de mots. En exemple le texte (2) du scripteur 15-B-3GB2 :

(2)



L'enchaînement des étapes consiste à (i) parser le texte, (ii) construire l'arbre de dépendances, (iii) calculer la mesure de complexité.
Voici en exemple la phrase (3) du scripteur 32-B-4P-27/02/2020, classé dans le profil mauvaise syntaxe dans la typologie multidimensionnelle :

> (3) On a mis la table et on veut manger parce que on avait vraiment faim donc je dis à ma grand-mère j'ai faim et je mange avec ma grand-mère, mon père, ma belle mére, et ma sœur et on a fini de manger de plat, on a envie de mangé de fromage et ma belle-mère donne des ordre à ma grande soeur quand elle regarde son telphone et ma belle mère donne les ordres à ma soeur après ma soeur se mit à pleurer parce que elle n'aime pas que ma belle-mère, que elle chit sur ma soeur après ma soeur rentrer dans la chambre de ma grand-mère et elle pleure après ma belle-mère rentre dans la chambre dans la chambre de ma grand-mère après elle continue à mettre à crie après mon papa rentre dans la chambre de ma grand-mère il va chercher ma belle-mère qui dedans la chambre de ma grand-mère après il tire le maillot de ma belle mère après ma belle-mère et mon père sont paris sans nous. »

On remarque que la séquence est particulièrement longue : des conjonctions de coordination sont observées là où l'on attendrait des signes de ponctuation. Une segmentation est donc un préalable nécessaire pour parser le texte. L'ajout des signes de ponctuation et la correction des fautes d'orthographe constituent également une étape incontournable pour la bonne exécution de l'algorithme. Cette étape n'est pas univoque et peut donc prêter à critique.
L'algorithme de calcul de complexité utilisé ici repose sur la taille maximale des sous-flux disjoints, également appelés « weights » ou poids du flux (Kahane et al., 2017).
Nous illustrons ces notions dans la Figure 1 par un exemple de calcul de poids sur une phrase courte, hors Alpha, de façon à pouvoir visualiser le processus. Le flux de dépendance dans une position donnée (entre deux mots dans une phrase) est l'ensemble des dépendances qui relient un mot à gauche de cette position à un mot à droite (Kahane & Yan 2019). On indique les positions inter-mots du flux de dépendances par les lignes verticales pointillées, et on peut lire les poids du flux associés par les nombres placés en dessous.

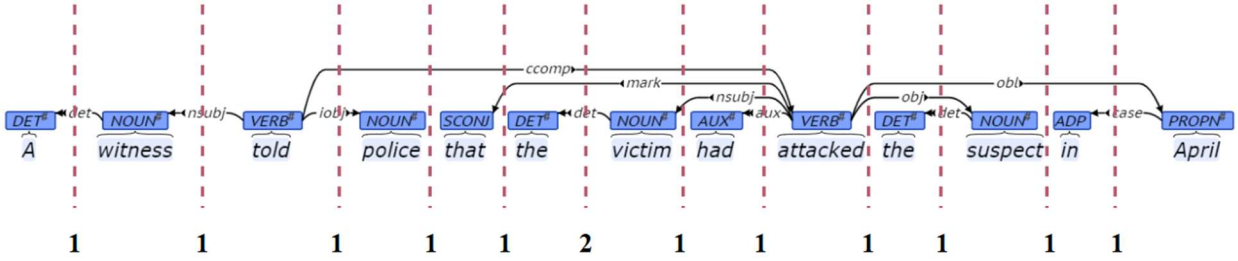


Figure 1. Les positions du flux dans une phrase

En résultats sur les quatre classes de scripteurs de Alpha, nous obtenons pour tous les profils, 1 comme score minimal de complexité, et 2 comme score maximal. Mais en moyenne, nous trouvons les scores suivants pour chacun des profils:
bonne syntaxe : 1, 24 ; assez bonne syntaxe :1, 22 ; assez mauvaise syntaxe : 1, 11 ; mauvaise syntaxe : 1, 09.
Notre hypothèse semble donc validée : les scores de complexité syntaxique sont ordonnés selon la qualité de la syntaxe (au sens de la typologie syntaxique en sortie de l'apprentissage non supervisé).
Il reste à comprendre en profondeur cette adéquation de la mesure de complexité syntaxique à l'acception de la qualité syntaxique mise en lumière par la typologie préalable des scripteurs. Ce travail est à compléter aussi dans diverses directions : âge des scripteurs, nécessité d'un découpage révisé des textes à l'étape parseur, échantillons contrôle. Nous n'avons en effet pas de résultats sur la complexité syntaxique en français écrit pour des entendants de l'âge des sourds Alpha. Enfin si l'on peut comparer l'ordre des scores, on ne sait pas encore les évaluer en absolu, par exemple par un test de significativité statistique pour variables ordinales, en rapport aux scores calculés sur d'autres corpus (Candito et al., 2014).

Bibliographie :

Candito, M., Perrier, G., Guillaume, B., Ribeyre, C., Fort, K., Seddah, and de la Clergerie. E. (2014). Deep Syntax Annotation of the Sequoia French Treebank. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 2298–2305, Reykjavik, Iceland. European Language Resources Association (ELRA)

Jepsen, J. B., De Clerck, G., Lutalo-Kiingi, S., & McGregor, W. B. (Éds.) (2015). Sign Languages of the World : A Comparative Handbook. DE GRUYTER. https://doi.org/10.1515/9781614518174

Kahane, S., Yan, C., & Botalla, M. A. (2017). What are the limitations on the flux of syntactic dependencies? Evidence from UD treebanks. In 4th international conference on Dependency Linguistics (Depling) (pp. 73-82).Marillier, (2010). La parataxe, entre indépendance et intégration, Edition scientifique Internationale Berne, 333-352

Kahane, K., Chunxiao Yan, C. (2019). Advantages of the flux-based interpretation of dependency length minimization In Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019), pages 98–109, Paris, France. Association for Computational Linguistics

Nivre, J., de Marneffe, M.C., Ginter, F. Goldberg, Y., Hajič, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. & Zeman, D. (2016). Universal Dependencies v1: A Multinlingual Treebank Collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA)

Prenovec, R. (2022) La littératie chez les sourds : états des lieux sociolinguistique et linguistique, perspectives didactiques hal-03641243 , version 1 (08-09-2022)

Prenovec (2024), La littératie sourde : le Corpus ALPHA et son exploration dans le domaine de compétence en orthographe lexicale, Journée scientifique MoDyCo, 5 septembre 2024, Université Paris Nanterre

Yan, C. (2021). Complexité syntaxique et flux de dépendance : études quantitatives dans les treebanks universal

Disentangling Structural and Developmental Complexity in the Acquisition of *C'est*-clefts in L1 French

Tess Wensink[1], Karen Lahousse[1], Cécile De Cat[2], Katerina Palasis[3], Béatrice Busson[1]
[1]KU Leuven, [2]University of Leeds, [3]Université Côte d'Azur
Research Foundation - Flanders (FWO)

Keywords: cleft sentences, structural complexity, developmental complexity, acquisition, French

This talk explores the relation between structural complexity, which refers to the formal properties of language, such as the number of elements and the relational patterns between them, and developmental complexity, which concerns the emergence and mastery of linguistic structures during acquisition (Pallotti, 2015). While structurally complex constructions are frequently developmentally complex and thus emerge late in language development (Ellis, 2009; Hamann & Tuller, 2014), this talk provides empirical evidence for the importance of distinguishing between the two forms of complexity in linguistic research.

The empirical focus of this talk is the development of structural (morphosyntactic) complexity of *c'est*-clefts in L1 French (1). Such *c'est*-clefts are structurally complex due to their bi-clausal nature, consisting of a matrix clause and an embedded clause (2).

(1) C' est moi qui lis.
It is me who read
'It's me who is reading.'

(2) C' est moi_i qui t_i lis.
it is me who read

METHOD. Our analysis involves a quantitative analysis of the syntactic properties of 1087 spontaneously produced c'est-clefts in L1 French, drawing on:

- 479 clefts from 16 children (ages 2-6) from the Palasis (2009) corpus,
- 356 clefts from 4 children (ages 1-5) from the York corpus (De Cat & Plunkett, 2002), and
- 252 clefts from 3 children (ages 1-4) from the Lyon corpus (Demuth & Tremblay, 2008).

Structural complexity is operationalized based on degrees of morphosyntactic complexity, specifically whether the cleft includes a Cleft-Relative Clause (CRC), a resumptive pronoun, and a complementizer (Table 1). Developmental complexity is operationalized as the proportion of different formal cleft types (Table 1) at the group level, and the order of appearance of each formal cleft type in individual children.

RESULTS. Our study reveals that children produce various formal types of c'est-clefts, each with differing degrees of structural complexity (Table 1; see also Lahousse and Jourdain (2023)). Full clefts, which involve an embedded clause with a complementizer, are structurally the most complex (6) (7). Clefts with a CRC-attempt lack a complementizer, making them less complex (4) (5). Lastly, reduced clefts without CRC are the least complex (3). We observe a clear developmental trajectory in the emergence of these cleft types: children younger than 2,5 years old produce reduced clefts without CRC (3) and clefts with a CRC-attempt [- C] (4) (5) before full clefts [+ C] (6) (7). Hence, structurally more complex clefts emerge later in development, suggesting an alignment between structural and developmental complexity. However, we also find that some children continue to produce (non-adultlike) clefts with a CRC-attempt [- C] (4) (5) after having produced a full cleft [+ C]. This persistence in using structurally less complex constructions, despite having demonstrated the ability to produce more complex ones, may be attributed to the morphosyntactic properties of clefts or to children's processing limitations. Further research is necessary to investigate these possibilities. This finding thus suggests that factors beyond mere structural complexity play a role in language development. Hence, this study underscores the necessity of differentiating between structural and developmental complexity to fully understand the nuances of linguistic acquisition.

Table 1. Formal types of c'est-clefts

Reduced clefts [- CRC]	Clefts with a CRC-attempt [- C]		Full clefts [+ C]	
	[- resumptive]	[+ resumptive]	[+ resumptive]	[- resumptive]
(3) *C'est moi.* it's me	(4) *C'est moi* it's me *lis.* read	(5) *C'est moi*$_i$ *je*$_i$ it's me I *lis.* read	(6) *C'est moi*$_i$ *que* it's me that *je*$_i$ *lis.* I read	(7) *C'est moi*$_i$ *qui*$_i$ it's me that *lis.* read
adultlike	non-adultlike	adultlike in informal French	adultlike in some varieties	adultlike

Bibliography :

De Cat, C., & Plunkett, B. (2002). QU'est ce qu'i (l) dit, celui+ Là?: notes méthodologiques sur la transcription d'un corpus francophone. Romanistische Korpuslinguistik: Korpora und gesprochene Sprache/Romance Corpus Linguistics: Corpora and Spoken Language. Narr.

Demuth, K. & Tremblay, A. (2008). Prosodically-conditioned variability in children's production of French determiners. Journal of Child Language, 35, 99–127.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. Applied Linguistics 30, 474–509.

Hamann, C., & Tuller, L. (2014). Genuine versus superficial relatives in French: the depth of embedding factor. Revisita di grammatica generativa, 36, 146–181.

Lahousse, K., & Jourdain, M. (2023). The emergence and early development of c'est 'it is' clefts in French L1. In C. Bonan & A. Ledgeway (Eds.), It-Clefts: Empirical and Theoretical Surveys and Advances (pp. 135–156). Walter de Gruyter.

Palasis, K. (2009). Syntaxe générative et acquisition: Le sujet dans le développement du système linguistique du jeune enfant [Doctoral Dissertation, Université de Nice-Sophia Antipolis]. https://doi.org/10.13140/2.1.2722.0806

Pallotti, G. (2015). A simple view of linguistic complexity. Second Language Research, 31(1), 117–134. https://doi.org/10.1177/026758314536435

Définir et mesurer la complexité en acquisition du Français Langue Étrangère et Seconde

Nathalie Gettliffe
UR2310 LISEC, Université de Strasbourg, de Haute-Alsace et de Lorraine

En acquisition des langues étrangères et secondes, les mesures de complexité linguistique sont souvent associées à deux autres mesures à savoir, l'aisance à communiquer et la précision linguistique afin de définir différents stades de développement linguistique (Skehan, 1998 ; Wolfe-Quintero & coll., 1998). Toutefois, d'un point de vue conceptuel, la notion de complexité n'est pas encore stabilisée (Deng & coll., 2021). En effet, certains chercheurs hésitent entre une définition qui se centrerait sur la complexité linguistique inhérente à la langue (ex : le passif) (Housen & coll., 2012) alors que d'autres invoquent la complexité cognitive pour expliquer pourquoi certaines structures sont apprises plus tardivement par les apprenants en langue étrangère (Larsen-Freeman, 2009). De plus, les modèles explicatifs (Robinson, 2001 : multiple resources attentional model ; Skehan, 2009 : limited attentional model ; Larsen-Freeman, 2009 : interconnection model) qui tentent de rendre compte de la complexité croissante des productions des apprenants ne positionnent pas la complexité au même niveau (finalité ou variable d'ajustement au détriment de l'aisance à communiquer et de la précision linguistique). Finalement, les mesures traditionnellement mobilisées (complexité syntaxique et complexité lexicale) non seulement ne semblent pas opérationnelles pour tous les niveaux et toutes les langues (ex : les propositions dépendantes) (Norris & Ortega, 2009 ; Yu, 2021) mais ce sont souvent toujours les mêmes mesures qui sont utilisées dans les études (Proposition/Unité T ; Proposition dépendante/Proposition ; types de mot/nombre de mots ; index de Guiraud) au détriment de mesures portant sur la morphologie inflexionnelle, les colocations, la coordination, … (Housen & coll., 2012).

Afin d'enrichir le débat, nous nous proposerons dans un premier temps de définir d'un point de vue conceptuel ce qui pourrait caractériser la complexité linguistique dans le contexte de l'appropriation d'une langue étrangère ou seconde et comment elle pourrait s'articuler avec les autres mesures de développement que sont l'aisance à communiquer et la précision linguistique. Plus précisément, nous lierons la complexité linguistique au paradigme des mémoires (Baddeley & coll., 2020) et expliqueront comment le contexte d'appropriation (implicite ou explicite) pourrait influencer le développement de la complexité linguistique. Enfin, nous nous attacherons à préciser pour la langue française, des mesures de complexité linguistique qui pourraient permettre de rendre compte des développements linguistiques d'apprenants sur plusieurs niveaux et dans plusieurs contextes d'appropriation dans le prolongement des recherches de Ågren et coll. (2012) et de Véronique et coll. (2009).

Bibliographie :

Ågren, M., Granfeldt, J. & Schlyter, S. (2012). The growth of complexity and accuracy in L2 French: Past observations and recent applications of developmental stages. Dans A. Housen, F. Kuiken & I. Vedder (dir.), Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA. (p.21-46). John Benjamins Publishing Company.

Baddeley, A., Eysenck, M. & Anderson, M. (2020). Memory. Routledge.

Deng, Y., Lei, L. & Liu, D. (2021). Calling for More Consistency, Refinement, and Critical Consideration in the Use of Syntactic Complexity Measures for Writing. Applied Linguistics, 42(5), 1021–1028. https://doi.org/10.1093/applin/amz069

Housen, A., Kuiken F. and Vedder I. (2012) Complexity, accuracy and fluency: Definitions, measurement and research. Dans A. Housen, F. Kuiken & I. Vedder (dir), Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA. (p.2-20). John Benjamins Publishing Company.

Larsen-Freeman, D. (2009). Adjusting Expectations : The Study of Complexity, Accuracy, and Fluency in Second Language Acquisition. Applied Linguistics, 30(4), 579 589. https://doi.org/10.1093/applin/amp043

Norris, J.M. & Ortega, L. (2009). Measurement for understanding: An organic approach to investigating complexity, accuracy, and fluency in SLA. Applied Linguistics, 30(4), 555–578.

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), Cognition and Second Language Instruction (1st ed., pp. 287–318). Cambridge University Press. https://doi.org/10.1017/CBO9781139524780.012

Skehan, P. (1998). A cognitive approach to language learning. Oxford University Press.

Skehan, P. (2009). Modelling Second Language Performance: Integrating Complexity, Accuracy, Fluency, and Lexis. Applied Linguistics, 30(4), 510–532. https://doi.org/10.1093/applin/amp047

Véronique, G.D., Carlo, C., Granget, C., Kim, J.-O. & Prodeau, M. (2009). L'acquisition de la grammaire du français langue étrangère. Didier.
Wolfe-Quintero, K., Inagaki, S., Kim, H. Y., & Kim, H. Y. (1998). Second language development in writing : Measures of fluency, accuracy, & complexity. Second Language Teaching & Curriculum Center
Yu, Q. (2021). An Organic Syntactic Complexity Measure for the Chinese Language: The TC-Unit. Applied Linguistics, 42(1), 60–92. https://doi.org/10.1093/applin/amz064

Comment évaluer la complexité syntaxique des productions orales enfantines ?

Christophe Parisse[1], Loïc Liégeois[2], Christophe Benzitoun[3], Caroline Masson[4]
et Christine da Silva-Genest[5]
[1]Université de Nanterre & UMR 7114 MoDyCO, [2]Université Clermont Auvergne & UR 999 LRL,
[3]Université de Lorraine & UMR 7118 ATILF, [4]Université Sorbonne Nouvelle & EA 7345,
[5]Université de Lorraine & UR 3450 DevAH

Mots-clés : acquisition du langage – trouble développemental du langage – évaluation du langage – complexité syntaxique

Évaluation du langage enfantin

Une des caractéristiques principales du langage de l'enfant est qu'il se complexifie au fur et à mesure des années. L'évaluation de cette complexité est souvent vue comme primordiale car l'enfant doit acquérir un langage dit complexe pour décrire, persuader, inférer, etc. C'est pourquoi, la longueur moyenne des énoncés (MLU, Brown, 1973) a été et reste encore le plus souvent l'étalon de mesure du développement du langage. Mais cette mesure grossière, bien qu'efficace (Klee et al., 1989), n'est plus fiable au-delà de l'âge de 48 mois (Klee & Fizgerald, 1985 ; Blake et al., 1993). Pour autant, le langage continue de se complexifier. La proportion d'énoncés complexes s'accroit largement après l'âge de 4 ans (Diessel, 2004) et de nombreux instruments de mesure du développement du langage de l'enfant ont été proposés comme la clausal density (Berman, 1996), le Complex Syntax Type-Token Ratio (CS-TTR Witkowska et al., 2022). Il existe également des mesures plus globales du développement morphosyntaxique telles que l'IpSyn (Index of Productive Syntax, Scarborough, 1990) ou le Developmental Sentence Scoring DSS (Lee, 1974) pour l'anglais. Malgré l'utilisation de ces mesures pour évaluer le langage d'enfants plus âgés, il est difficile de s'accorder sur une mesure syntaxique pertinente et corrélée à un degré de complexité. Par ailleurs, la manière de recueillir les données est un facteur important à considérer car elle a une influence directe sur la complexité syntaxique observée. Par exemple, les tâches de récit et de restitution d'histoire induisent, chez des enfants de 5 à 7 ans, un langage plus complexe qu'une conversation (Westerveld & Vidler, 2016 ; Westerveld et al., 2004.).
C'est pour ces raisons qu'en dépit de son intérêt, les mesures de complexité syntaxique du langage enfantin s'effacent devant l'utilisation de tests de langage qui privilégient des situations de production très contraintes et dirigées. Notre travail vise à réhabiliter l'utilisation d'une mesure de la complexité de production spontanée qui fonctionne au-delà de l'âge de 4 ans.

Complexité syntaxique

L'absence d'un outil efficace et simple à utiliser pour mesurer la complexité syntaxique peut s'expliquer par le côté très intuitif de cette notion, définie de manière floue et qui fluctue en fonction des approches théoriques ou de l'âge cible des enfants observés. Une définition unique et consensuelle de la complexité n'existe pas dans les sciences du langage (Monville-Burston, 2013 ; Martinot, 2013a). La complexité peut se définir d'un point de vue structural comme une proposition indépendante et une ou plusieurs propositions subordonnées (Marinellie, 2004), comme des phrases constituées de plusieurs propositions contenant chacune un verbe conjugué (Leeman, 2002), comme une construction marquée moins fréquente, avec une structure et un traitement plus complexes (Givón, 1995), ou bien par la diversité des structures (De Clercq, 2016), par la profondeur de la structure (Ferreira, 1991), voire à l'interface entre la syntaxe et le discours en tenant compte du degré d'accessibilité des référents du discours (Gibson, 1998). La complexité peut également s'appréhender en termes de coût cognitif ou de difficulté cognitive pour l'usager de la langue. Ce qui n'est pas encore acquis est plus difficile à traiter pour le locuteur et donc plus complexe. La complexité est alors différente d'un locuteur à un autre en fonction de ses acquisitions, de son expérience langagière ou de son degré d'exposition à des constructions complexes (Canut et al., 2010 ; Martinot, 2013b).

Mesurer la complexité du langage produit par un enfant

La notion de complexité syntaxique nous amène donc à nous interroger sur ses critères définitoires, ses caractéristiques (y a-t-il une seule forme de complexité ?) et sa généralité (est-elle propre à une communauté ou à un individu ?). Notre proposition a pour objectif de combler ces manques en proposant de nouveaux outils pertinents pour apprécier un niveau de développement syntaxique d'enfants âgés de plus de 4 ans et de s'approprier la notion de complexité syntaxique à la

lumière des enjeux de l'évaluation du langage oral d'enfants. Elle ne vise pas à répondre à des questions théoriques sur ce qu'est la complexité syntaxique en général, mais à effectuer, dans du langage produit par l'enfant, des mesures de production syntaxique qui traduisent une utilisation inhabituelle ou exceptionnelle du langage. Ces mesures seront comparées à des évaluations faites par des spécialistes du langage de l'enfant pour contrôler si elles peuvent reproduire une expertise humaine.

Deux pistes ont été suivies, qui seront comparées entre elles. Une première piste « M16 (règles ad-hoc) » se base sur les caractéristiques de la grammaire de l'oral (Blanche-Benveniste,1990) et celles des productions enfantines orales spontanées. Nous avons créé une typologie du degré d'élaboration langagière et avons retenu 16 critères de complexité (e.g. présence d'un sujet nominal, d'adjectifs épithètes ou de temps verbaux spécifiques tels que le futur simple ou l'imparfait). Cette typologie est ensuite appliquée automatiquement au discours des enfants par des méthodes de traitement automatique du langage. Nous avons considéré comme complexe les énoncés qui comprennent au moins un des 16 critères. Une deuxième piste « Apprentissage TTK » est basée sur l'utilisation de mesures lexico-syntaxiques automatiques réalisées à l'aide d'un outil développé dans le projet ANR TextToKids (Blandin et al., 2020).

Ces mesures sont appliquées sur le corpus Colaje (Morgenstern et Parisse, 2012) et le corpus EVALANG (da Silva-Genest et al., 2023) constitué de productions d'enfants avec et sans trouble du développement du langage (TDL). Ce corpus a été recueilli (Liégeois et al., 2024, da Silva-Genest et al., 2023) dans des situations de production de langage spontané chez des enfants ayant 5 à 7 ans. Les 20 mesures les plus efficaces (notamment celles qui sont corrélées avec l'âge de développement du langage) sont utilisées pour calculer combien de mesures sont inhabituellement élevées dans un énoncé (plus que le 80ème quantile). A la différence de la méthode M16, la méthode TTK n'induit pas de biais tiré de l'expertise des spécialistes et se base donc sur une méthode d'apprentissage non-supervisé.

Application

Les méthodes M16 et TTK ont été appliquées aux productions spontanées de 6 enfants avec ou sans TDL âgés de 5 à 7 ans dans des situations de jeu et de récit d'expériences personnelles. Ces enfants sont issus du corpus XXX, et n'ont pas participé à l'apprentissage TTK. Cela correspond à près de 1500 énoncés. L'apprentissage TTK a été réalisé avec le reste du corpus XXX, et aussi avec le corpus COLAJE. Les résultats obtenus (énoncés complexes ou non) ont été comparé avec un codage manuel réalisé par les auteurs de l'étude. Les résultats sont présentés dans la table ci-dessous.

Mesure (méthode de traitement)	**Rappel**	**Précision**	**F1**
M16 (règles ad-hoc)	0.70	0.58	0.63
Apprentissage TTK (corpus XXX)	0.57	0.52	0.54
Apprentissage TTK (corpus COLAJE)	0.60	0.49	0.54
Combinaison de méthodes (ad-hoc et apprentissage)	0.85	0.50	0.63

La dernière ligne du tableau de résultats (COMBINAISON) contient une mesure combinant les résultats de M16 et TTK. La mesure de Rappel indique quelle proportion d'énoncés complexes est bien repérée. 0.85 indique que 85% des énoncés complexes ont été identifiés. La mesure de Précision indique combien d'énoncés considérés complexes le sont effectivement (e.g. 0.50 soit 50% des énoncés sont effectivement complexe, par notre mesure automatique). La mesure F1 tient compte à la fois du rappel et de la précision. C'est donc la meilleure mesure de la qualité des algorithmes.

Nous pouvons donc voir que la méthode ad hoc M16 est la meilleure sur notre corpus. Nous voyons aussi que la méthode TTK obtient des résultats similaires sur les deux corpus d'apprentissage. Même si elle est moins efficace, elle pourrait être plus robuste sur d'autres corpus. Ces méthodes produisent des résultats qui ne se recouvrent pas exactement. En les combinant, nous récupérons plus d'énoncés complexes au prix d'un filet qui attrape plus d'énoncés non complexes.

Conclusion

Les méthodes proposées ne sont pas parfaites, mais leur qualité n'est pas si différente de celle de nombreux outils de mesure du langage de l'enfant (Conti-Ramsden et al., 2001). Elles fonctionnent de manière automatique et rapide et sont donc facilement utilisables par des orthophonistes. Elles permettent d'identifier un nombre important d'énoncés clés avec lesquels un ou une professionnelle pourra travailler pour mieux évaluer et prendre en charge le langage d'un enfant. Nous proposons de mettre cet outil à disposition en ligne sur internet pour qu'il puisse être utilisé à ces fins.

Bibliographie :

Berman, R. A. (1996). “Form and function in developing narrative abilities: the case of ‘and’,” in Social Interaction, Context, and Language: Essays in Honor of Susan Ervin-Tripp, eds D. Slobin, J. Gerhardt, A. Kyratzis, and J. Guo (Mahwah: Lawrence Erlbaum), 343–367.

Blake, J., Quartaro, G., & Onorati, S. (1993). Evaluating quantitative measures of grammatical complexity in spontaneous speech samples. Journal of Child Language, 20(1), 139–152. https://doi.org/10.1017/S0305000900009168

Blanche-Benveniste, C. (1990). Le français parlé. Éditions du CNRS.

Blandin A., Lecorvé, G., Battistelli, D. et Étienne, A. (2020). Age Recommendation for Texts. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 1431–1439, Marseille, France. European Language Resources Association.

Brown, R. W. (1973). A first language: The early stages. Harvard University Press.

Canut, E., Bocérean, C. & André, V. (2010). De l’apprentissage au développement : une approche interactionniste de l’acquisition des constructions syntaxiques complexes chez l’enfant de 3 à 6 ans. In Neveu, F., Muni Toke, V., Durand, J., Kingler, T., Mondada, L., Prévost, S. (éds.). Congrès Mondial de Linguistique Française – CMLF 2010. DOI 10.1051/cmlf/2010092.

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). Journal of child psychology and psychiatry, 42(6), 741-748.

da Silva-Genest, C., Masson, C., & Le Mené Guigourès, M. (2023). Évaluer les compétences morphosyntaxiques et discursives d’enfants d’âge scolaire : Analyse d’une situation de jeu libre. Éla. Études de linguistique appliquée, 210(2), 179-195. https://doi.org/10.3917/ela.210.0053

De Clercq, B. (2016). Le développement de la complexité syntaxique en français langue seconde : complexité structurelle et diversité. Congrès Mondial de Linguistique Française – CMLF 2016, SHS Web of Conferences 27, 07006, DOI: 10.1051/SHS 2 shsconf/20162707

Diessel, H. (2004). The Acquisition of Complex Sentences. Cambridge University Press.

Ferreira F. (1991) “Effects of Length and Syntactic Complexity on Initiation Times for Pre-pared Utterances”, in Journal of Memory and Language, vol. (30/2).

Gibson E. (1998) “Linguistic complexity: Locality of syntactic dependencies”, 1998 vol. 68

Givón T., 1995, Functionalism and Grammar. Amsterdam & Philadelphia, John Benjamins.

Klee, T., Schaffer, M., May, S., Membrino, I., & Mougey, K. (1989). A comparison of the age-MLU relation in normal and specifically language-impaired preschool children. Journal of Speech and Hearing Disorders, 54, 226–233.

Klee, T., & Fitzgerald, M. D. (1985). The relation between grammatical development and mean length of utterance in morphemes. Journal of Child Language, 12, 251–269.

Lee, L. L. (1974). Developmental sentence analysis: A grammatical assessment procedure for speech and language clinicians. Northwestern University Press.

Leeman D., 2002, La phrase complexe. Les subordinations, Bruxelles, Duculot.

Liégeois, L., da Silva-Genest, C., Benzitoun, C., Masson, C. et Parisse, C. (2024). « Méthodologie(s) de segmentation des transcriptions de l’oral en vue de la création d’une mesure d’évaluation des productions verbales enfantines ». Colloque international « La linguistique de l’oral spontané à travers les langues : création, annotation et analyse de corpus, segmentation du discours. 23-24 mai, Paris.

Marinellie S. A., (2004). Complex syntax used by school-age children with specific language impairment (SLI) in child–adult conversation, Journal of Communication Disorders, Volume 37(6), 517-533, https://doi.org/10.1016/j.jcomdis.2004.03.005.

Martinot, C. (2013a). Acquisition de la complexité en français langue maternelle et étrangère. Travaux de Linguistique, 66(1) : 7-14.

Martinot, C. (2013b) Les phénomènes complexes de la langue sont-ils complexes pour tous les enfants ? A.N.A.E 124 : 279-287.

Monville-Burston, M. (2013). Complexité et transfert dans l'acquisition du français langue étrangère : le cas des apprenants chypriotes du FLE. Travaux de linguistique, 66, 97-134. https://doi.org/10.3917/tl.066.0097

Morgenstern, A., & Parisse, C. (2012). The Paris Corpus. Journal of French Language Studies, 22(Special Issue 01), 7–12. https://doi.org/10.1017/S095926951100055X

Scarborough, H. S. (1990). Index of productive syntax. Applied Psycholinguistics, 11, 1-22.

Westerveld, M. F., Gillon, G. T., & Miller, J. F. (2004). Spoken language samples of New Zealand children in conversation and narration. Advances in Speech Language Pathology, 6(4), 195–208. https://doi.org/10.1080/14417040400010140
Westerveld, M. F., & Vidler, K. (2016). Spoken language samples of Australian children in conversation, narration and exposition. International journal of speech-language pathology, 18(3), 288–298. https://doi.org/10.3109/17549507.2016.1159332
Witkowska, D., Lucas, L., Jelen, M. B., Kin, H., & Norbury, C. (2022). Development of complex syntax in the narratives of children with English as an Additional Language and their monolingual peers. PsyArXiv. https://doi.org/10.31234/osf.io/96dx